Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers

Johan Hattne[1], Nathaniel Echols[1], Rosalie Tran[1], Jan Kern[1], Richard J Gildea[1,10], Aaron S Brewster[1], Roberto Alonso-Mori[2], Carina Glöckner[3], Julia Hellmich[3], Hartawan Laksmono[4], Raymond G Sierra[4], Benedikt Lassalle-Kaiser[1], Alyssa Lampe[1], Guangye Han[1], Sheraz Gul[1], Dörte DiFiore[3], Despina Milathianaki[2], Alan R Fry[2], Alan Miahnahri[2], William E White[2], Donald W Schafer[2], M Marvin Seibert[2], Jason E Koglin[2], Dimosthenis Sokaras[5], Tsu-Chien Weng[5], Jonas Sellberg[5,6], Matthew J Latimer[5], Pieter Glatzel[7], Petrus H Zwart[1], Ralf W Grosse-Kunstleve[1], Michael J Bogan[2,4], Marc Messerschmidt[2], Garth J Williams[2], Sébastien Boutet[2], Johannes Messinger[8], Athina Zouni[3,9], Junko Yano[1], Uwe Bergmann[2], Vittal K Yachandra[1], Paul D Adams[1] & Nicholas K Sauter[1]

X-ray free-electron laser (XFEL) sources enable the use of crystallography to solve three-dimensional macromolecular structures under native conditions and without radiation damage. Results to date, however, have been limited by the challenge of deriving accurate Bragg intensities from a heterogeneous population of microcrystals, while at the same time modeling the X-ray spectrum and detector geometry. Here we present a computational approach designed to extract meaningful high-resolution signals from fewer diffraction measurements.

The ~40-fs XFEL pulse can deliver diffraction information on time scales that outrun radiation damage, which allows studies of macromolecular reaction dynamics under functional physiological conditions[1–3], and the small beam focus size permits investigation of extremely small and weakly diffracting microcrystals[4–6]. Unlike the case with single-crystal X-ray diffraction experiments performed at conventional synchrotron radiation (SR) sources, in XFEL studies the sample is destroyed with a single pulse. This requires the full data set to be assembled from a series of still diffraction shots of individual microcrystals, a technique known as serial femtosecond crystallography (SFX).

As with conventional crystallography, the objective of SFX is to obtain a complete set of structure-factor amplitudes through the measurement of Bragg spot intensities (coherent scattering of X-rays described by Bragg's law) to as high a diffraction angle as possible. The high-resolution signal is ultimately limited by noise, and the background (e.g., from solvent) often dominates the diffraction pattern for all but the most intense low-resolution (low-angle) Bragg spots[7]. At SR sources, accurate sampling of the diffraction at the limit of detectability is accomplished by optimally modeling the diffraction experiment, including the relationship between real space (the crystal) and reciprocal space (the diffracted X-ray collected on the detector). The most intense Bragg spots are used to deduce the best-fitting lattice model (indexing), which is then used to predict exactly which pixels on each image to examine for Bragg spot integration, even though a signal may not be visually discernible from background. The same fundamental approach is applicable to the analysis of XFEL data. Here we describe such a data-processing approach for XFEL data, which enables weak signals to be measured from many fewer crystal specimens than possible with the previously available program CrystFEL[8]. We have added the method to our open-source software suite, the cctbx.xfel component of the computational crystallography toolbox[9]. A primer and tutorial are available at http://cci.lbl.gov/xfel, with code archived at http://cctbx.sf.net.

We tested our method for processing SFX diffraction patterns against data collected at the coherent X-ray imaging (CXI) instrument of the LCLS, using the Cornell-SLAC pixel array detector (CSPAD). We derived a structural model for the metalloprotein thermolysin (**Fig. 1a**, and **Supplementary Tables 1** and **2**) that was comparable in quality to structures determined by conventional SR X-ray diffraction at a similar resolution of 2.1 Å. The electron density of the native calcium and zinc ions (omitted from the phasing model) in the difference map (**Fig. 1b**,**c**) indicates that the metal positions were determined by the processed data and are not the result of bias from the phasing model. We also reprocessed 1.9 Å–resolution lysozyme data[10] (**Supplementary Tables 1** and **3**)

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [2]Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, Menlo Park, California, USA. [3]Max-Volmer-Laboratorium für Biophysikalische Chemie, Technische Universität, Berlin, Germany. [4]Stanford PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, California, USA. [5]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, California, USA. [6]Department of Physics, AlbaNova, Stockholm University, Stockholm, Sweden. [7]European Synchrotron Radiation Facility, Grenoble, France. [8]Institutionen för Kemi, Kemiskt Biologiskt Centrum, Umeå Universitet, Umeå, Sweden. [9]Institut für Biologie, Humboldt Universität zu Berlin, Berlin, Germany. [10]Present address: Diamond Light Source, Harwell Science and Innovation Campus, Didcot, Oxfordshire, UK. Correspondence should be addressed to N.K.S. (nksauter@lbl.gov).

RECEIVED 4 AUGUST 2013; ACCEPTED 4 FEBRUARY 2014; PUBLISHED ONLINE 16 MARCH 2014; DOI:10.1038/NMETH.2887

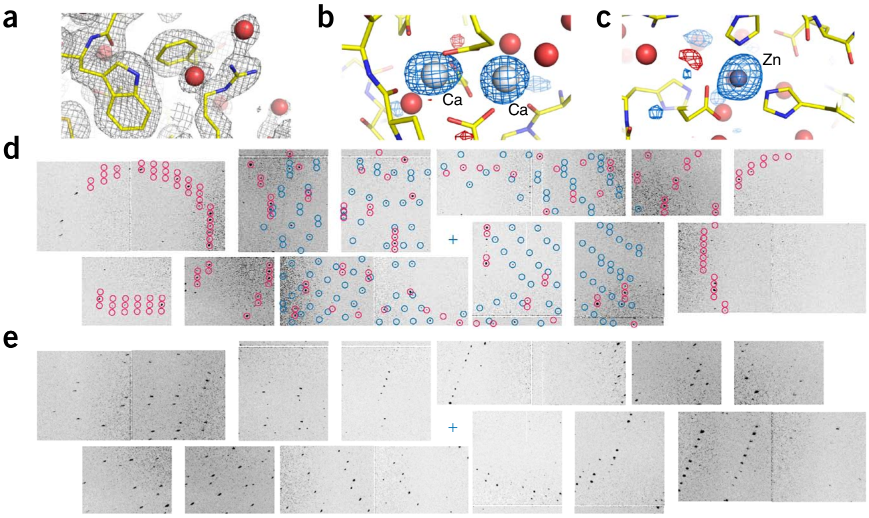

Figure 1 | Thermolysin structure determination at 2.1 Å resolution. (**a**) Likelihood-weighted electron density map calculated with coefficients $2mF_o - DF_c$, where F_o and F_c are the observed and modeled structure factor amplitudes, m is the figure of merit and D is derived from coordinate-error estimates, contoured at 1 s.d. (gray mesh). Water molecules are shown as red spheres. (**b**,**c**) $mF_o - DF_c$ difference density map contoured at +3 s.d. (blue mesh) and −3 s.d. (red mesh), which shows binding sites for two of the four calcium ions (**b**) and the single zinc ion (**c**). (**d**) Detail of two crystal lattices found on the same diffraction image. Modeled spot positions assigned to the different lattices are shown in red and blue, respectively. The sample-detector distance of 135 mm corresponds to a resolution of 2.15 Å at the edges. (**e**) Detail from a different diffraction image. Increasing radial spot elongation was observed with distance from the beam center (blue cross).

previously processed with the software suite CrystFEL[8], to compare the two programs.

We found that cctbx.xfel processed about twice as many diffraction lattices from individual crystals as has been reported for CrystFEL[10] (**Supplementary Table 1**). The indexing algorithm[11], which identifies unit-cell dimensions and crystal orientations, searches for directional vectors that describe the observed rows of Bragg spots, from which three are chosen to form the unit cell. Several factors make this a difficult problem. First, the CSPAD detector consists of 64 pixel array readouts (**Fig. 1d**,**e**) that are periodically disassembled. Thus, the metrology (the relative positions and orientations) of the readouts must be redetermined with sufficient accuracy (**Fig. 2a**), as even small subpixel offsets can diminish the number of images from which lattices can be indexed (**Fig. 2b**). Second, the destruction of each crystal after one XFEL shot removes the ability to view the diffracted lattice from various directions, hindering the selection of unit-cell basis vectors. To compensate, we supplied additional information to the indexing algorithm in the form of a target unit cell, from either an isomorphous crystal form or a preliminary round of indexing. This target unit cell permits us to choose a group of three vectors that best fits the known cell's lengths and angles, thus increasing the number of successfully indexed images. A final factor is the

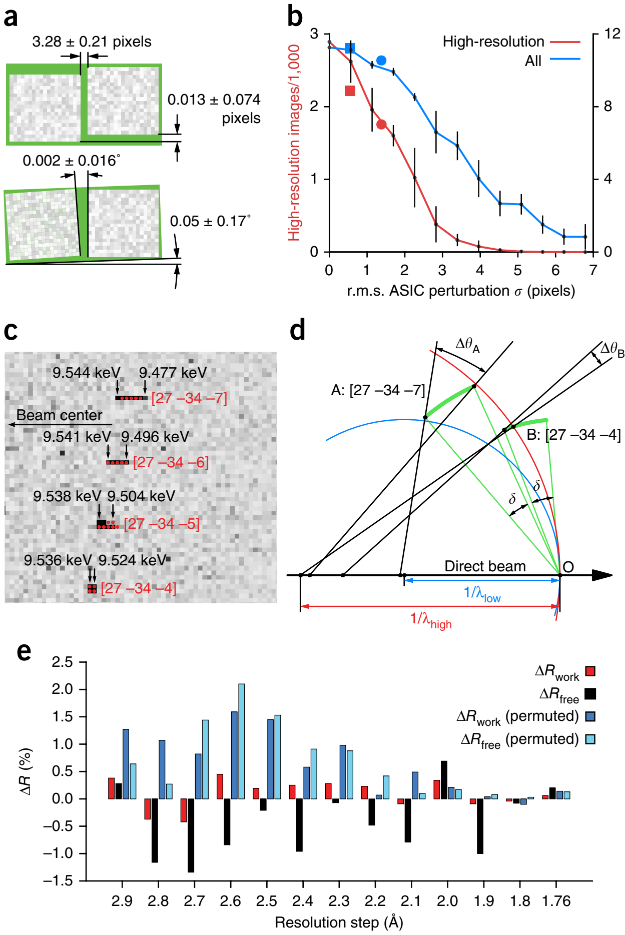

Figure 2 | Calibration and validation. (**a**) Aggregate relative positions (top) and rotations (bottom) of 32 pairs of application-specific integrated circuits (ASICs) on the CSPAD detector. The calibration results independently verify the manufacturing constraints (ASIC pairs are aligned along the long axis, separated by a 3.0-pixel gap). (**b**) Impact of positional accuracy on the indexing and integration success rate. Separately perturbing the ASICs away from their true positions reduced both the total number of indexed images (all) and the number of images that contain successfully integrated reflections at high (1.8–2.2 Å) resolution. Failure to apply final subpixel (squares) or whole-pixel (circles) corrections impaired processing as expected. σ, s.d. (**c**) Detail of four neighboring Bragg reflections from thermolysin (with Miller indices indicated in brackets), showing pronounced (7-pixel) radial elongation for the top reflection and lesser elongation for those nearby. Solution of Bragg's law for each pixel (vertical arrows) identifies the spread of photon energies that contribute to each reflection. Red disks delineate integration masks from a three-parameter model with wavelength limits λ_{high} = 1.297 (9.556 keV) and λ_{low} = 1.313 (9.443 keV) and full-width mosaic spread δ = 0.174°. (**d**) Reciprocal space diagram indicating how differently shaped reflections arise. Reciprocal lattice points (arcs) all have a constant angular extent δ owing to their mosaic spread. Points are in reflecting condition if they are within the zone between the high-energy (red) and low-energy (blue) Ewald spheres. Therefore, a greater fraction of the mosaic distribution from Bragg spot A is within the reflecting condition, leading to a reflection that subtends a greater radial angle $\Delta\theta$. (**e**) Paired refinements of the thermolysin structure and their impact on the reliability factors, ΔR. Shells of successively higher-resolution data are interpreted as improving the refinement results as long as ΔR_{free} is continuously negative, i.e., out to 2.1 Å.

high density of crystals delivered to the X-ray beam, which often produces diffraction patterns containing more than one lattice (**Fig. 1d**). Although software exists for modeling multiple lattices in SR diffraction[12,13], previous XFEL approaches[14] effectively filter these data away, by requiring that 80% of observed spots be covered by a single model. However, we found it straightforward to treat XFEL data with two lattices. The full set of bright candidate Bragg spots was used to derive the first lattice. Candidate spots falling on this lattice were then removed, and the remaining subset was used to find the second lattice, as has been described for SR data[12]. Spot overlaps among multiple lattices were rare, so we ignored the minimal inaccuracies in the integrated signal resulting from overlap.

The outcome of data integration depends critically on the ability to exactly target the pixels that actually contain signal. A too-inclusive model will capture adjacent pixels that contain only background noise, thus diluting the signal-to-noise ratio of the measurement. Conversely, overly discriminating models fail to include all of the signal. A crucial first step for data processing, therefore, is to tailor the model to the data at hand. We explain why there is a need for new data-modeling algorithms, beyond what is implemented by CrystFEL, in the **Supplementary Note**. In short, microscopic 'mosaic' domains in the crystal produce Bragg spots shaped like concentric arcs, and the spread of energies in the self-amplified spontaneous emission (SASE) pulse streaks spots radially.

For cctbx.xfel, we tested two approaches to model the Bragg spots. Although spots vary in size and shape across the lattice (**Fig. 1e**), they tend to be locally similar. This suggests that an empirical approach can be used whereby integration masks are chosen based on the shapes of nearby bright spots. We chose this method, which captures spot shapes of all extremes including both concentric arcs and radial streaks, as the default treatment for data analysis (**Supplementary Tables 1–3**). A deeper inspection of the data (**Fig. 2c**) revealed cases in which Bragg reflections adjacent to each other nonetheless had very distinct radial widths. These differing widths are explained by the fact that for the full spread of SASE energies to be recorded in the diffraction pattern, Bragg's law demands that the crystal contains microscopic (mosaic) domains with a distribution of either orientations or unit-cell dimensions. Wide radial-width spots were produced for reflections that satisfy Bragg's law for the full distribution of mosaic domains (given the crystal orientation and range of incident energies), and narrow radial-width spots were observed for those reflections that only satisfy the reflecting condition for a subset of domains (**Fig. 2d**). Modeling three parameters (high and low bandpass limits plus mosaicity) predicted approximately which pixels to target for signal integration (**Fig. 2d**). The key benefit of this second, parametric approach is that it roughly accounts for the size and shape differences of adjacent Bragg spots, thus helping the integration mask conform to the actual signal. Although the three-parameter model does not give an exact match to the spot shape (**Fig. 2d**), refinement of additional parameters could improve the approach.

We next tested how best to determine the resolution limits of the data set. An important consequence of shot-to-shot variability is that each lattice diffracts to a different limiting angle. Before merging the data into a single set of structure factors, we constructed Wilson plots (integrated Bragg spot intensity versus diffraction angle bin) in order to determine a separate cutoff angle for each lattice. Once the data had been merged, we employed an iterative paired-refinement technique[15] to determine the overall highest-resolution shell with a measurable information content (**Fig. 2e**). We found that at the highest resolution proven to contain significant signal (2.1 Å), only 1,700 lattices contributed to the thermolysin diffraction data, with an average multiplicity of observation of only 4.5 per structure factor (**Supplementary Table 2**). The size of this selected subset is much smaller than for previous high-resolution XFEL crystallography experiments; past experiments using CrystFEL have required $>10^4$ crystals to obtain reliable structure factors[6,10,16]. In cases in which only 10^2–10^3 diffracting crystals were available, data merging has been only partially successful[5,17]. Thus our results with cctbx.xfel are encouraging as XFEL progress has been limited by both the difficulty of preparing enough crystal specimens and the limited data-acquisition time at the light source.

In summary, our new developments implemented in cctbx.xfel include optimal indexing and retention of data from multiple lattices, separate determination of the resolution cutoff for individual lattices, better descriptions of the Bragg spot shape and accurate detector geometry to permit well-conforming spot-shape models. By carefully discriminating between image pixels known to contain diffraction signal and the surrounding pixels containing only background noise, we derived accurate structure factors with substantially fewer crystal specimen exposures.

We plan software developments to improve the final merged set of structure factors. A present limitation is that XFEL Bragg diffraction gives only a partial measurement of the structure factor, as the crystal is not fully rotated through the reflecting condition (**Fig. 2d**). We intend to implement postrefinement models[18,19] to allow the correction of intensity measurements to their full-spot equivalent. Such a correction requires detailed knowledge of the incident spectrum. In **Figure 2e**, we present the range of X-rays as a top-hat function, but in fact the SASE spectrum is stochastic and finely textured[20]. Although the X-ray spectra were not available for the data shown here, single-shot measurement of the spectrum is possible[20] and will be incorporated into our method in the future. Taken all together, our method will make it easier to use XFEL-based experiments to measure small structure-factor differences, such as those from anomalous scattering that will enable the *de novo* determination of macromolecular structures. Although SFX is presently a challenging technique, its potential payoff in terms of enabling specialized structural and dynamical studies of macromolecules is enormous.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Protein Data Bank: 4OW3 (structure factors and model for thermolysin); Coherent X-ray Imaging Data Bank ID23 (raw data streams for thermolysin).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health (NIH) grants GM095887 and GM102520 and Director, Office of Science, US Department of Energy (DOE) under contract DE-AC02-05CH11231 for data-processing methods (N.K.S.); Director, DOE Office of Science, Office of Basic Energy Sciences (OBES), Chemical Sciences, Geosciences and Biosciences Division (CSGB) under contract DE-AC02-05CH11231 (J.Y. and V.K.Y.); NIH grant GM055302 (V.K.Y.); and NIH grant P41GM103393 (U.B.). Sample injection was supported by LCLS (M.J.B. and D.W.S.) and the Atomic, Molecular and Optical Science program, CSGB Division, OBES, DOE (M.J.B.), and through the SLAC National Accelerator Laboratory Directed Research and Development program (M.J.B. and H.L.). J.M. was supported by the Artificial Leaf Project Umeå (K&A Wallenberg Foundation), the Solar Fuels Strong Research Environment Umeå (Umeå University), Vetenskapsrådet and Swedish Energy Agency (Energimyndigheten). Experiments were carried out at the LCLS at SLAC, an Office of Science User Facility operated for the DOE by Stanford University. We thank A. Perazzo, M. Dubrovin, I. Ofte, and A. Salnikov for collaboration on data analysis, and C. Kenney for expertise related to the CSPAD detector.

AUTHOR CONTRIBUTIONS

J. Hattne, J.K., J.Y., U.B., V.K.Y., P.D.A. and N.K.S. conceived of the new data-processing methods and analyzed the data; J. Hattne, N.E., R.J.G., A.S.B., R.W.G.-K., P.H.Z., M.M., P.D.A. and N.K.S. wrote the data-processing software; U.B., J.Y., V.K.Y., J.K., R.A.-M., J.M., A.Z., N.K.S., G.J.W., S.B., A.R.F., A.M., D.M., D.W.S., W.E.W. and M.J.B. designed the experiment; R.T., C.G., J. Hellmich, D.D., A.L., G.H., J.K. and A.Z. prepared samples; S.B., J.E.K., M.M., M.M.S., G.J.W. operated the CXI instrument; M.J.B., H.L., R.G.S., J.K., J.M., B.L.-K., S.G., R.T., C.G., J. Hellmich, J.S., D.W.S., A.M. and G.J.W. developed, tested and ran the sample delivery system; R.A.-M., U.B., M.J.B., S.B., N.E., R.J.G., P.G., C.G.,S.G., G.H., J.Hattne., J.Hellmich, J.K., J.E.K., H.L., A.L., B.L.-K., D.M., M.M., J.M., N.K.S., M.M.S., J.S., R.G.S., D.S., R.T., T.-C.W., G.J.W., V.K.Y., J.Y. and A.Z. performed the LCLS experiment; J. Hattne, N.E., J.K., J.Y., U.B., V.K.Y., P.D.A. and N.K.S. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Neutze, R. *et al. Nature* **406**, 752–757 (2000).
2. Alonso-Mori, R. *et al. Proc. Natl. Acad. Sci. USA* **109**, 19103–19107 (2012).
3. Kern, J. *et al. Science* **340**, 491–495 (2013).
4. Chapman, H.N. *et al. Nature* **470**, 73–77 (2011).
5. Koopmann, R. *et al. Nat. Methods* **9**, 259–262 (2012).
6. Redecke, L. *et al. Science* **339**, 227–230 (2013).
7. Bourenkov, G.P. & Popov, A.N. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 58–64 (2006).
8. White, T.A. *et al. J. Appl. Cryst.* **45**, 335–341 (2012).
9. Sauter, N.K. *et al. Acta Crystallogr. D Biol. Crystallogr.* **69**, 1274–1282 (2013).
10. Boutet, S. *et al. Science* **337**, 362–364 (2012).
11. Sauter, N.K., Grosse-Kunstleve, R.W. & Adams, P.D. *J. Appl. Cryst.* **37**, 399–409 (2004).
12. Sauter, N.K. & Poon, B.K. *J. Appl. Cryst.* **43**, 611–616 (2010).
13. Powell, H.R., Johnson, O. & Leslie, A.G. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1195–1203 (2013).
14. Kirian, R.A. *et al. Acta Crystallogr. A* **67**, 131–140 (2011).
15. Karplus, P.A. & Diederichs, K. *Science* **336**, 1030–1033 (2012).
16. Kirian, R.A. *et al. Opt. Express* **18**, 5713–5723 (2010).
17. Johansson, L.C. *et al. Nat. Methods* **9**, 263–265 (2012).
18. Winkler, F.K., Schutt, C.E. & Harrison, S.C. *Acta Crystallogr. A* **35**, 901–911 (1979).
19. Rossmann, M.G. *et al. J. Appl. Cryst.* **12**, 570–581 (1979).
20. Zhu, D. *et al. Appl. Phys. Lett.* **101**, 034103 (2012).

ONLINE METHODS

Sample preparation. Lyophilized thermolysin from *Bacillus stearothermophilus* (Hampton Research) was resuspended in 0.05 M NaOH at a concentration of 25 mg/ml. 300 µl of the protein stock was mixed in a 1:1 ratio with 40% PEG 2000, 100 mM MES pH 6.5 and 5 mM $CaCl_2$. Crystallization occurred within minutes. The obtained crystals were transferred into 10% PEG 2000, 100 mM MES pH 6.5, 5 mM $CaCl_2$ (buffer A) and then stepwise into buffer A containing 10%, 15%, 20% and 30% (w/v) glycerol, respectively. Thermolysin concentration was determined spectrophotometrically using an absorbance value A = 1.83 (1 mg/ml) at 277 nm (ref. 21) and a molecular mass of 34.6 kDa[22]. The final protein concentration of the crystal suspension was 20–24 mg/ml. The average size of the obtained crystals was 2 µm × 3 µm × 1 µm. As judged by microscope images of various batches, the size distribution was very narrow. Assuming an average crystal volume of 6 μm^3, 12 monomers per unit cell and a nominal unit cell volume of 1×10^6 Å^3, 6×10^5 unit cells/crystal gives a concentration of ~3.4×10^{10} crystals/ml.

Thermolysin data collection. Diffraction experiments were carried out at the CXI instrument at LCLS[23]. We had previously reported the use of a nanoflow liquid injector that markedly reduces the requirements on sample amount[24,25]. The suspension of thermolysin crystals was injected into the interaction region by this electrospun liquid jet, using a 1-m-long silica capillary of 50 µm inner diameter, 150 µm outer diameter, outer diameter tapered at both ends (New Objective) with one end in a pressurized cell outside the vacuum chamber of the CXI instrument, dipping into a vial with 100 µl of the crystal suspension. A potential of +2,500 V (relative to a counter electrode below the interaction region) was applied to the suspension by means of a bare Pt electrode inside the sample vial. The flow rate was on the order of 0.5 µl/min by applying a backing pressure of 124.1 kPa to the suspension.

The CXI instrument was operated at energies of 9.56 keV and 9.77 keV (**Supplementary Table 1**), and the beam intensity was 6×10^{11} photons/pulse, with a mean pulse duration of 47 fs and a frequency of 120 Hz. The beam was focused to a size of 2.25 μm^2 full-width half maximum (FWHM) at the interaction point. Diffraction was measured using the front CSPAD detector[26] of the CXI instrument. The detector has a pixel size of 110 µm × 110 µm and a total of 1,516 × 1,516 pixels.

Resolution of this particular experiment was limited by geometric factors and not the intrinsic strength of the diffraction signal. Several combinations of sample-to-detector distance and incident wavelength were used for data collection, but with the most aggressive choice (detector distance = 135 mm, λ = 1.30 Å), geometric limits were 2.15 Å at the detector edge and 1.75 Å in the corner, thus accounting for the falloff in data completeness at high resolution in **Supplementary Table 2**.

Raw data streams have been deposited into the Coherent X-ray Imaging Data Bank[27] (CXIDB), along with an exact list of the images that were merged (**Supplementary Tables 1** and **2**) to form the structure factor intensities. A tutorial on accessing information from the raw data files is available at http://cci.lbl.gov/xfel.

Lysozyme data. To afford a fair comparison between CrystFEL and cctbx.xfel, our only tractable option was to reprocess raw data that had been previously analyzed by the CrystFEL software developers. We obtained data from the CXIDB, which archives the raw data streams from the 1.9 Å–resolution structure determination of lysozyme[10] under accession ID 17. To select data for the comparison, we chose only those run numbers (305–327) that yielded the 12,247 images used in ref. 10, as documented in a list maintained at the CXIDB website (**Supplementary Table 1**). For those run numbers, the CXI instrument was operated at 9.39 keV and the pulse duration was 40 fs. With a detector distance of 93 mm, the geometric limits were 1.74 Å at the detector edge and 1.46 Å in the corner, both well beyond the 1.9 Å resolution limit that we imposed in order to perform a direct comparison with the published results.

Data processing. Data were processed with our package cctbx.xfel[9]. After subtraction of a dark-run average image, bright candidate Bragg spots were chosen with the Spotfinder component of cctbx[28], with settings being adjusted by trial and error specifically for these data; e.g., the minimum spot area was set at two square pixels, and the criteria for accepting spots was set to allow spot picking to an outer resolution limit of about 2.5 Å for thermolysin and 1.9 Å for lysozyme. Images were indexed (unit cell dimensions and crystal orientations determined) with the Rossmann data-processing system (DPS) algorithm[29,30] as implemented in our program Labelit[11]. Unit-cell dimensions modeled by the indexing algorithm varied from crystal to crystal; population means and s.d. for thermolysin are reported in **Supplementary Table 1**. A small number of thermolysin lattices (233, ~2%) did not conform to hexagonal Bravais symmetry using our standard criteria[31]; these were removed from further processing and are not included in the reported population. Similarly, 321 non-tetragonal lysozyme lattices were removed (~1%). For previous data analyses with photosystem II[3,32], we also removed lattices whose unit-cell lengths were highly non-isomorphous (differing by >10%) compared to the mean, in order to avoid merging data from nonidentical crystal structures[33,34]. However, for the thermolysin and lysozyme data, none of the unit-cell lengths were rejected as outliers.

Improving indexing by using a target unit cell. Destruction of each crystal after one XFEL shot makes indexing difficult. Accuracy is much greater at SR sources, where it is possible to mount the crystal on a goniometer and view the diffracted lattice from two different crystal orientations ~90° apart[11]. In contrast, the liquid-jet method delivers samples in random, unknown, orientations. Furthermore, the XFEL diffraction images examined here varied extensively in quality (resolution and number of Bragg spots), with a less successful indexing outcome from poorer images. With degraded data, the DPS algorithm can fail by choosing three candidate unit cell axes that individually appear to describe periodicity in the diffraction pattern, but when combined do not adequately cover the lattice. To avoid this failure mode, we supplied additional information to the indexing algorithm in the form of a target unit cell taken from isomorphous crystal forms (Protein Data Bank (PDB) codes 2TLI for thermolysin and

doi:10.1038/nmeth.2887

4ET8 for lysozyme). Groups of three candidate axes from the DPS algorithm are evaluated to find the best fit to the known cell lengths and angles. By requiring this approximate similarity, we increased the number of successfully indexed images from ~8,000 to ~11,600 for thermolysin. A similar approach was used previously by others to identify the lattice within noisy data[35,36]. We expect that this method will be generally applicable to XFEL data and not limited to cases in which an isomorphous crystal form is known. Data can be treated in two passes, first to determine a consensus unit cell from the highest-quality diffraction images where indexing is readily achieved, and second to use this consensus cell as a target for indexing the entire data set. In support of this idea, we note that the population s.d. of the thermolysin unit-cell lengths (**Supplementary Table 1**) is quite narrow (0.3–0.4%), and even for previous low-resolution photosystem II data[3] the s.d. values (0.9–1.9%) were reasonably low.

Relationship between indexing and hit rates. We had previously described the use of cctbx.xfel to provide detailed feedback on the diffraction quality within minutes of data acquisition[9]. For this initial analysis, the Spotfinder component of cctbx[28] is used to classify a diffraction pattern as a 'hit' if it contains 16 or more candidate Bragg spots with dark-subtracted peak heights above 450 analog-digital units (on the CSPAD high-gain setting) out to a resolution limit of 4.0 Å. This peak height criterion is chosen by trial and error to best identify Bragg spots for the thermolysin data set, and the level can easily be changed in a configuration file for other data sets. In **Supplementary Figure 1** we show the final outcome: 77% of the initial low-resolution 'hits' are successfully integrated and merged into structure factors, with a slightly lower success rate (65%) for hits containing the lowest number of candidate spots. Reasons for the residual failure rate are still to be determined and will likely vary from case to case in future experiments.

Empirical approach to modeling the spot shape. Bragg spots from both data sets (thermolysin data are illustrated in **Fig. 1e**) were observed to vary in size and shape both within a single lattice and also from image to image. Therefore, the previously published CrystFEL model that treats spots as uniformly round and equally sized in reciprocal space[14] was judged to be a poor fit to these data. As described in the **Supplementary Note**, the underlying phenomenon treated by that model (large λ/a ratio, where λ is the wavelength of the incident light, and a is the crystal width) does not apply for high-resolution experiments. In fact, it is not possible to identify a single criterion to describe the spot shape throughout the data sets; some images exhibit concentric arcs consistent with mosaic spread[37] (data not shown), whereas other images contain elongation that is chiefly radial (**Fig. 1e**). We do note, however, that whatever the behavior, spots tend to be locally similar in size and shape in each lattice (with one exception, see below). This suggests an empirical approach to determining the spot model. First, easily identified high-intensity Bragg spots (using the program Spotfinder[28]) are used to index the lattice. Next, at each predicted lattice position on the image, a mask is constructed consisting of a union of the ten nearest spot shapes from the Spotfinder set, similar to the approach taken by some SR data-reduction programs[38]. This mask determines the set of pixels to be used for signal summation (integration). Taking a union of all nearby spot masks helps to increase the number of pixels assigned to each Bragg spot, to avoid missing pixels that actually contain signal. This is necessary because the predicted spot positions are slightly inaccurate due to the use of a monochromatic model; in fact the incident light has a 0.2–0.5% bandpass[39] (as described below). This simple empirical approach was used to derive all the structure factor measurements in **Supplementary Tables 1–3**.

Parametric approach to modeling the spot shape. Given the theoretical framework of Bragg's law, it is possible to interpret the shape and size of Bragg spots in terms of more fundamental experimental properties including the spectral dispersion, the crystal size and the internal crystal disorder[40–47]. Thus, although the above empirical approach is adequate for the present, a deeper understanding of XFEL Bragg spot shapes may be possible. In images of both thermolysin (**Figs. 1e** and **2c**) and lysozyme we observed radial spot elongation that is most pronounced at higher diffraction angles. This is consistent with the protein crystals acting as spectral analyzers, such that each Bragg reflection disperses the broad bandpass SASE pulse (typically 0.2–0.5% bandpass)[39] over a radial line up to several pixels wide. Furthermore, we observe that reflections adjacent to each other (**Fig. 2c**) can nonetheless have very distinct radial widths. The explanation is rooted in the fact that for a spread of energies to be recorded in the diffraction pattern, Bragg's law demands that the crystal contain microscopic (mosaic) domains with a distribution of either orientations or unit cell dimensions. In **Figure 2d** we represent each Bragg spot as a spherical cap in reciprocal space (shown as an arc) representing a spread of orientations, as has been done previously[48]. In our experiment, wide spots are produced for reflections that satisfy Bragg's law for the full distribution of mosaic domains in the crystal (given the crystal orientation and range of incident energies), whereas narrow spots are seen for those reflections that only satisfy the reflecting condition for a subset of microscopic domains (**Fig. 2d**). By modeling three parameters (high and low bandpass limits, plus mosaicity) we could predict approximately which pixels to target for signal integration for each Bragg reflection (**Fig. 2d**). The key benefit of this approach is that it roughly accounts for the size and shape differences of adjacent Bragg spots, reducing the inclusion of non-signal pixels in the integration mask and thus helping to extract weak signals. Although the three-parameter model in **Figure 2d** does not give an exact match to the spot shape, we believe that further development will improve the approach. Important additional parameters that could be refined include the spectral shape and unit cell variation, whereas others such as crystal size and beam divergence are probably negligible for experiments performed at the CXI 1 μm focus.

Signal integration and error estimation. Signal intensity I for each Bragg spot was integrated over a set of pixels determined by empirical mask construction as described above. A surrounding set of pixels, twice the size of the signal set, and separated from it by a guard zone two pixels wide, was designated for measuring the local background. This background set was used to fit a least-squares plane for background subtraction as described[49]. The estimated variance $\sigma^2(I)$ of the signal measurement was based on counting statistics[49], using a rough estimate for the

doi:10.1038/nmeth.2887

CSPAD high-gain value of 7.5 analog-to-digital units per photon. Integrated intensities were then corrected for polarization[50]. It was realized that the data set contained numerous intensity measurements at large negative multiples of σ (I), from which we concluded that Poisson statistics did not adequately model the experimental error. Error estimates from each diffraction pattern were therefore inflated by assuming that negative values of I/σ (I) are actually decoy measurements (noise only) with a Gaussian distribution centered at zero and with a s.d. of 1, thus providing a lower bound on modeling errors. This inflation factor is determined separately for each image, and acts to increase the initially determined errors from counting statistics. Negative I values were then removed from the data set, and data on each image were scaled to a reference data set derived from an isomorphous structure (see below). When later merging multiple measurements of the same Miller index, the error was modeled simply by propagating the per-measurement σ (I) values in quadrature. As the systematic error contributions for XFEL data are not fully understood, no other systematic correction or error normalization was attempted. The error model derived here is believed to be entirely different than that used in CrystFEL; therefore, the respective I/σ (I) values for the two programs in **Supplementary Tables 1**–**3** cannot be compared.

Scaling. Integrated intensities from separate images were scaled to intensities derived from an isomorphous reference structure (PDB codes 2TLI for thermolysin and 4ET8 for lysozyme); this scaling step helped to correct for specimen-to-specimen variation in crystal size and pulse power. For projects for which no isomorphous reference structure is available, we propose an iterative procedure wherein the data are merged once without scaling to gain an approximate set of merged intensities, which are then used as the reference for rejecting poorly correlated images in the next round.

Different resolution cutoffs for each lattice. An important consequence of shot-to-shot variability is that each lattice diffracts to a different limiting angle; this can be illustrated even within a single image (**Fig. 1d**) where one lattice (red) extends to higher resolution than a second one (blue). For data reduction, we choose a separate limit for integrating each lattice. Integration relies on having an accurate crystal orientation model, which in turn depends on the set of bright candidate Bragg spots found in our case by the program Spotfinder[28]. For example, if Spotfinder spots extend only to 4 Å on a particular image, the orientational model is not accurate enough to predict the positions of weak spots at 2.5 Å resolution. We verified this general result through studies on simulated data (data not shown). A very conservative approach is therefore used for integration: for each image separately, the radius of integration is extended slightly past the Spotfinder limit, and a Wilson plot is constructed (integrated Bragg spot intensity versus diffraction angle bin), to identify a resolution limit at which average intensity falls below average noise (based on counting statistics). The radius is increased until such a crossover point is found, at which point it is concluded that either there is no more signal to be found or the model has diverged from the data. When merging multiple measurements together, it would be counterproductive to include high-resolution integrated measurements from beyond this limit where there is no signal, as this would degrade the overall signal-to-noise ratio. Allowing separate resolution cutoffs for each image leads to a final merged data set with high multiplicity of observation at low resolution and lower multiplicity at high resolution (**Supplementary Table 2**), yet there is confidence that the highest-resolution shell contains real signal.

The quality of the reflections merged in this fashion was assessed by calculating the correlation coefficient of semi-data sets merged from odd- and even-numbered images ($CC_{1/2}$)[15]. We note that our multiplicity statistics (**Supplementary Tables 2** and **3**) differ from those in previously published high-resolution XFEL analyses[6], which report uniform multiplicity counts over all resolution bins, which is the result of applying a single global resolution limit.

Validation of the resolution cutoff. As the data quality gradually decreases at the highest resolution (**Supplementary Table 2**), it would be advantageous to derive a convenient statistical 'rule of thumb' to determine the highest resolution that contains valid, merged structure factors. There must be some reasonable cutoff as the multiplicity of observation and the internal correlation coefficient $CC_{1/2}$ decrease, but it needs to be established which cutoff values should be chosen. To provide an objective criterion, we used the iterative paired-refinement technique suggested in ref. 15. Each iteration compares the result of two atomic structure refinements, the first using data only out to a conservative resolution limit, and the second including reflections in the next, higher-resolution shell. The two models are then evaluated against the smaller, low-resolution set of reflections, and the two reliability factors are computed (R_{work} and R_{free}[51]). As long as R_{free} decreases, the added data contribute useful information to the refinement. An increase in R_{work} but unchanged R_{free} indicates that the model has become less overfit. As a negative control, the model is refined a third time adding the same higher-resolution intensities, but with randomly permuted (incorrect) Miller indices in the shell. Analysis of the thermolysin data starting at 3.0 Å, and progressing in steps of 0.1 Å toward the highest-resolution limit (1.76 Å) showed that the refinement results improve (i.e., R_{free} decreases) out to at least 2.1 Å (**Fig. 2e**), whereas randomly permuted Miller indices nearly always increase the R factors, as expected. At the 2.1-Å cutoff, the average observational multiplicity of each structure factor was only 4.5, and the correlation coefficient between semi-data sets was 17.0%.

Relationship between resolution and accurate detector model. The empirical and parametric approaches to constructing Bragg spot profiles as outlined above place very stringent requirements on the geometrical modeling (metrology) of the detector. Many diffraction patterns (**Fig. 1**) exhibited Bragg spots that are only one or 2 square pixels in area, particularly at low resolution. For spot modeling to work as proposed, therefore, the position of each pixel in space must be known to substantially better accuracy than the pixel dimension, but this is a difficult goal for current XFEL detectors owing to their unique construction as a mosaic of pixel array sensors[26,52]. We took a bootstrapping approach starting with approximately known sensor positions, followed by the use of Bragg observations from the entire data set (either thermolysin or lysozyme), to derive more accurate sensor positions and orientations by iterative nonlinear least-squares positional

doi:10.1038/nmeth.2887

refinement (see below). This improved metrology allowed us to model the Bragg spots with an r.m.s. deviation (observed spot position versus modeled position) of 0.65 pixels and 1.00 pixels for thermolysin and lysozyme, respectively. Any well-diffracting set of protein crystals would have sufficed for this procedure; it was not necessary for the unit cell or structure to be known ahead of time.

To assess the general importance of accurate detector metrology we carried out an analysis in which the accurately refined sensor positions were intentionally perturbed, with shifts drawn from a two-dimensional normal distribution with s.d. σ_r (**Fig. 2b**). Five repetitions were performed for each σ_r magnitude. Indexing success depended weakly on metrology (half of the images could still be indexed with a positional perturbation of 3.5 pixels); but high-resolution integration was strongly dependent, with a 30% loss of high-resolution signal resulting from a perturbation of just a single pixel. This is exactly as expected; our empirically determined integration masks conformed very tightly to the spot shape; therefore for the method to work, the positions of individual detector tiles need to be accurately known.

We arrived at the same conclusion, by a different route, if we simply reversed the refinement steps of our detector calibration. This outcome (for the thermolysin data) was also plotted in **Figure 2b**. Reversing the final step of iterative nonlinear least-squares positional refinement left us with sensor positions 0.55 pixels away from their true positions, with consequent loss in both high-resolution and overall data. Reversing the penultimate step (where we determined the nearest whole-integer pixel positions without any sensor rotations) put the sensors 1.38 pixels away from true, with a further degradation in the results.

Refinement of the detector geometry model (metrology). The CSPAD detector used at the CXI instrument is laid out in a mosaic arrangement consisting of four groups (quadrants) of eight silicon pixel-array sensors[26]. As the quadrants can be translated on mechanical rails, a coarse determination of their relative positions must be made before any Bragg patterns can be analyzed. Pseudo-powder patterns were synthesized for this purpose by summing a large number of thermolysin diffraction images, all recorded at the same sample-detector distance. A graphical application was written, permitting the manual adjustment of the quadrant locations to align the observed powder rings with overlaid circular fiducial rings. This program is also suitable for calibrating the detector quadrants with silver behenate[53] powder patterns.

Before the experiment, sensor positions and orientations (in each quadrant) were characterized optically at the LCLS to within tens of micrometers, but this calibration did not necessarily achieve the accuracy required for spot modeling, nor did it probe the actual readouts that are bump-bonded to the sensors. Each sensor was bonded to a pair of side-by-side 194 × 185 pixel application-specific integrated circuits (ASICs)[26]. Detailed positions and orientations of the 64 ASIC readouts were refined by nonlinear least-squares refinement of the target functional

$$f = \sum_{\substack{\text{ASICs,}\\ \text{crystals,}\\ \text{spots}}} (\mathbf{r}_{\text{obs}} - \mathbf{r}_{\text{calc}})^2$$

where $\mathbf{r}_{\text{obs}}$ is the observed detector position of the Bragg spot centroid determined with the program Spotfinder[28], $\mathbf{r}_{\text{calc}}$ is the modeled position after indexing, and the sum is over all Spotfinder spots (on all images and ASICs) that correspond to modeled spots. Variable parameters in the refinement included the positions and rotations of all ASICs, the position of the direct beam and crystal-to-detector distance for each crystal shot, and the orientation and unit cell dimensions for each crystal. Correct performance of this algorithm was monitored by considering the refined placement of pairs of ASICs bonded to the same silicon sensor, which are thought to be exactly aligned by a mechanical guide piece during the manufacture process. These internal controls derived from the thermolysin data (**Fig. 2a**) showed that the ASIC pairs are mutually aligned to an r.m.s. rotation of 0.016° and an r.m.s. displacement perpendicular to the long sensor axis of 0.074 pixels; we interpreted these values as the accuracy limits of our refinement method. The tolerances were similar for the lysozyme data, 0.030° and 0.072 pixels, respectively. In addition, we found that on the particular detector used for thermolysin, the 32 sensors had an r.m.s. tilt of 0.17° in the plane of the detector, and that the separation between same-sensor ASIC pairs varied with an r.m.s. deviation of 0.21 pixels (**Fig. 2a**).

Refinement of the detector distance. We calibrated the absolute distance between crystal sample and imaging detector to an accuracy of ~1 mm. Fortunately the indexing algorithm and indeed the entire data-processing pipeline was robust to this level of uncertainty, with small errors in the distance being absorbed by other modeled parameters (unit-cell dimensions, wavelength). We determined the distance by grid search around an initial estimate: an entire run collected at a fixed distance was reprocessed several times with calibration offsets differing by 0.5 mm, which were then scored by counting the number of images successfully indexed (**Supplementary Fig. 2**). Offsets of ± 8 mm from the best value reduced the indexing rate by roughly a factor of 2.

An alternate distance calibration is possible by observing circular powder patterns from silver behenate as noted above, and the cctbx.xfel software can facilitate this analysis. Such a calibration might offer improved accuracy as it uses a recognized standard, but as a practical matter, given the time constraints of collecting data at LCLS, it was more efficient to simply use the thermolysin or lysozyme data to estimate the distance as shown in **Supplementary Figure 2**.

Structure solution. Merged structure factors were phased by molecular replacement using Phaser[54] within the Phenix[55] system. For thermolysin, the search model consisted of thermolysin (PDB code 2TLI[56]) from which all nonprotein atoms were removed; for lysozyme the model was taken from PDB code 4ET8 (ref. 10). New models were built into the resulting maps using phenix.autobuild[57], and refined using phenix.refine[58]. Refinement statistics are shown in **Supplementary Table 1**. The molecular clashscore (number of bad all-atom overlaps per thousand atoms) and Ramachandran stereochemical statistics were calculated with MolProbity[59].

Crystallographic *R* factors for the refined thermolysin model are comparable in quality to synchrotron structures that have been determined at a similar resolution (2.1 Å). To determine this, we used the program phenix.r_factor_statistics[60,61] to print the *R* factor distribution from 2,271 PDB structures at resolutions in the range 2.05–2.15 Å. Our thermolysin values of R_{work} = 22.2%

doi:10.1038/nmeth.2887

and R_{free} = 26.5% were within 1 s.d. of the mean (R_{work} = 20.1 ± 2.4%; R_{free} = 24.6 ± 2.6%). The R-factor distribution was derived by taking coordinates, structure factors and R_{free} flags from the PDB, and using the Phenix toolbox to derive the R factors. As a result, the distributions can be directly compared with our refinements, which were also performed with Phenix.

Similarly, for the 1.9 Å lysozyme structure, we considered 3,578 PDB structures at resolutions in the range 1.85–1.95 Å. Our Phenix-refined values of R_{work} = 18.7% and R_{free} = 22.9% for the cctbx.xfel structure factors, and R_{work} = 17.7% and R_{free} = 22.0% for the CrystFEL structure factors, were each within 1 s.d. of the mean (R_{work} = 19.3 ± 2.3%; R_{free} = 23.2 ± 2.6%).

21. Inouye, K. *J. Biochem.* **112**, 335–340 (1992).
22. Titani, K. *et al. Nature* **238**, 35–37 (1972).
23. Boutet, S. & Williams, G.J. *New J. Phys.* **12**, 035024 (2010).
24. Sierra, R.G. *et al. Acta Crystallogr. D Biol. Crystallogr.* **68**, 1584–1587 (2012).
25. Bogan, M.J. *Anal. Chem.* **85**, 3464–3471 (2013).
26. Hart, P. *et al. Proc. SPIE* **8504**, 85040C (2012).
27. Maia, F.R.N.C. *Nat. Methods* **9**, 854–855 (2012).
28. Zhang, Z. *et al. J. Appl. Cryst.* **39**, 112–119 (2006).
29. Steller, I., Bolotovsky, R. & Rossmann, M.G. *J. Appl. Cryst.* **30**, 1036–1040 (1997).
30. Rossmann, M.G. & van Beek, C.G. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1631–1640 (1999).
31. Sauter, N.K., Grosse-Kunstleve, R.W. & Adams, P.D. *J. Appl. Cryst.* **39**, 158–168 (2006).
32. Kern, J. *et al. Proc. Natl. Acad. Sci. USA* **109**, 9721–9726 (2012).
33. Giordano, R. *et al. Acta Crystallogr. D Biol. Crystallogr.* **68**, 649–658 (2012).
34. Diederichs, K. & Karplus, P.A. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1215–1222 (2013).
35. Paithankar, K.S. *et al. Acta Crystallogr. D Biol. Crystallogr.* **67**, 608–618 (2011).
36. White, T.A. *et al. Acta Crystallogr. D Biol. Crystallogr.* **69**, 1231–1240 (2013).
37. Nave, C. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 848–853 (1998).
38. Otwinowski, Z. & Minor, W. *Methods Enzymol.* **276**, 307–326 (1997).
39. Emma, P. *et al. Nat. Photonics* **4**, 641–647 (2010).
40. Greenhough, T.J. & Helliwell, J.R. *J. Appl. Cryst.* **15**, 338–351 (1982).
41. Greenhough, T.J. & Helliwell, J.R. *J. Appl. Cryst.* **15**, 493–508 (1982).
42. Greenhough, T.J., Helliwell, J.R. & Rule, S.A. *J. Appl. Cryst.* **16**, 242–250 (1983).
43. Ren, Z. & Moffat, K. *J. Appl. Cryst.* **28**, 461–481 (1995).
44. Dauter, Z. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1703–1717 (1999).
45. Diederichs, K. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 535–542 (2009).
46. Schreurs, A.M.M., Xian, X. & Kroon-Batenburg, L.M.J. *J. Appl. Cryst.* **43**, 70–82 (2009).
47. Porta, J. *et al. Acta Crystallogr. D Biol. Crystallogr.* **67**, 628–638 (2011).
48. Bolotovsky, R. & Coppens, P. *J. Appl. Cryst.* **30**, 65–70 (1997).
49. Leslie, A.G.W. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 48–57 (2006).
50. Kahn, R. *et al. J. Appl. Cryst.* **15**, 330–337 (1982).
51. Brünger, A.T. *Nature* **355**, 472–475 (1992).
52. Strüder, L. *et al. Nucl. Instrum. Methods Phys. Res. A* **614**, 483–496 (2010).
53. Huang, T.C. *et al. J. Appl. Cryst.* **26**, 180–184 (1993).
54. McCoy, A.J. *et al. J. Appl. Cryst.* **40**, 658–674 (2007).
55. Adams, P.D. *et al. Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
56. English, A.C. *et al. Proteins* **37**, 628–640 (1999).
57. Terwilliger, T.C. *et al. Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
58. Afonine, P.V. *et al. Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).
59. Chen, V.B. *et al. Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
60. Urzhumtseva, L. *et al. Acta Crystallogr. D Biol. Crystallogr.* **65**, 297–300 (2009).
61. Afonine, P.V. *et al. J. Appl. Cryst.* **43**, 669–676 (2010).

doi:10.1038/nmeth.2887