BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

A revised partiality model and post-refinement algorithm for X-ray free-electron laser data

Helen Mary Ginn,[a] Aaron S. Brewster,[b] Johan Hattne,[b] Gwyndaf Evans,[c] Armin Wagner,[c] Jonathan M. Grimes,[a,c] Nicholas K. Sauter,[b] Geoff Sutton[a] and David Ian Stuart[a,c]*

[a]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [b]Physical Biosciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA, and [c]Diamond House, Harwell Science and Innovation Campus, Fermi Avenue, Didcot OX11 0QX, England. *Correspondence e-mail: dave@strubi.ox.ac.uk

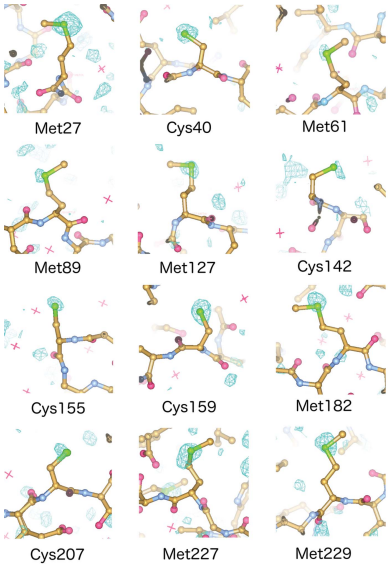
Received 12 March 2015
Accepted 6 April 2015

Keywords: post-refinement; free-electron laser; partiality.

PDB reference: 1.46 Å resolution XFEL structure of CPV17, 4zqx

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

Research towards using X-ray free-electron laser (XFEL) data to solve structures using experimental phasing methods such as sulfur single-wavelength anomalous dispersion (SAD) has been hampered by shortcomings in the diffraction models for X-ray diffraction from FELs. Owing to errors in the orientation matrix and overly simple partiality models, researchers have required large numbers of images to converge to reliable estimates for the structure-factor amplitudes, which may not be feasible for all biological systems. Here, data for cytoplasmic polyhedrosis virus type 17 (CPV17) collected at 1.3 Å wavelength at the Linac Coherent Light Source (LCLS) are revisited. A previously published definition of a partiality model for reflections illuminated by self-amplified spontaneous emission (SASE) pulses is built upon, which defines a fraction between 0 and 1 based on the intersection of a reflection with a spread of Ewald spheres modelled by a super-Gaussian wavelength distribution in the X-ray beam. A method of post-refinement to refine the parameters of this model is suggested. This has generated a merged data set with an overall discrepancy (by calculating the $R_{\rm split}$ value) of 3.15% to 1.46 Å resolution from a 7225-image data set. The atomic numbers of C, N and O atoms in the structure are distinguishable in the electron-density map. There are 13 S atoms within the 237 residues of CPV17, excluding the initial disordered methionine. These only possess 0.42 anomalous scattering electrons each at 1.3 Å wavelength, but the 12 that have single predominant positions are easily detectable in the anomalous difference Fourier map. It is hoped that these improvements will lead towards XFEL experimental phase determination and structure determination by sulfur SAD and will generally increase the utility of the method for difficult cases.



OPEN ACCESS

1. Introduction

A number of structures have been solved using serial femtosecond crystallography (SFX) at an X-ray free-electron laser (XFEL; Redecke *et al.*, 2013; Liu *et al.*, 2013; Boutet *et al.*, 2012; Tenboer *et al.*, 2014; Kern *et al.*, 2013, 2014; Ginn *et al.*, 2015), which have benefited from a large number of indexable snapshots from the crystalline samples. The recent structure of photoactive yellow protein (Tenboer *et al.*, 2014), capturing time-resolved high-resolution intermediates, used between 22 678 images and 64 496 images to generate structures, achieving a discrepancy of 6.5% between two half data sets ($R_{\rm split}$) for the latter. Such studies benefit from an abundant supply of crystalline sample, allowing the use of Monte Carlo integration (Kirian *et al.*, 2010, 2011). This method has been successfully used to observe the anomalous signal from the S atoms in lysozyme (Barends *et al.*, 2013) using 43 840 indexed patterns collected at a wavelength of 1.7 Å. Difference Fourier

peaks associated with a methionine S atom were observed to a maximum of 4.5σ at 3.5 Å resolution using the *CrystFEL* software suite (White *et al.*, 2012; Boutet *et al.*, 2012).

The structure of photosystem II (Kern *et al.*, 2013) was solved from data processed using the *cctbx.xfel* software suite, which is used, in part, in this study. The pipeline in *cctbx.xfel* has since been improved to include crystal-orientation refinement, used in the later processing of photosystem II (Kern *et al.*, 2014; Sauter *et al.*, 2014; Sauter, 2015). In systems where the supply of crystals is restricted, a high indexing rate and careful modelling of XFEL-derived data is of paramount importance (Hattne *et al.*, 2014).

Several algorithms and models have been developed, including an algorithm to minimize the distance of modelled reflections from the Ewald sphere (termed the Ewald offset correction; Kabsch, 2014), which improves the orientation matrices. This was also achieved by Sauter *et al.* (2014), successfully improving crystallographic *R* factors during refinement. The problem of assigning accurate orientation matrices to diffraction data is not unique to XFEL studies, and is helped by applying post-refinement, which was first developed by Rossmann *et al.* (1979) and Winkler *et al.* (1979) for oscillation data and essentially uses a reference set of intensities (often obtained by a preliminary merging of the current data set) as a target for the improved modelling of partially recorded reflections. White (2014) developed a post-refinement and partiality model and applied it to simulated data, seeking to assign partialities to match the fact that no reflection is fully recorded, resulting in a large improvement in merging statistics. In a separate method, Sauter (2015) applied a partiality model and several post-refinement algorithms to thermolysin data, refining a simple scale factor, linear isotropic *B* factor and orientation matrix angles all together, which improved the anomalous signal of the Zn atom in the structure. Changes to aspects of the experiment including the treatment of multiple lattices, resolution-cutoff consideration and better spot-shape models have improved the *cctbx.xfel* pipeline, without even needing to consider partiality (Hattne *et al.*, 2014). We have also previously reported a method of improving the orientation matrices and describing these partialities, which provided high-quality data (an $R_{\rm split}$ of 11.7% from 5787 crystals) and a structure determination at 1.75 Å resolution (Ginn *et al.*, 2015). However, the data used to solve this structure were calculated with only refinement of the orientation matrix, without the help of a reference data set. Here, we largely discuss refinement of the parameters which contribute to the partiality model using a reference data set.

We argue that the value of the modelled partiality is acutely dependent on accurate definition of the orientation matrices to describe XFEL data, which have been gradually improving as processing techniques progress. Here, we present an updated partiality model and a post-refinement algorithm applied to 7225 images; together these lead to more accurately defined orientation matrices and more reliable data. This partiality model is distinguished by taking into account both the size of the reciprocal-lattice point (rlp) and the multiple Ewald spheres which are intersected, which are in turn defined by the wavelength of the beam and its energy spread. We have used images generated from cypovirus polyhedra, previously solved to 1.75 Å resolution (Ginn *et al.*, 2015), as a well behaved test case. In the 1.46 Å resolution structure obtained, the differences in peak density for O, N and C atoms are better resolved and data quality is sufficient to clearly see the weak anomalous signal of 12 of the 13 ordered S atoms within the asymmetric unit, despite the X-ray energy being very far from the sulfur *K* edge.

2. Materials and methods

2.1. Data collection

SFX experiments were carried out at the Linac Coherent Light Source of the SLAC National Accelerator Laboratory, Menlo Park, California, USA using the CSPAD detector on the CXI instrument to solve the structure of CPV17 at approximately 1.46 Å wavelength (Ginn *et al.*, 2015). Data collected at a later stage of the same visit (18 March 2013) used the same experimental setup and beamline parameters except that the wavelength was decreased to 1.3 Å, allowing a resolution of 1.65 Å at the edge and 1.42 Å in the corner of the detector. The latter data were used in this study. The transmittance of the beam was set to 3%. A total of 66 564 diffraction patterns were recorded, of which 7225 were indexed using the *cctbx.xfel* pipeline. The flow rate of the injector was increased by 33% relative to the data collected at 1.46 Å, leading to an enrichment in images for which multiple crystals were within the beam compared with the previously reported data set.

2.2. Detector geometry

The CSPAD detector comprises 64 pixel-array panels, and is periodically rebuilt. Over the course of an experiment, when the detector is rebuilt or the physical positions of the XFEL components are changed the relative positions and rotations of the panels may change, but it is difficult to accurately measure and incorporate this information into the image metadata at the time of collection. We refer to the correction of these variations as detector geometry, but it is known in *cctbx.xfel* as metrology. *cctbx.xfel* uses a method of metrology correction based on information from multiple images to correct the metrology to subpixel accuracy (Hattne *et al.*, 2014), but in this case the metrology correction was not extracted from the images, and an alternative method to correct for geometry errors is described here.

Integration (see §2.5) was initially carried out over a subset of 500 images. The integration mask was shifted to the local maximum intensity within a 21 × 21 window centred on the initial predicted spot position. We recorded, for each panel, the average shift for every spot on the detector which falls within the coordinates of that panel. These shifts were then plotted, as in Fig. 1(*a*), showing a clear preference for spots to shift in a characteristic way for each panel. Streaking of the plot (Fig. 1*b*) occurs when there are errors in the detector

distance or integration wavelength, which is discussed in §2.3. To calculate the translational shift for each panel, a moving window of a 2 × 2 pixel box scanned the plot as shown in Fig. 1(*a*), with a step size of 0.1 pixels. The coordinates of the window that lies over the largest number of shifts were stored for future use. The rotations of the panel along the *X* and *Y* axes in the plane of the detector were also refined: for the *X* axis the position of a reflection on the panel can be expressed as a fraction of the panel width, and we found that the recorded shift for a reflection is partially dependent on this fraction. If a panel is rotated on the *X* axis, the left-hand side of the panel may be closer to the crystal than the right-hand side. This will manifest by pushing reflections on both the left-hand and the right-hand sides of the panels towards the edges of the panel. *X* and *Y* rotation parameters take into account the extra shift caused by the dependence on the predicted position on the panel along the *X* or *Y* axis. This was calculated empirically by plotting the horizontal axis shift against the horizontal coordinates of the predicted position as a fraction of the panel width and fitting a regression line, the gradient of which is the *X* rotation parameter. This was similarly calculated for the *Y* rotation and also stored for future use. Later integration events then applied the most common shift to each reflection and applied the rotation parameters based on their position on the panel using the geometry-corrected shifts for all spot integration, locking weak reflections to the correct coordinates.

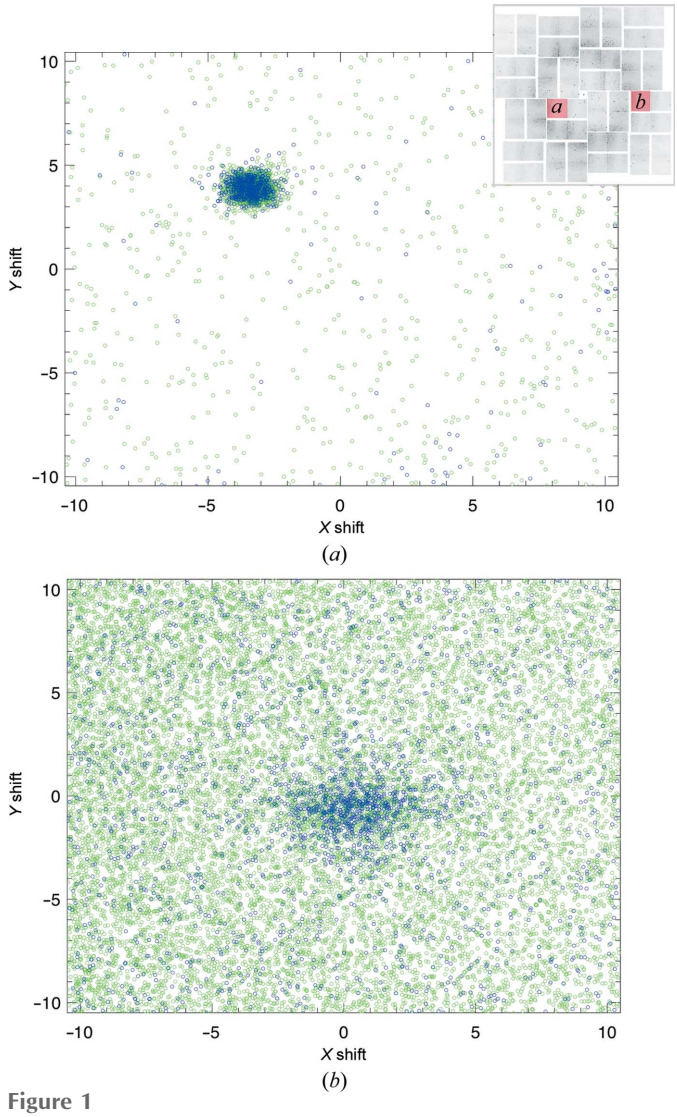


Figure 1
Spot positions were allowed to migrate around a 21 × 21 box centred on the original predicted position. The pixel shift in *X* and *Y* coordinates from the original starting point was retained for each reflection. These pixel shifts were aggregated for each panel. The *X* and *Y* pixel shifts of each reflection on a single panel were plotted against each other, as above. The average intensity counts were calculated for each panel. Green reflections are those below the average, whereas blue reflections are those above the average intensity. (*a*) This panel is the third panel from the left and five panels down from the top of the detector. The most common shift in this panel was easily resolved. The applied shift for this panel was (−3.2, 4.4) pixels in *X*, *Y* coordinates. (*b*) This panel is the second panel from the right and five panels down from the top of the detector, plotted at an incorrect detector distance to show streaking along the *X* axis, but the most common shift could still be resolved and was at (0.2, −0.5) pixels. Because the data were weaker than those in (*a*), more spots incorrectly focused on noise and deviated from the common shift of (0.2, −0.5).

2.3. Refinement of global parameters

The unit-cell length was set to 106.1 Å, which had been determined previously from room-temperature powder diffraction studies (Ginn *et al.*, 2015). Any residual variation in the unit cell was accounted for by refinement of the mean wavelength of each pulse, which is perfectly correlated with the dimensions of a cubic lattice, $a = b = c$.

If the proposed distance from the crystal to the detector was too short, reflections appeared closer to the centre of the image than they actually were, which was reflected in the shift plot. A shift of an individual reflection on a panel was greater if the reflection was furthest away from the centre of the image, whilst lower resolution reflections were less affected by errors in the proposed detector distance. This caused a broadening of shifts along the vector from the centre of the image to the midpoint of a particular panel. This occurred whether the proposed detector distance was too close or too far. Because the streaking of the plot (Fig. 1*b*) was accentuated when the error in detector distance was greater, minimizing this streaking effect could be used to manually adjust the detector distance by eye. The optimum wavelength for integration could then be extracted after initial orientation-matrix refinement (see §2.4). This was an iterative process, with successive changes in detector distance and wavelength progressively improving the panel shift plots.

2.4. Initial orientation-matrix refinement

Initial orientation-matrix refinement was carried out as described previously (Ginn *et al.*, 2015) for each image in order to produce an initial orientation matrix as close as possible to the true matrix without requiring a reference data set. Reflections were classed as strong if they reached an $I/\sigma(I)$ of greater than 12 (rather than using an absolute intensity threshold, as was used previously), but the counting errors were not included in further refinement. Strong reflections

were used in initial refinement of the orientation matrix. An $I/\sigma(I)$ of 12 is indeed a high threshold, but was established after a few manual trials as a value which balances obtaining a large number of strong reflections with excluding noise.

2.5. Integration

Integration of reflections involved summation of foreground pixel photon counts minus background pixel photon counts according to the masks outlined in Fig. 2(*a*) during initial orientation-matrix refinement (see §2.4). Starting spot positions were derived from the *cctbx.xfel* orientation matrices. The midpoint of the mask was initially centred on the highest intensity pixel within a 21 × 21 window, before better detector geometry had been calculated, after which the spots were not allowed to wander but were focused on the most common shift for each individual panel (see §2.2). Counting statistics were disregarded owing to potential interference from the uncertain and nonlinear gain of the detector and we found, empirically, that including counting statistics in the merging process reduced the quality of the final data. Errors in individual reflections were set to unity, divided by the partiality and multiplied by the scale factors for individual images.

After initial orientation-matrix refinement and immediately prior to generating the reflection list which continues into post-refinement, unique integration boxes were generated for individual reflections. The broadening on one axis for a given spot was caused by the broad range of wavelengths, and therefore range of Ewald sphere centres, which will cause illumination of a spot. Spots were stretched along the broadened axis by a number of pixels calculated by ray-tracing from limiting Ewald sphere centres to the detector and finding the resultant pixel shift. Each individual pixel can contribute a proportion between 0 and 1 towards the foreground signal for a given spot. The background reading was taken from a square padded by one pixel away from the ellipse, as shown in Fig. 2(*b*). The mean intensity of the background pixels was calculated and this value was subtracted from each foreground pixel. If any part of the integration shoebox or background pixels fell off the edge of a panel it was excluded from integration. This rejected approximately 30% of reflections.

2.6. Creating an initial data set to seed refinement

For the initial merge, each image was overpredicted by using an artificially wide energy bandwidth of 3.0%, and an initial partiality correction was applied to reflections within this range. The parameters for this partiality correction were set as the pre-determined initial starting values, or in the case of the mean wavelength by choosing the Ewald sphere wavelength at which the largest number of strong reflections were illuminated from the initial orientation-matrix refinement. Approximately 2000 reflections were predicted for each image using this method. These overpredicted reflections were kept in memory in order that they could contribute to future alterations in the orientation matrix during post-refinement. There was an indexing ambiguity in space group *I*23 which was broken by using a modified version of the algorithm of Brehm & Diederichs (2014), as has been described previously (Ginn *et al.*, 2015). Intensities were weighted by the unrefined partiality correction and scale factors to bring the average intensity of each image to 1000 ADU (analogue-to-digital units) before merging.

2.7. Nature of the partiality model

We have used a partiality model based on that described in Ginn *et al.* (2015) with some alterations. Two functions were defined, one describing the profile of the energy spectrum, E, and one describing the profile of the rlp, P. The final partiality ρ for each reflection is defined in (1), where p is the fraction of the rlp along the curve of constant length from the origin, from 0 (lowest wavelength) to 1 (highest wavelength), as defined in Rossmann *et al.* (1979),

$$\rho = \int_0^1 (EP)\, dp. \qquad (1)$$

The function E was defined as a Gaussian distribution in our previous version of the partiality model, but this has been changed to a super-Gaussian model of mean wavelength μ and standard deviation σ with an exponent N (Decker, 1995) set to 1.5, closer in resemblance to a Christmas tree, as shown in (2). λ_p represents the Ewald sphere wavelength of the reciprocal-space coordinate at fraction p along the rlp. μ is the mean wavelength of the beam and σ is the standard deviation of the beam wavelength (the bandwidth of the pulse). The shape of

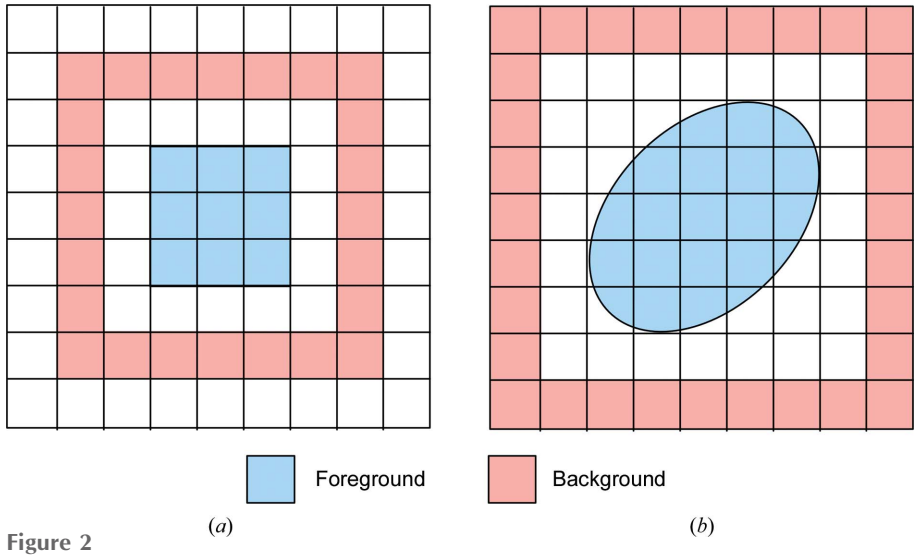


Figure 2
Foreground and background masks used to calculate the integrated signal of each reflection, using the simple shoebox during initial orientation-matrix refinement (*a*) and an elliptical shoebox based on the energy bandwidth (*b*).

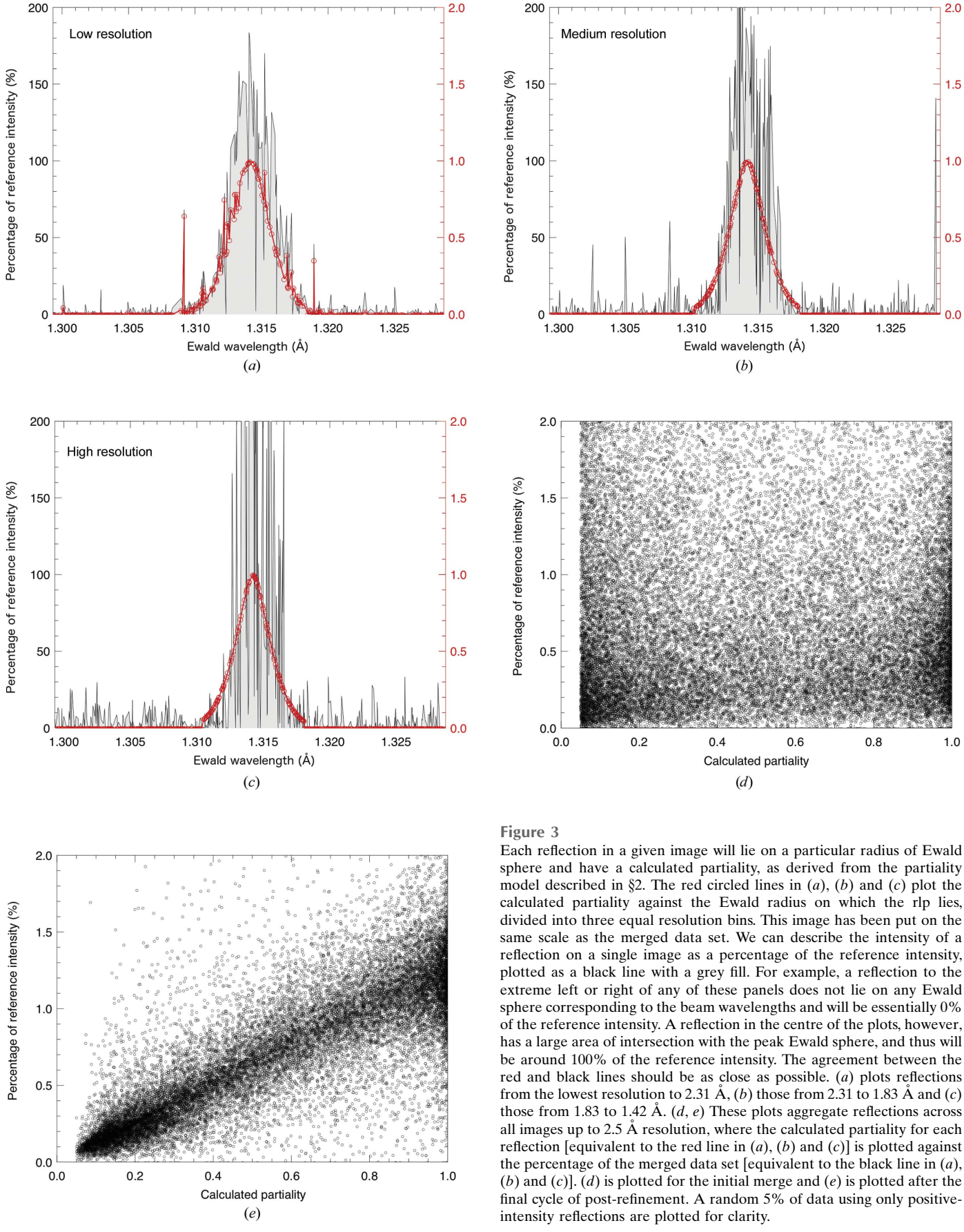


Figure 3
Each reflection in a given image will lie on a particular radius of Ewald sphere and have a calculated partiality, as derived from the partiality model described in §2. The red circled lines in (*a*), (*b*) and (*c*) plot the calculated partiality against the Ewald radius on which the rlp lies, divided into three equal resolution bins. This image has been put on the same scale as the merged data set. We can describe the intensity of a reflection on a single image as a percentage of the reference intensity, plotted as a black line with a grey fill. For example, a reflection to the extreme left or right of any of these panels does not lie on any Ewald sphere corresponding to the beam wavelengths and will be essentially 0% of the reference intensity. A reflection in the centre of the plots, however, has a large area of intersection with the peak Ewald sphere, and thus will be around 100% of the reference intensity. The agreement between the red and black lines should be as close as possible. (*a*) plots reflections from the lowest resolution to 2.31 Å, (*b*) those from 2.31 to 1.83 Å and (*c*) those from 1.83 to 1.42 Å. (*d*, *e*) These plots aggregate reflections across all images up to 2.5 Å resolution, where the calculated partiality for each reflection [equivalent to the red line in (*a*), (*b*) and (*c*)] is plotted against the percentage of the merged data set [equivalent to the black line in (*a*), (*b*) and (*c*)]. (*d*) is plotted for the initial merge and (*e*) is plotted after the final cycle of post-refinement. A random 5% of data using only positive-intensity reflections are plotted for clarity.

this function can be seen in Figs. 3(*a*), 3(*b*) and 3(*c*), where the higher resolution panels (Figs. 3*b* and 3*c*) show minimal interference from the function *P*,

$$E = \frac{1}{k(2\pi\sigma)^{1/2}} \exp\left(\frac{-|\lambda_p - \mu|^N}{2\sigma^N}\right). \quad (2)$$

The scale factor k is the maximum achievable partiality for a reflection of the given rlp radius, if it were centred in the middle of the beam. This normalized partialities and made them comparable between resolution shells so they always fell between 0 and 1, and prevented per-cycle inflation of high-resolution reflection intensities during post-refinement. No Lorentz correction was applied to the data, and as a result the normalization applied above led to higher resolution reflections being somewhat underestimated, which was approximately corrected for by applying a *B* factor to the data. The function *P* describes the cross-section of the rlp as a proportion of the maximum cross-sectional area as in (3),

$$P = 4p(1-p). \quad (3)$$

Intensities and errors were divided by the partiality to inflate their values to an estimate of the true value. A minimum partiality cutoff of 0.05 was defined, and intensities with a partiality of less than 0.05 were not included in refinement. Applying this partiality model resulted in an average of 280 accepted reflections above the cutoff being merged for the images which refined correctly. Below a partiality of 0.05 the contribution of each reflection was so low, as the reflections were weighted by their partiality, that their contribution to the final data set would be negligible.

2.8. Refining individual images

There were four parameters to refine per individual image: two rotation angles, which allow correction of the orientation matrix along two axes perpendicular to the beam and to each other, the rlp size (governed by the size of the crystal exposed in the beam) and the mean wavelength of the SASE pulse. The target function was defined as the *R* factor between the image and the reference, which we aimed to minimize by altering the values of these parameters. Initial starting values were taken to be 0° for the two rotation angles and 1×10^{-4} Å^{-1} for the rlp size, which is equivalent to a 1 µm crystal size. The initial wavelength was taken as the mean of all Ewald sphere wavelengths where the intensity count was greater than 200 ADU (analogue-to-digital units).

The rlp size and mean wavelength parameters were altered repeatedly by testing the current value and the current value plus or minus a defined step length. The new value for the parameter was taken as the value corresponding to the lowest value of the target function. If the value remained unchanged from the previous alteration, the step length was divided by two. The rotation angles were altered in a combined, two-dimensional grid search, where nine possible values were tested and both step lengths were reduced simultaneously. The convergence criteria were 1×10^{-4} degrees for rotation angles, 1×10^{-5} Å for the mean wavelength and 1×10^{-5} Å^{-1} for the rlp size. The bandwidth was set to 0.26% (the spread of four standard deviations), equivalent to 25 eV for the experimental wavelength used. Mosaicity was set to zero, as it is very low for these crystals (from analysis of synchrotron data processing; Gildea *et al.*, 2014) and separating the effects of bandwidth expansion and increased mosaicity is difficult when the mosaicity parameter is so low. In this case, the effect of mosaicity was subsumed in the bandwidth effect. Correlation using each indexing choice was checked in order to make sure images were not mis-indexed during initial separation of indexing choices, and images were corrected if necessary.

2.9. Rejecting outliers

Outlier rejection occurred in both the individual image-refinement stages and at the merging stages. For images with a correlation weaker than 99% with the reference data set, up to three reflections per cycle were rejected if the removal of these reflections caused an upwards shift in the correlation coefficient CC, where the value of (1 − CC) decreased by more than 6% (for example, if a correlation coefficient of 90.0% increases to 90.6% or more). Once reflections had been rejected by this method, they were not reintroduced into the system until the final merge, in which the rejected reflections were recalculated afresh. This resulted in an average of 21.5 rejected reflections per image over the course of the full refinement process. The second stage of outlier rejection occurred just prior to merging. The mean and σ for the intensity of each reflection were calculated from the independent observations weighted by the individual partialities and scale factors. A rejection cutoff of 1.8σ was chosen in order to reject less than 10% of reflections, assuming a Gaussian distribution of observations per unique reflection. Reflections which were more than 1.8σ away from the mean were rejected, but these could be reincluded in subsequent rounds of refinement.

2.10. Merging

Observations were reduced to the unique index for the asymmetric unit of space group *I*23. All images above a certain threshold of correlation with the reference data set were included in the final merge for each macrocycle. This threshold was set to exclude images which failed to refine correctly, in this case 0.9 for all merges apart from the final merge, which was made more stringent (this was set to 0.95). Images were also rejected if the final number of reflections above the partiality cutoff of 0.05 was 100 or fewer. Scale factors were generated for each individual image. Friedel pairs were not separated for the purposes of generating scale factors for each image nor for outlier rejection, as the anomalous differences were considered to be negligible. Friedel pairs were maintained separate only to generate anomalous data on the final merge. The shared reflections between the individual image and the reference data set were plotted, and the best-fit gradient forced through the origin was calculated. The scale factor was set to the reciprocal of this gradient, so that following application of this scale factor the gradient would be

recalculated as unity. For each unique reflection, individual intensity observations were corrected for partiality and the mean and σ were calculated, excluding rejected reflections. The σ value is calculated from the distribution of observations for a given reflection.

Refinement was allowed to continue for at least seven cycles until the R factor between the most recent reference data set and the previous reference data set was 1% or less. After finishing refinement, the data were merged three times using scale factors generated using the previous best merge. The last merge was taken as the final data set.

2.11. Generation of anomalous difference Fourier map

In the case of merging to preserve the anomalous data, which occurred at the final merge, the rejection criteria were set to more stringent values at 1.0σ as these improved the spherical shape of the anomalous difference intensity for these atoms. The resolution for anomalous signal was cut arbitrarily to 1.8 Å. Anomalous difference Fourier maps were generated using phase estimates from *ANODE* (Thorn & Sheldrick, 2011).

3. Results

cctbx.xfel (Sauter *et al.*, 2013; Hattne *et al.*, 2014) identified 7225 hits from a 9 min run of data collection from CPV17 polyhedrin crystals on the CXI instrument at the LCLS (Emma *et al.*, 2010). These data were collected with pulses of X-ray wavelength 1.3 Å, which allowed spots to be integrated up to 1.42 Å resolution in the corner of the detector. Initial orientation matrices were generated by *cctbx.xfel* and were refined according to the protocols defined in Ginn *et al.* (2015). There was no uneven sampling of reciprocal space detected (*Appendix A*). Integration and resolution of indexing ambiguity was followed by a post-refinement algorithm to remove residual errors in the orientation matrix and refinable experimental parameters. The post-refinement algorithm comprises macrocycles, each composed of refinement against a reference data set followed by merging of a new reference data set, as outlined in §2. The overarching program architecture is illustrated in Fig. 4.

3.1. Integration boxes

Owing to the highly reproducible nature of CPV17 crystals, the unit cells were fixed to dimensions $a = b = c = 106.1$ Å, allowing the detector distance to be easily refined manually. Changing the shoebox to an ellipse decreased the highest resolution (1.485–1.460 Å) $R_{\rm split}$ from 51.4 to 42.5%. By fixing the unit cell to 106.1 Å and refining only the detector distance and wavelength this was further reduced to 40.6%.

3.2. Detector geometry

The diffraction images were not automatically corrected for detector-geometry errors, and our previous method corrected for geometry errors by centring on a local maximum photon count. This was a significant contribution to high $R_{\rm split}$ values at high resolution, as it overestimated the weak data. It was also found that by increasing the flow rate of the injector the number of multiple lattice images was increased compared with the data used in Ginn *et al.* (2015); this in turn increased the danger that a spot allowed to 'wander' would focus on an incorrect local maximum such as noise or a neighbouring lattice point. We applied a new type of geometry correction, defined in §2.2, and found that this increased the quality of the data. Without geometry correction, the $R_{\rm split}$ over all resolution shells is 3.53%, whereas with geometry correction this decreases to 3.15%.

3.3. Initial merge

The initial merge was carried out by applying the partiality model, in the absence of any post-refinement of parameters, to each individual image. The $R_{\rm split}$ for the initial merge was 9.20% and the $CC_{1/2}$ was 0.986. These values are similar to the statistics for our earlier 1.75 Å resolution structure, where 5787 crystals were merged to generate an $R_{\rm split}$ of 11.7%, when the small increase in image number is taken into account. However, this included the rejection of 69.3 reflections per

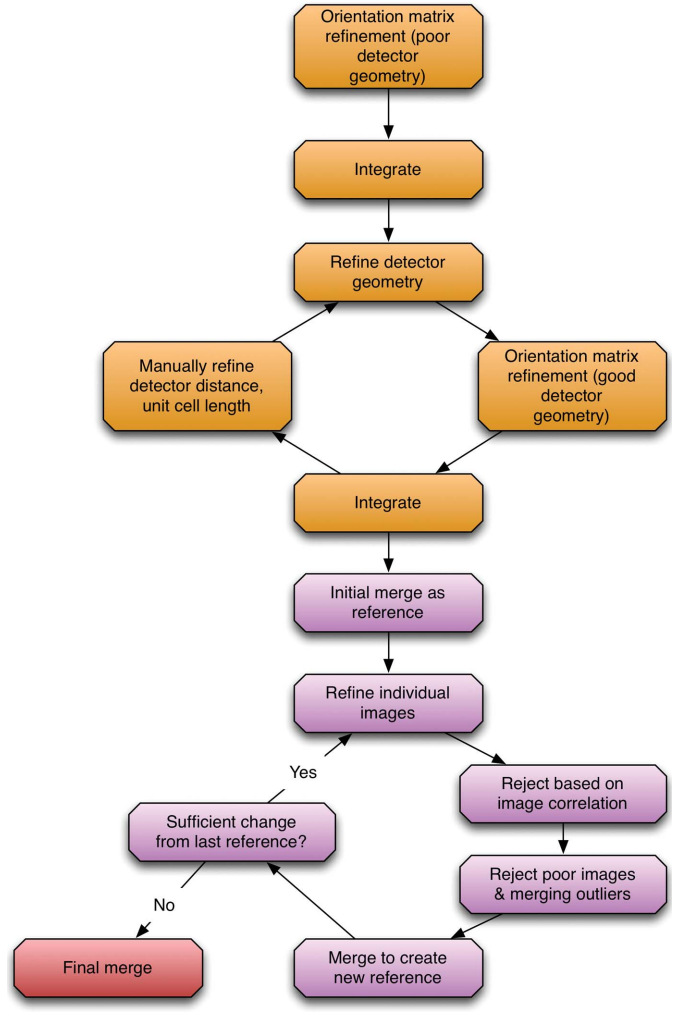


Figure 4
Diagram showing the flow of software during the post-refinement of XFEL data.

Table 1
Calculated $CC_{1/2}$ and R_{split} for each cycle of refinement.

The final merge uses a correlation threshold of 0.95 and fully recalculates rejected reflections.

Cycle No.	$CC_{1/2}$	R_{split} (%)	Rejected reflections per image
Initial merge	0.9860	9.20	69.3
Cycle 1	0.9987	4.76	22.2
Cycle 2	0.9989	4.04	22.9
Cycle 3	0.9990	3.85	23.6
Cycle 4	0.9990	3.78	24.1
Cycle 5	0.9991	3.73	24.4
Cycle 6	0.9990	3.71	24.6
Cycle 7	0.9991	3.70	24.8
Final merge	0.9995	3.15	21.5

image, indicative of the unrefined models producing large numbers of outliers. The major differences between these two processing strategies were the super-Gaussian modelling of the bandwidth, compared with a Gaussian model for the 1.75 Å resolution structure, and the fixing of the energy spread in the 1.46 Å resolution structure, compared with a variable parameter in the previous processing. This initial merge provided a good starting reference against which the data set was refined, even for low numbers of images (see §3.8).

3.4. Refining individual images

The rotation angles, rlp size and mean X-ray wavelength were refined according to §2.8. The average rotation in the orientation matrix was 0.026°. A successfully refined image had a good match between the predicted partiality and the integrated counts as a percentage of the reference data set, which was considered to be a measure proportional to the true partiality of the reflection. This is shown in Fig. 3. The super-Gaussian exponent of 1.5 was chosen on the basis that it became clear from plots such as Fig. 3 that there was some variation in the shape of the energy bandwidth which was not accounted for by an exponent of 2. After refining these parameters, the calculated partiality can be seen to agree well with the estimate of the true partiality. In the higher resolution shells the quality of the data was acutely dependent on the orientation matrix, detector-geometry correction and spot integration being as accurate as possible. The combination of these effects and the weaker data explains the diminishing $CC_{1/2}$ and increasing R_{split} at high resolution.

3.5. Merging after each cycle

In the first cycle after the initial merge, 4.5% of images had their indexing choice corrected. During the final merge, 9.5% of images were rejected owing to individual image-exclusion criteria. For the images which were included in the final merge, the average correlation coefficient with the reference was 0.988. Rejection of individual reflections based on individual image correlation alone rejected an average of 6.3 reflections per image, whereas rejections based on merging statistics rejected 20.9 reflections prior to the final merge. Inspection of individual diffraction patterns showed that these rejections were largely a result of overlapping or close reflections from multiple crystals in the beam. Although excluded reflections were recalculated on each merge, these were also excluded from individual images in the next round of image refinement. To calculate the progress of refinement, $CC_{1/2}$ and R_{split} were calculated on each merge and these values converged in seven cycles, as shown in Table 1, to a final R_{split} of 3.15%.

Table 2
Crystallographic refinement data for new processing of CPV17.

Values in parentheses are for the highest resolution shell or are standard deviations (s.d.) where specified.

Total diffraction patterns	65564
No. of indexed diffraction patterns	7225
No. of patterns used in final merge	6537
Space group	*I*23
Unit-cell parameter (Å)	$a = b = c = 106.1$
Resolution (Å)	25.0–1.46 (1.485–1.460)
Completeness (%)	99.5 (92.8)
Multiplicity	45.4 (9.14)
Unique reflections	34369
Total observations	1830360
Reflections per image	280
No. of reflections rejected/mean No. per image	8.9/21.5
Mean wavelength (s.d.) (Å)	1.3150 (0.00211)
Mean crystal dimension (s.d.) (µm)	1.22 (0.43)
Mean rotation correction (s.d.) (°)	0.026 (0.025)
R_{split} (%)	3.15 (40.6)
$CC_{1/2}$	0.9993 (0.3311)
R_{meas} (%)	19.0 (71.4)
$R_{p.i.m.}$ (%)	1.13 (7.78)
No. of atoms	2243
Protein residues (total/observable)	237/236
R_{work}/R_{free} (%)	11.1/15.8

3.6. Properties of the final data set

Including post-refinement produced a much stronger processing algorithm than that used for the previous solution of CPV17. The initial merge, without post-refinement, achieved an R_{split} of 9.20%, as shown in Table 1, albeit with a large number of rejections. After post-refinement, the R_{split} value was 3.15%, which means that each image has a 12-fold improvement in lowering the R_{split} value, and the number of rejected reflections was reduced. This increase is the squared ratio between the prior R_{split} and the best R_{split} when corrected for number of images, assuming that R_{split} is largely proportional to N, where N is the number of images. Crystallographic refinement statistics for this data set are shown in Table 2. The final resolution cutoff was chosen as the highest resolution shell for which $CC_{1/2}$ was greater than 0.3 (1.46 Å). The increase in reflection number allowed anisotropic *B*-factor refinement, resulting in crystallographic *R* factors of R_{work} = 11.1% and R_{free} = 15.8% using *PHENIX* (Adams *et al.*, 2010).

The improved quality of the high-resolution information can be seen in the R_{work} and R_{free} values for the high-resolution shell. In the post-refined data set, the highest resolution shell, 1.485–1.46 Å, had an R_{work} and R_{free} of 29.0 and 33.9%, respectively, whereas the same shell in the non-post-refined data (cycle 0) had an R_{work} and R_{free} of 34.8 and 40.8%, respectively, and the highest resolution shell for the previous 1.75 Å resolution structure solution, 1.81–1.75 Å, had an R_{work} and R_{free} of 43 and 46%, respectively. Thus, the added

quality of the data was strongly reflected in the markedly improved model-refinement statistics.

3.7. Electron-density maps

Electron-density maps with *B*-factor sharpening for the post-refined 1.46 Å resolution data set clearly showed a marked improvement over the 1.75 Å resolution structure and over the 1.46 Å resolution data set before post-refinement (the initial merge), with visibly distinguishable atomic numbers of C, N and O atoms. The peak density for O and N atoms was sufficiently different to resolve the conformations of glutamine and asparagine residues. The difference in density is shown in Fig. 5. Extra water molecules, as well as alternative conformations in the main chain, were added to the atomic model.

This data set was collected using a 1.3 Å wavelength beam, which was very far from the sulfur *K* edge of 5 Å, and would result in an f'' of 0.42 electrons per S atom. This would produce an anomalous signal of an average 0.5% change in amplitude, compared with a background intensity $R_{\rm split}$ of 3.15%. Nevertheless, an anomalous difference Fourier map calculated on the post-refined data clearly highlighted 12 of the 13 ordered S atoms present in the asymmetric unit of the crystal, as shown in Fig. 6. The only missing sulfur anomalous density was for Cys142 (not shown), which was in two alternative conformers each with half occupancy. One conformation formed a disulfide bond with Cys155. The mean anomalous peak for the 12 anomalous S atoms was 3.87σ, with a maximum of 5.43σ for Met182. The anomalous difference map was generated using *ANODE* (Thorn & Sheldrick, 2011) to produce cleaner maps, although using a basic estimate of the native phase shifted by 90° was also sufficient to observe all of these peaks. Seven of the 12 sulfur anomalous densities were at least 4.2σ at the peak and within half an angstrom of the corresponding S atom.

This anomalous density was compared with the anomalous density achieved by the initial merge with no application of post-refinement. In this case there was barely any density associated with methionine S atoms. The closest peaks to S atoms occurred 0.7–2.8 Å away from the sulfurs at an average peak density of 2.62σ, and the highest peak associated with sulfur was the 244th highest peak in the entire electron-density map. Hence, these were lost within noise and were not specifically associated with S atoms, whereas the application of post-refinement reduced the errors sufficiently to observe the anomalous density, as demonstrated above.

3.8. Effect of the number of images

In order to observe the effect of number of images on $R_{\rm split}$, the above procedure was repeated on subsets of 200 to 4000 images (see Fig. 7). Although the low-image-number subsets were hampered by poor initial reference data sets, they still recovered a substantial amount of information. Using the Monte Carlo relationship $R_{\rm split} \propto 1/N^{1/2}$ (Kirian *et al.*, 2010) and the $R_{\rm split}$ of 24.2% for 500 images, we would expect an $R_{\rm split}$ of 6.36% for 7225 images. However, owing to the post-refinement method using information across many images to improve the parameters for individual images, the $R_{\rm split}$

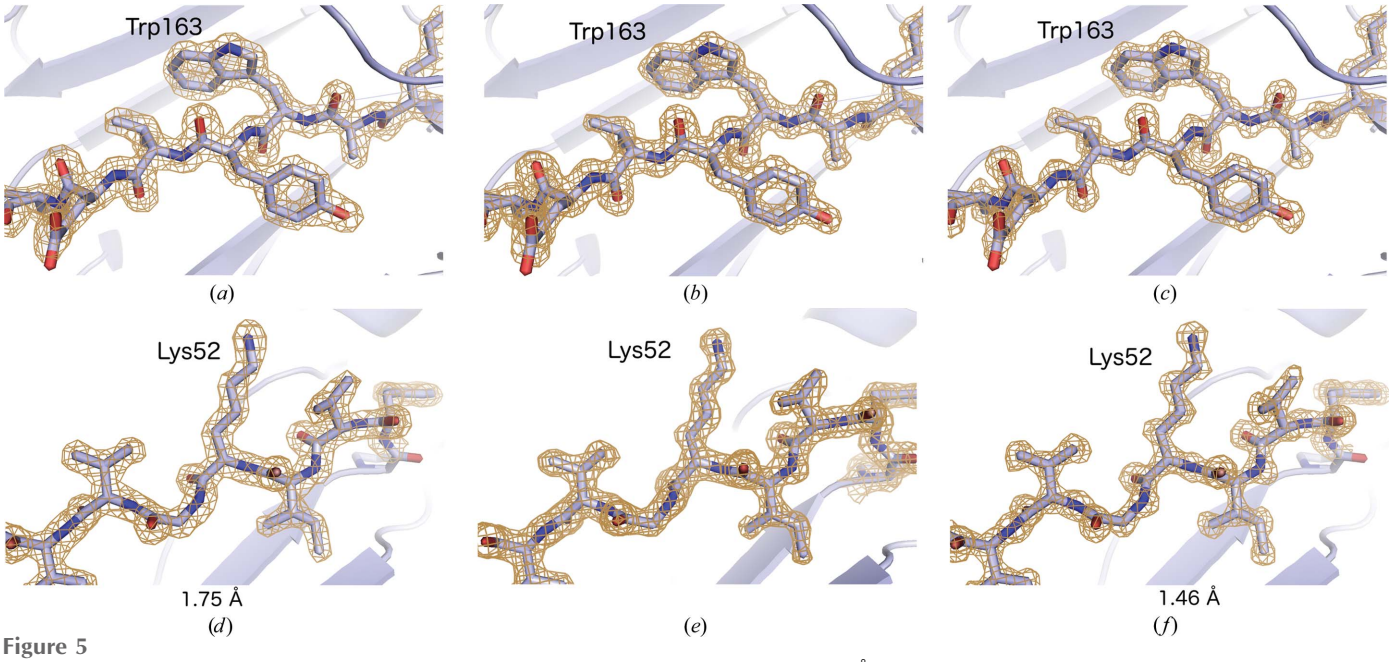


Figure 5
Electron-density map ($2mF_{\rm o} - DF_{\rm c}$) from reflections associated with PDB entry 4s1l to 1.75 Å resolution (Ginn *et al.*, 2015) associated with Trp163 (*a*) and Lys52 (*d*), their corresponding electron density in the initial merge for the higher resolution data set, (*b*) and (*e*), and the final presented 1.46 Å resolution structure, (*c*) and (*f*), at a σ of 1.5. The Lys52 side chain shows the prominence of the H atoms on the methylene groups, which are not pronounced on the perpendicular profile. The N, C and O atoms are distinguishable compared with the 1.75 Å resolution structure. The high-resolution information beyond 1.75 Å and post-refinement of this data set appear to have separate sequential improvements on the quality of the electron density compared with the 4s1l structure.

achieved was actually 3.70% (prior to the final merge), thereby recovering information beyond the effects of simply adding more data to reduce nonsystematic errors and essentially producing a data quality equivalent to what might be expected from a data set three times the size.

4. Discussion

The choice of an appropriately shaped integration box markedly improved the high-resolution information for this set of images when analysed using the methodology described in Ginn *et al.* (2015), but errors remained at high resolution, most likely owing to residual errors in the detector geometry and the comparative weakness of the data, as well as subtle aberrations enhanced at high resolution caused by errors in detector distance and wavelength. Nonetheless, the resolution of the analysis was limited by the geometry of the detector. Outlier rejection was instrumental in lowering the $R_{\rm split}$ owing to the presence of multiple lattices on a large number of images. In order to reject more rigorously, the multiple lattices on a given image should all be indexed, as has been performed to generate a synchrotron structure for crystals of this polyhedrin (Gildea *et al.*, 2014) and as has been carried out on images of thermolysin (Hattne *et al.*, 2014), to identify neighbouring or overlapping reflections from separate lattices. Processing multiple lattices would also increase the number of observations for assembling the data set.

Overall, the post-refinement algorithm has resulted in a marked improvement in the extraction of information from the diffraction patterns. The electron-density maps were significantly improved and the weak anomalous signal from the S atoms was revealed, despite the experiment being performed far from the optimal wavelength. We hope that this will lead the way to allowing routine crystal structure solution by sulfur SAD on serial femtosecond lasers, especially when data are collected at an optimum wavelength for sulfur SAD

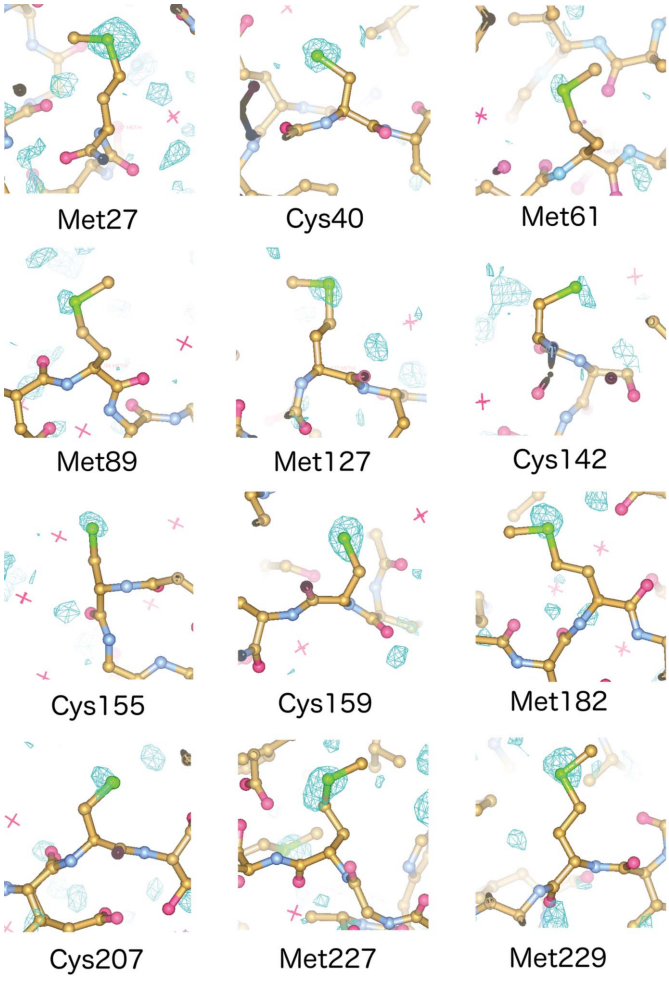


Figure 6
Anomalous signal contoured at 2.4σ highlighted for 12 S atoms as labelled. The highest density is 5.43σ (Met182) and the lowest is 2.53σ (Met70 and Met89). The average peak σ is 3.87.

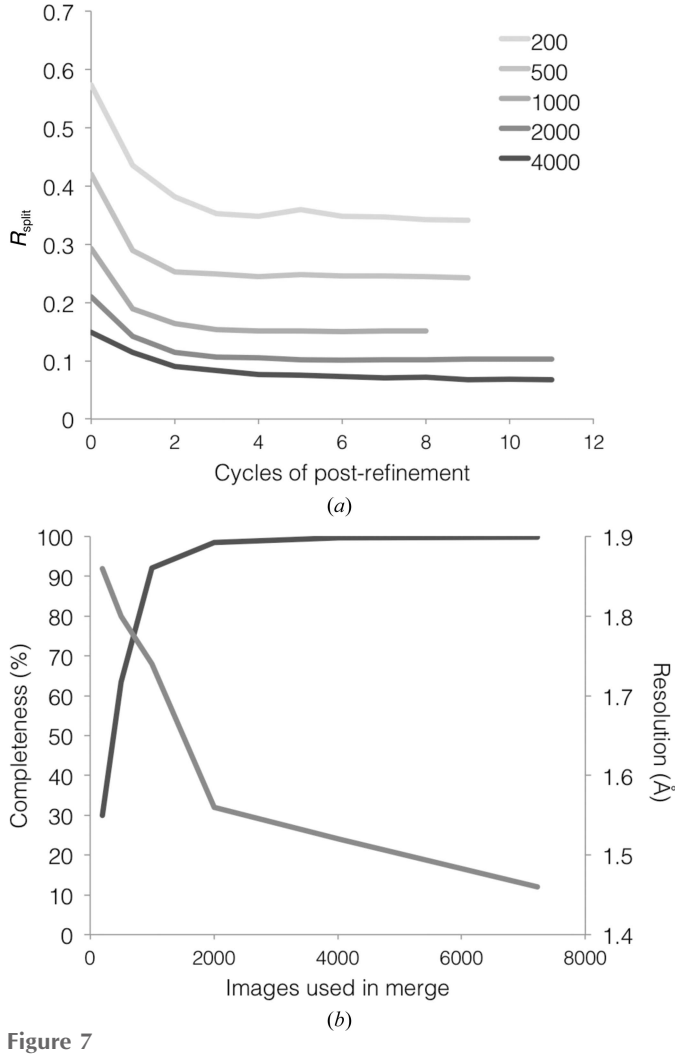


Figure 7
(*a*) Calculated $R_{\rm split}$ values for subsets of images between 200 and 4000 in number, using a final correlation merge threshold of 0.9. (*b*) Maximum resolution (light grey line) denotes the first resolution shell beyond which $CC_{1/2}$ falls below 0.3. Completeness (dark grey line) is calculated from low resolution to the resolution cutoff. This suggests that even 1000 crystals will give a useful data set comparable to that reported for data collected at a synchrotron from a similar number of crystals (Gildea *et al.*, 2014).

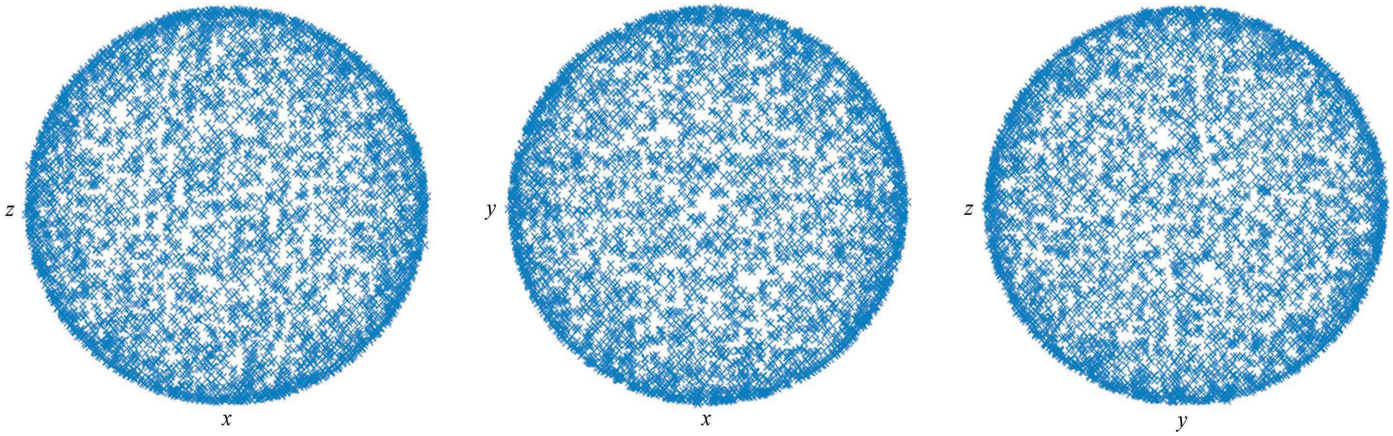


Figure 8
Orientation matrices for 1500 crystals were applied to a (1, 1, 1) coordinate and the point was plotted against the corresponding perpendicular axes, including symmetry-related points owing to cubic space-group symmetry. This shows an even sampling of crystal orientations.

studies. By markedly reducing the number of diffraction images required to provide a high-quality data set, such methods should also open up the method to more challenging problems and increase the efficiency with which the scarce resource of XFEL beam time can be used. We are aiming shortly to fold the code written for this purpose into *cctbx.xfel* and *DIALS* (Waterman *et al.*, 2013).

Note added in proof: An alternate approach to post-refinement of X-ray free-electron laser data has recently been published by Uervirojnangkoorn *et al.* (2015).

APPENDIX *A*
Crystal orientation

Analysis of the orientation matrices for 1500 crystals showed no preferred orientations, as shown in Fig. 8.

Acknowledgements

DIS was supported by the Medical Research Council, grant G1000099. HMG was supported by the Wellcome Trust (studentship 075491/04). ASB, JH and NKS were supported by US National Institutes of Health grants GM095887 and GM102520. 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. We are very grateful for the expert support for the operation of the sample injector provided by the group of Ilme Schlichting (Max Planck Institute for Medical Research, Heidelberg, Germany), in particular Sabine Botha, R. Bruce Doak and Robert L. Shoeman. We are very grateful to the LCLS–CXI staff, Marc Messerschmidt, Sébastien Boutet, Garth Williams and Dan Deponte. Admin support was received from the Wellcome Trust, grant 090532/Z/09/Z. This is a contribution from the Oxford Instruct Centre. Code is available on request, and will be folded into *cctbx.xfel* and *DIALS*.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Barends, T. R. *et al.* (2013). *Acta Cryst.* D**69**, 838–842.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Brehm, W. & Diederichs, K. (2014). *Acta Cryst.* D**70**, 101–109.
Decker, F.-J. (1995). *AIP Conf. Proc.* **333**, 550.
Emma, P. *et al.* (2010). *Nature Photonics*, **4**, 641–647.
Gildea, R. J., Waterman, D. G., Parkhurst, J. M., Axford, D., Sutton, G., Stuart, D. I., Sauter, N. K., Evans, G. & Winter, G. (2014). *Acta Cryst.* D**70**, 2652–2666.
Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Nature Commun.* **6**, 6435.
Hattne, J. *et al.* (2014). *Nature Methods*, **11**, 545–548.
Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.
Kern, J. *et al.* (2013). *Science*, **340**, 491–495.
Kern, J. *et al.* (2014). *Nature Commun*, **5**, 4371.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C., 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.
Liu, W. *et al.* (2013). *Science*, **342**, 1521–1524.
Redecke, L. *et al.* (2013). *Science*, **339**, 227–230.
Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.
Sauter, N. K. (2015). *J. Synchrotron Rad.* **22**, 239–248.
Sauter, N. K., Hattne, J., Brewster, A. S., Echols, N., Zwart, P. H. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 3299–3309.
Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. (2013). *Acta Cryst.* D**69**, 1274–1282.
Tenboer, J. *et al.* (2014). *Science*, **346**, 1242–1246.
Thorn, A. & Sheldrick, G. M. (2011). *J. Appl. Cryst.* **44**, 1285–1287.
Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y., Hattne, J., Brewster, A. S., Sauter, N. K., Brunger, A. T. & Weis, W. I. (2015). *eLife*, **4**, e05421.
Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K. & Evans, G. (2013). *CCP4 Newsl. Protein Crystallogr.* **49**, 16–19.
White, T. A. (2014). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **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.
Winkler, F. K., Schutt, C. E. & Harrison, S. C. (1979). *Acta Cryst.* A**35**, 901–911.