JOURNAL OF SYNCHROTRON RADIATION

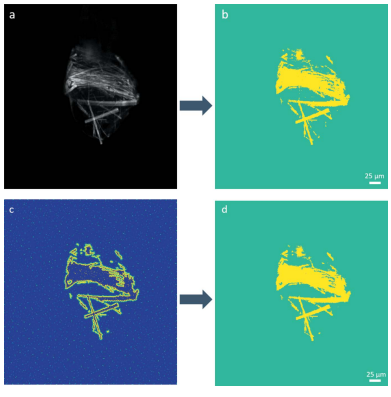
ISSN 1600-5775

Received 30 June 2016
Accepted 11 October 2016

Edited by R. W. Strange, University of Essex, UK

Keywords: dynamic sampling; supervised learning approach; X-ray diffraction; nonlinear optical microscopy; two-photon-excited fluorescence; second-harmonic generation.

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



Dynamic X-ray diffraction sampling for protein crystal positioning

Nicole M. Scarborough,[a] G. M. Dilshan P. Godaliyadda,[b] Dong Hye Ye,[b] David J. Kissick,[c] Shijie Zhang,[a] Justin A. Newman,[a] Michael J. Sheedlo,[a] Azhad U. Chowdhury,[a] Robert F. Fischetti,[c] Chittaranjan Das,[a] Gregery T. Buzzard,[d] Charles A. Bouman[b] and Garth J. Simpson[a]*

[a]Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA, [b]Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA, [c]GM/CA@APS, X-ray Science Division, Argonne National Laboratory, Lemont, IL 60439, USA, and [d]Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA. *Correspondence e-mail: gsimpson@purdue.edu

A sparse supervised learning approach for dynamic sampling (SLADS) is described for dose reduction in diffraction-based protein crystal positioning. Crystal centering is typically a prerequisite for macromolecular diffraction at synchrotron facilities, with X-ray diffraction mapping growing in popularity as a mechanism for localization. In X-ray raster scanning, diffraction is used to identify the crystal positions based on the detection of Bragg-like peaks in the scattering patterns; however, this additional X-ray exposure may result in detectable damage to the crystal prior to data collection. Dynamic sampling, in which preceding measurements inform the next most information-rich location to probe for image reconstruction, significantly reduced the X-ray dose experienced by protein crystals during positioning by diffraction raster scanning. The SLADS algorithm implemented herein is designed for single-pixel measurements and can select a new location to measure. In each step of SLADS, the algorithm selects the pixel, which, when measured, maximizes the expected reduction in distortion given previous measurements. Ground-truth diffraction data were obtained for a 5 µm-diameter beam and SLADS reconstructed the image sampling 31% of the total volume and only 9% of the interior of the crystal greatly reducing the X-ray dosage on the crystal. Using *in situ* two-photon-excited fluorescence microscopy measurements as a surrogate for diffraction imaging with a 1 µm-diameter beam, the SLADS algorithm enabled image reconstruction from a 7% sampling of the total volume and 12% sampling of the interior of the crystal. When implemented into the beamline at Argonne National Laboratory, without ground-truth images, an acceptable reconstruction was obtained with 3% of the image sampled and approximately 5% of the crystal. The incorporation of SLADS into X-ray diffraction acquisitions has the potential to significantly minimize the impact of X-ray exposure on the crystal by limiting the dose and area exposed for image reconstruction and crystal positioning using data collection hardware present in most macromolecular crystallography end-stations.

1. Introduction

X-ray diffraction (XRD) at synchrotron facilities is the most widely used approach for generating high-resolution structures of macromolecules. Synchrotron and X-ray free-electron lasers exhibit much higher fluxes and tighter localization than benchtop sources, allowing for substantial improvements in signal-to-noise and total analysis time. A key step in this pipeline is accurately positioning the protein crystal prior to diffraction analysis. Challenges in crystal positioning are exacerbated by current trends towards serial crystallography and nanocrystal analysis using synchrotron sources and fixed

targets, in which the reduced dimensions of the protein crystals present challenges for conventional imaging approaches (Aishima *et al.*, 2010; Andrey *et al.*, 2004; Cherezov *et al.*, 2009; Moukhametzianov *et al.*, 2008; Pothineni *et al.*, 2006; Stepanov, Hilgart *et al.*, 2011). Positioning becomes even more challenging when the crystals are in a turbid medium such as lipidic mesophase, in which techniques currently in use for crystal detection, *i.e.* bright-field imaging (Andrey *et al.*, 2004; Jain & Stojanoff, 2007; Pothineni *et al.*, 2006) and UV fluorescence imaging (Pohl *et al.*, 2004; Pothineni *et al.*, 2006; Vernede *et al.*, 2006), routinely offer poor discrimination between a crystal and its surroundings.

Nonlinear optical imaging methods are particularly promising as they are capable of detecting protein crystals rapidly through turbid media without inducing damage from X-rays or UV radiation (Kissick *et al.*, 2010, 2011). This multimodal nonlinear optical microscope combines complementary optical imaging techniques, such as second-harmonic generation (SHG), two-photon-excited ultraviolet fluorescence (TPE-UVF), two-photon-excited fluorescence (TPEF), single-photon visible fluorescence and laser transmittance bright-field imaging, in a single platform (Newman *et al.*, 2016). SHG, or the frequency doubling of light, provides contrast for noncentrosymmetric crystalline material, with no signal for amorphous protein aggregate (Kissick *et al.*, 2010). In TPE-UVF, a 532 nm laser is used to excite fluorescence from aromatic residues in proteins such as tryptophan (Madden *et al.*, 2011). TPEF and single-photon fluorescence provide an additional complimentary fluorescence mechanism able to detect proteins with color centers, as well as proteins which may have undergone oxidation (Padayatti *et al.*, 2012). A recent study also suggests that, under cryogenic conditions, TPEF can excite intrinsic fluorescence from cryogenically stabilized conjugated double bonds within a protein (Lukk *et al.*, 2016). However, an instrument with these capabilities has only been implemented at one beamline (Madden *et al.*, 2013; Newman *et al.*, 2016), limiting widespread access.

X-ray rastering has found the widest use as it requires no additional hardware other than that already in place for diffraction analysis (Cherezov *et al.*, 2009; Hilgart *et al.*, 2011; Song *et al.*, 2007; Stepanov, Hilgart *et al.*, 2011; Aishima *et al.*, 2010). With the emergence of high-speed direct-detection array sensors, raster scanning can be performed in reasonable timeframes for manual and automated crystal positioning (Broennimann *et al.*, 2006). In this method, diffraction is used to identify the crystal positions based on the detection of Bragg-like peaks in the scattering patterns. However, this additional X-ray exposure prior to data collection has the potential to contribute to crystal damage (Dettmar *et al.*, 2015). X-ray exposure can produce both specific damage (*e.g.* to disulfide bonds) (Burmeister, 2000; Holton, 2009; Nave & Garman, 2005) and global loss in diffraction power (Holton, 2009; Garman, 2010). Although X-ray rastering can be performed on smaller crystals it requires a beam with a higher flux and smaller diameter, increasing both the overall measurement time and the X-ray exposure to the crystals (Sanishvili *et al.*, 2011). In principle, one could avoid such complications by solving the structures directly from the data acquired from raster scanning over many crystals at fixed orientations. Even with a single orientation per diffraction pattern, data can be merged from the pool to recover protein structure, as is now regularly carried out in diffraction measurements using X-ray free-electron laser sources (Schlichting, 2015; Martin-Garcia *et al.*, 2016). Alternative sampling strategies with synchrotron sources span between the extreme cases of one orientation each with many crystals to full data sets on single crystals. However, in practice X-ray damage typically extends significantly beyond the exposed region of targeted analysis under cryogenic conditions, such that samples in neighboring voxels can still suffer damage prior to analysis for all these sampling strategies.

The advantages of reduced total X-ray exposure prior to data collection are even more pronounced in room-temperature data collection. In these cases, free radicals produced by photoelectrons can migrate over much longer distances than in measurements under cryogenic conditions. Exposures of regions void of proteins still produces radicals that can result in loss of protein integrity and diffraction resolution, even when the exposures are tens of micrometers or more away from the protein crystal locations (Warkentin *et al.*, 2013).

An adaptation of X-ray rastering is proposed here, in which dynamic sampling greatly reduces the X-ray exposure of the crystals prior to data collection. In brief, the set of preceding localized diffraction measurements are used to select the next most informative location for diffraction scanning image reconstruction, such that crystals can be localized with a much smaller net X-ray exposure of the sample. Ground-truth diffraction data were obtained for a 5 µm-diameter beam and were in excellent agreement with higher-resolution positioning measurements carried out using TPEF and SHG. The supervised learning approach for dynamic sampling (SLADS) was assessed for crystal localization to reduce both the total dose to the sample and the specific dose to the crystals. The anticipated reductions in exposure were also assessed for measurements with a 1 µm beam diameter using TPEF images as surrogates for XRD with a 1 µm-diameter beam. SLADS was then implemented, without ground-truth images, into a beamline at Argonne National Laboratory.

2. Experimental methods

Full length mCherry was cloned into pGEX6P1 and transformed into Rosetta cells using standard cloning protocols. Cells were grown to an optical density of 0.4–0.6 and induced by the addition of 200 µ*M* IPTG, at which point the temperature was decreased to 18°C for 16–18 h. The cultures were harvested by centrifugation and lysed *via* a French press. Resulting lysates were then cleared by centrifuging at 100000*g* for 1 h and the protein purified following standard GST purification protocols. The sample was further purified by size-exclusion chromatography and concentrated to 20 mg ml^{-1} to be used in crystallization experiments. The crystals used in these studies were grown using both sitting drop and hanging drop vapor diffusion methods in mother liquor containing

100 m*M* Tris pH 8.0, 100 m*M* sodium acetate and 30% PEG 4000 at room temperature, as has been previously reported (Shu *et al.*, 2006). The crystals grew over the course of 1–4 d and formed large clusters of rod-shaped crystals, which would eventually be broken apart for individual experiments.

Lysozyme crystals were grown using a Hampton lysozyme kit; 20 mg of lysozyme was solubilized in 1 ml of 0.02 *M* sodium acetate trihydrate, pH 4.6 (HR7-108). Crystals were grown by hanging drop vapor diffusion using the reagent 30% (*w*/*v*) polyethylene glycol monomethyl ether 5000, 1.0 *M* sodium chloride and 50 m*M* sodium acetate trihydrate pH 4.5 (HR2-805).

XRD raster scan images were acquired for mCherry crystals looped and cryo-cooled. XRD raster images were acquired using the 5 µm minibeam collimator with a 5 µm × 5 µm cell size. Total X-ray exposure time was 1 s unattenuated integrating over a 1° rotation of the sample during the raster scan acquisition. In total, 3200 cells were interrogated corresponding to an area of 200 µm × 400 µm over a period of 4 h. Dynamic sampling of the lysozyme crystals was performed using a 5 µm collimator with a total X-ray exposure of 0.1 s per sampled pixel with 20× attenuation and no integration over rotation angles. During SLADS data acquisition, calculation of the optimal position for subsequent sampling was completed in approximately 1 ms. Random access time per pixel averaged 1 s.

Diffraction raster in *JBluIce* (Stepanov, Makarov *et al.*, 2011) currently employs the program *DISTL* (Zhang *et al.*, 2006). *DISTL* is a part of the package *LABELIT* (Sauter *et al.*, 2004), which estimates potential Bragg candidates. There are three steps involved: (i) isolating diffraction-like peaks from the background in a diffraction image considering the noise variability in the local environment; (ii) validating the isolated peaks from the rejection of possible sources of ice-rings, salt particles or crystal disorder, and (iii) gauging size and shapes of each peak. *DISTL* estimates diffraction peaks more quickly than full-blown indexing and processing of diffraction data [normally performed with programs such as *XDS* (Diederichs, 2006; Kabsch, 2010), *MOSFLM* (Leslie & Powell, 2007), *HKL2000* (Minor *et al.*, 2000)].

The nonlinear optical images were acquired using an integrated multi-modal nonlinear optical microscope system at The National Institute of General Medical Sciences and National Cancer Institute Structural Biology Facility beamline 23-ID-B at the Advanced Photon Source (GM/CA@APS) as described previously (Madden *et al.*, 2013; Newman *et al.*, 2016). A Fianium FemtoPower 1060 ultrafast fiber laser was utilized to generate ∼160 fs pulses centered at ∼1060 nm, with a 50 MHz repetition rate. The maximum laser power sent to the sample was ∼90 mW. SHG, TPE-UVF, TPEF and fluorescence signals were collected with a 25 mm lens and then separated from the fundamental by a dichroic mirror. The TPEF and fluorescence signals were separated from the SHG and TPE-UVF by a second dichroic mirror and detected using a photomultiplier tube (PMT). The SHG and TPE-UVF were then separated from each other by a third dichroic mirror and detected by two separate PMTs. The 512 × 512 SHG, TPE-UVF, TPEF, fluorescence and bright-field images were acquired concurrently using a PCI Express Digitizer (ATS9440, AlazarTech).

3. Theoretical methods

In dynamic sampling a new measurement location is selected using previous measurements. The different dynamic sampling methods in the literature differ primarily in the definition of a measurement and in the criteria used for measurement selection. For this application, dynamic sampling for XRD, the supervised learning approach for dynamic sampling (SLADS) algorithm, presented by Godaliyadda *et al.* (2016) and illustrated in Fig. 1, was used.

SLADS was selected primarily because it is designed for single-pixel measurements and has shown potential for measurement selections on application similar to XRD such as electron backscatter diffraction (Godaliyadda *et al.*, 2016). In the following simulations, reconstructions with normalized distributions, defined in equation (9), below 4×10^{-3} were achieved by acquiring just 7–30% of all available measurements within the image. Furthermore, SLADS can select a new location to measure in 1–50 ms a practical timescale for high-throughput XRD raster scanning. The theory of the SLADS algorithm follows.

To explain the formulation of SLADS, the underlying object is denoted as $X \in \mathbb{R}^N$. Here, N is the number of pixels in X. Now assume that k measurements have already been

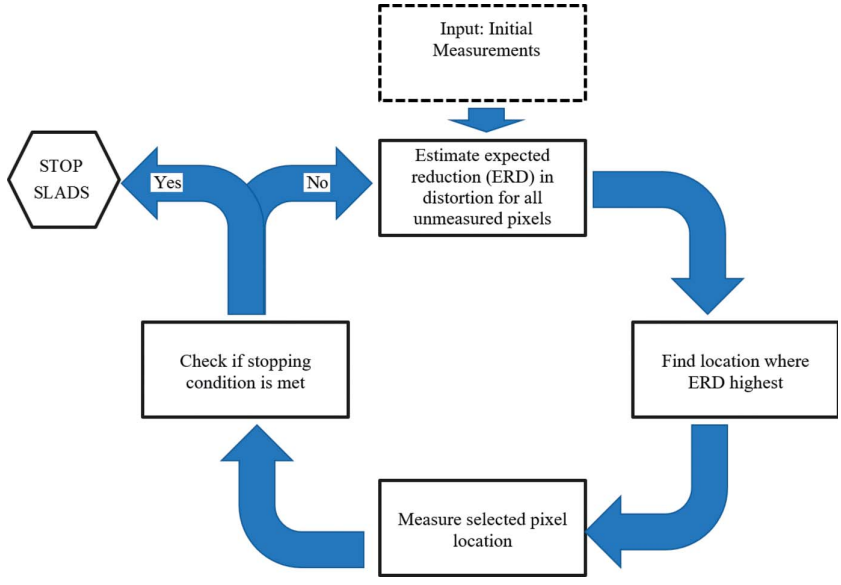


Figure 1
Illustration of the SLADS algorithm. The inputs to the function are initial measurements and the parameters from training. SLADS runs until a predefined stopping condition is met.

acquired at a set of locations $S = \{s^{(1)}, s^{(2)}, \ldots, s^{(k)}\}$. These measurements are an $N \times 2$ matrix $Y^{(k)}$,

$$Y^{(k)} = \begin{bmatrix} s^{(1)}, X_{s^{(1)}} \\ \vdots \\ s^{(k)}, X_{s^{(k)}} \end{bmatrix}, \quad (1)$$

where $X_{s^{(i)}}$ refers to the value of pixel location $s^{(i)}$. The goal is to find the location $s^{(k+1)}$ that most reduces the reconstruction distortion. This reconstruction distortion is a value quantifying the difference between the underlying image and the reconstructed image, *e.g.* the number of pixels in the reconstruction that do not match their corresponding pixels in the underlying image. The distortion between the images X and $\hat{X}^{(k)}$, the image reconstructed using $Y^{(k)}$, is defined as $D(X, \hat{X})$, where

$$D(X, \hat{X}) = \sum_{r \in \Omega} D(X_r, \hat{X}_r). \quad (2)$$

For this particular problem where there is a binary image, let

$$D(X_r, \hat{X}_r) = \begin{cases} 0 & \text{if } X_r = \hat{X}_r, \\ 1 & \text{if } X_r \neq \hat{X}_r. \end{cases} \quad (3)$$

Assuming another measured pixel s, then presumably $\hat{X}^{(k;s)}$ which is the reconstruction performed using the measurement X_s and $Y^{(k)}$ is a better estimate of X when compared with $\hat{X}^{(k)}$. Hence, the reduction in distortion can be defined after the location s is measured as

$$R^{(k;s)} = D\left(X, \hat{X}^{(k)}\right) - D\left(X, \hat{X}^{(k;s)}\right). \quad (4)$$

However, as the underlying image, X, is unknown during image acquisition, the expected reduction in distortion (ERD), $\bar{R}^{(k;s)}$, for every pixel can be computed, which is given by

$$\bar{R}^{(k;s)} = E\left[D\left(X, \hat{X}^{(k)}\right) - D\left(X, \hat{X}^{(k;s)}\right) \middle| Y^{(k)}\right]. \quad (5)$$

Then the location of the next measurement $s^{(k+1)}$ is given by

$$s^{(k+1)} = \underset{s \in \{\Omega \setminus S\}}{\arg\max} \left\{ E\left[D\left(X, \hat{X}^{(k)}\right) - D\left(X, \hat{X}^{(k;s)}\right) \middle| Y^{(k)}\right]\right\}, \quad (6)$$

where Ω is the set of all locations in the image. The pixel which corresponds to the largest ERD is then sampled.

Now in order to compute the ERD during dynamic sampling, a relationship between the measurements and the ERD must be found. In SLADS, this relationship is a simple regression function that is computed using an offline training process. In order to reduce computation time during training the reduction in distortion is approximated by

$$R^{(s)} \approx \sum_{r \in \Omega} h_r^{(s)} D(X_r, \hat{X}_r). \quad (7)$$

Here $h_r^{(s)}$ is defined as

$$h_r^{(s)} = \exp\left\{-\frac{c}{2(\sigma^{(s)})^2}\|r - s\|^2\right\}, \quad (8)$$

where $\sigma^{(s)} = \min_{t \in S}\{\|s - t\|^2\}$. The reasons behind choosing this particular approximation and the need for it are also detailed by Godaliyadda *et al.* (2016). In training, a linear relation between R^2 and $Y^{(k)}$ is found and this learned relation is used to compute the ERD. The ERD can then be computed in real time during dynamic sampling and, as a result, a new sampling location can be found in 1–10 ms for the measurements herein using a 5 µm-diameter X-ray beam.

The SLADS algorithm also incorporates a stopping condition that allows sampling to stop when a desired reduction in distortion has been achieved. If the underlying image is known, dynamic sampling can be stopped when the normalized distortion (ND) is below a threshold T, *i.e.* when

$$\frac{1}{|\Omega|} D\left(X, \hat{X}^{(k)}\right) \leq T. \quad (9)$$

However, as the image is unknown, the ND cannot be computed. Therefore, the recursion computed shown below at each step of SLADS and a threshold were set instead,

$$\varepsilon^{(k)} = (1 - \beta)\varepsilon^{(k-1)} + \beta D\left(X_{s^{(k)}}, \hat{X}_{s^{(k)}}^{(k-1)}\right) \leq \tilde{T}(T). \quad (10)$$

This threshold $\tilde{T}(T)$ was computed during training so as to correspond to a desired normalized distortion level T. The procedure to find this threshold $\tilde{T}(T)$ is also described by Godaliyadda *et al.* (2016).

4. Results and discussion

Dynamic sampling experiments were performed directly on ground-truth X-ray diffraction data as well as nonlinear optical images serving as high-resolution surrogates for crystal position. Analyses are included both for a 5 µm-diameter beam in §4.1 consistent with the ground-truth diffraction data, as well as for simulations corresponding to a 1 µm-diameter beam using the nonlinear optical measurements as surrogates for diffraction images in §4.2. When implemented into the beamline at Argonne National Laboratory, SLADS ran without a ground truth, §4.3. A comparison of SLADS with conventional approaches can be found in §4.4.

4.1. SLADS experiment on 5 µm XRD images

In the previous section, it was mentioned that SLADS requires training using representative images with known positions. As the final reconstructions are not particularly sensitive to the nature of the training images, surrogates were created for the XRD images by modifying high-resolution TPEF measurements of protein crystals [Figs. 2(*a*) and 2(*c*)]. The TPEF images for training were acquired at a resolution of 1 µm, but the ground-truth image had a resolution of 5 µm. Therefore, first the 1 µm resolution image was downsampled by a factor of five to create an image that corresponds to measurements made at 5 µm resolution. In order to correct for aliasing artefacts that result from direct decimation the image was smoothed using a low-pass filter and converted to a binary image resulting in Fig. 2(*b*). The SLADS algorithm also requires more images to find the stopping condition. For this purpose, the same procedure was implemented on the image

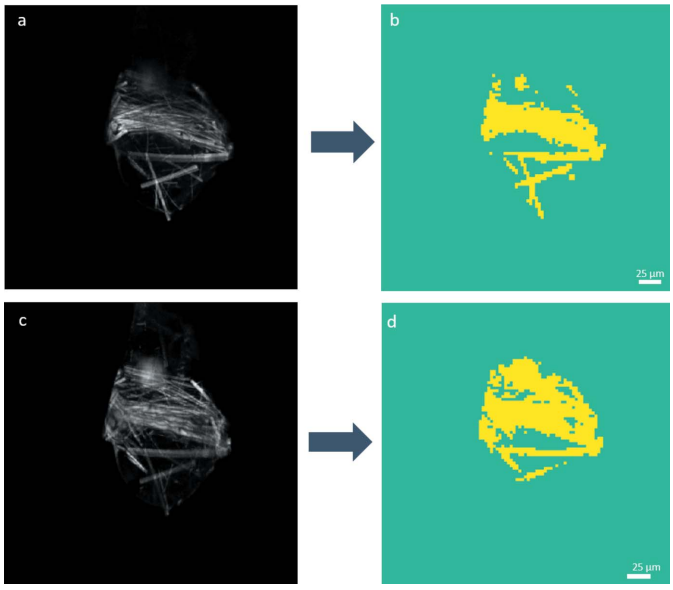


Figure 2
(*a*, *c*) TPEF of mCherry crystal acquired at 1 µm resolution. (*b*, *d*) Synthetic binary XRD image created by thresholding, filtering the TPEF image using a Gaussian kernel and downsampling the image five times. A training database for the algorithm was created with (*b*), whereas (*d*) was used to determine a threshold to stop SLADS.

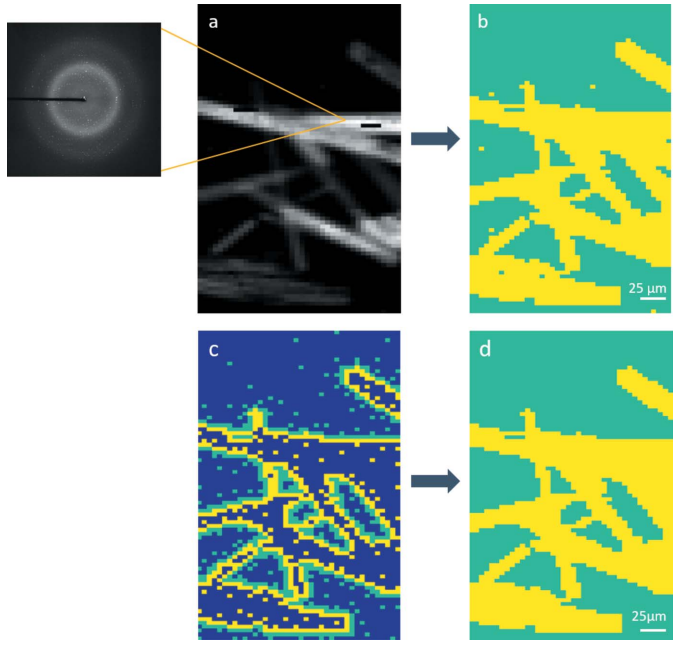


Figure 3
(*a*) XRD image of mCherry crystals with accompanying diffraction pattern. (*b*) Binary-image ground-truth imageconstructed by setting a threshold of 5×10^5. (*c*) Location and measurements that were acquired; green = background; yellow = crystal; 30.8% of the image was sampled and 9.03% of the interior of the crystal was measured. (*d*) Image reconstructed from SLADS measurements in (*c*). The ND between the ground truth in (*b*) and reconstructed image in (*d*) was 4×10^{-3}.

shown in Fig. 2(*c*) to create another image that corresponds to measurements made at 5 µm resolution (Fig. 2*d*).

For crystal positioning prior to diffraction data collection, the measurement objective was squarely focused on locating the presence or absence of a crystal, such that the training images were converted into binary images corresponding to the presence or absence of protein-like diffraction. For this purpose, the intensity values of the downsampled TPEF images used for training were rescaled to an 8-bit intensity range and binarized based on threshold selection using Otsu's method (Sezgin, 2004). This approach maximizes the inter-class variance and minimizes the intra-class variance between the two classes in each image (*i.e.* class 0, crystal absent; class 1, crystal present). Next, the average between them was computed to determine the threshold that, when applied to both downsampled images, created the binarized images, Figs. 2(*b*) and 2(*d*). Fig. 2(*b*) was used to create the training database and Fig. 2(*d*) to find the threshold to stop SLADS. This same relative threshold (rescaled to the dynamic range of the diffraction measurements) was implemented to binarize the XRD image used for testing (*i.e.* ground truth).

A known 'ground truth' diffraction image was obtained (Fig. 3) for assessing the SLADS algorithm for crystal positioning. The grayscale brightness in the XRD map in Fig. 3(*a*) is proportional to the number of diffraction-like peaks identified in the original X-ray scattering pattern. With the aim of using SLADS to identify crystal position, the grayscale image was converted to a binary map based on a threshold imposed on the peak counts. The threshold was computed using training data (explained later in this section). The resulting binary image (Fig. 3*b*) consisted of two labels: label 1 (yellow) for pixels where sample was present and label 0 (green) for pixels where sample was absent. In this particular experiment the threshold was 1.6×10^5 counts, total integrated signal for *DISTL*, and Fig. 3(*b*) was used as the ground truth.

For the training phase of this experiment, the smoothing parameter c in equation (8) was empirically determined to be 8 and the weighted-mode interpolation method described by Godaliyadda *et al.* (2016) was used for all reconstructions. To compute the stopping condition in equation (10), $\beta = 0.006$ and the threshold in equation (10) corresponded to an ND [defined in equation (9)] of 2×10^{-3}.

The initial measurement mask was generated using the haltonset function in MATLAB resulting in a 1% random sampling of the ground-truth image. Then, the location of each subsequent measurement was determined according to the SLADS algorithm also run through MATLAB. The sampling procedure continued until the stopping condition was met. The results of the experiment are shown in Figs. 3(*c*) and 3(*d*). These images correspond to when the stopping condition was met, which in this experiment was when 30.8% of the image was sampled. Fig. 3(*c*) shows the locations and measurements that were acquired (yellow for a crystal pixel and green for a background pixel), Fig. 3(*d*) shows the image reconstructed using the measurements in Fig. 3(*c*).

As a measure of quality, the ND was calculated, defined in equation (9), between the ground truth, Fig. 3(*b*), and the reconstructed image, Fig. 3(*d*). The ND when SLADS stopped was 4×10^{-3} (0–1). The primary goal was to minimize the X-ray dosage and exposure time experienced by the crystal. To quantify exposure, the percentage of measurements made

within the sample (excluding the boundaries) was computed and found to be 9.0%; thereby, limiting the area of the crystal exposed to potentially harmful radiation before analysis.

4.2. SLADS experiment on 1 μm simulated XRD images

SLADS can also be implemented on higher-resolution measurement schemes. First, the algorithm was trained on similar images, once again using TPEF images as surrogates for diffraction. To construct the training database, the image shown in Fig. 4(*b*) was used which was created by thresholding the image in Fig. 4(*a*). Again the weighted mode interpolation method was used for all reconstructions and set $c = 8$ in equation (8). Then to find the threshold to stop SLADS, the image shown in Fig. 4(*d*) was used. This image was created by thresholding the image shown in Fig. 4(*c*). To compute the stopping condition β was set to 0.001. This β value is different from the previous experiment. The reason for this discrepancy is that the value of β is chosen according to the number of pixels in the image with larger images requiring a smaller β.

Due to the large amount of time it takes to acquire a full high-resolution (pixel size ∼1 μm) XRD image and the high value of beam time at facilities delivering 1 μm high-flux beams, a simulated 'ground-truth' XRD image was created using an already available high-resolution TPEF image with similar spatial resolution. The original TPEF image is shown in Fig. 5(*a*) and the simulated image created by thresholding this image is shown in Fig. 5(*b*). Once more, the threshold for creating the binary image was carried out by applying Otsu's method on the training images.

The results of the SLADS sampling are shown in Figs. 5(*c*) and 5(*d*). Fig. 5(*c*) shows the locations and measurements that were acquired, Fig. 5(*d*) displays the image reconstructed using the measurements in Fig. 5(*c*). In this experiment,

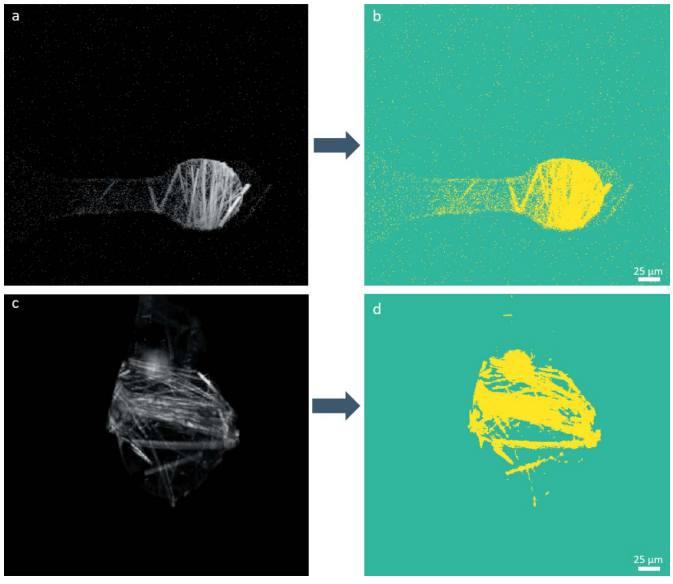


Figure 4
(*a*) TPEF image of mCherry crystals, (*b*) thresholded image for use as training set, (*c*) TPEF of mCherry crystals and (*d*) thresholded image to find stopping condition.

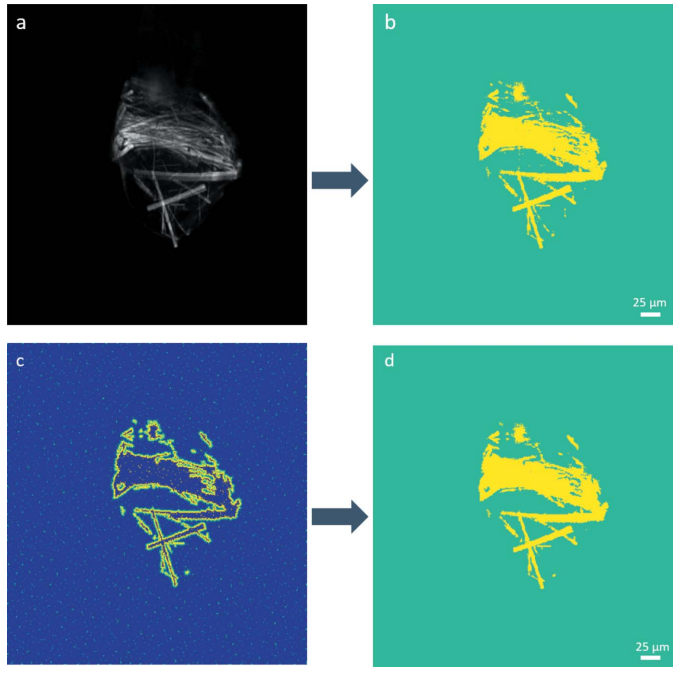


Figure 5
(*a*) TPEF image of mCherry crystals. (*b*) Synthetic image created by thresholding the image in (*a*) and used as ground truth for reconstruction. (*c*) Location and measurements that were acquired; green = background; yellow = crystal; 6.7% of the image was sampled and 11.5% of the interior of crystal was measured. (*d*) Image reconstructed from SLADS measurements in (*c*). The ND between the ground truth in (*b*) and reconstructed image in (*d*) was 1.7×10^{-3}.

SLADS stopped when just 6.7% of the image was sampled and the ND between the reconstructed image, Fig. 5(*d*), and ground truth, Fig. 5(*b*), was approximately 1.7×10^{-3}. Furthermore, only 11.5% of the interior of the crystal was sampled; therefore, limiting the area of the crystal exposed to X-rays before analysis.

4.3. Proof of concept: SLADS implementation at Argonne National Laboratory

In the previous two sections, simulations of dynamic sampling were performed on images that were already collected to demonstrate proof of concept. SLADS was then incorporated into the software at Argonne National Laboratory and run without ground-truth images.

Two experiments were performed: the same lysozyme crystal was used and the loop imaged at two positions defined at 0° and 90°. Both experiments were initialized by sampling first 1% of the image using low-discrepancy (pseudo sequence) sampling, then SLADS took over until 3% of the image was sampled.

The results are shown in Fig. 6 and the acquired videos with overlaid reconstruction (carried out after the acquisition) are shown in Videos S1 and S2 of the supporting information, whereas a bright-field image of the looped crystal is shown in Fig. S1. Figs. 6(*a*) and 6(*b*) shows the results for 0° orientation and Figs. 6(*c*) and 6(*d*) when the sample is rotated by 90°. Figs. 6(*a*) and 6(*c*) show the measured images, yellow for a

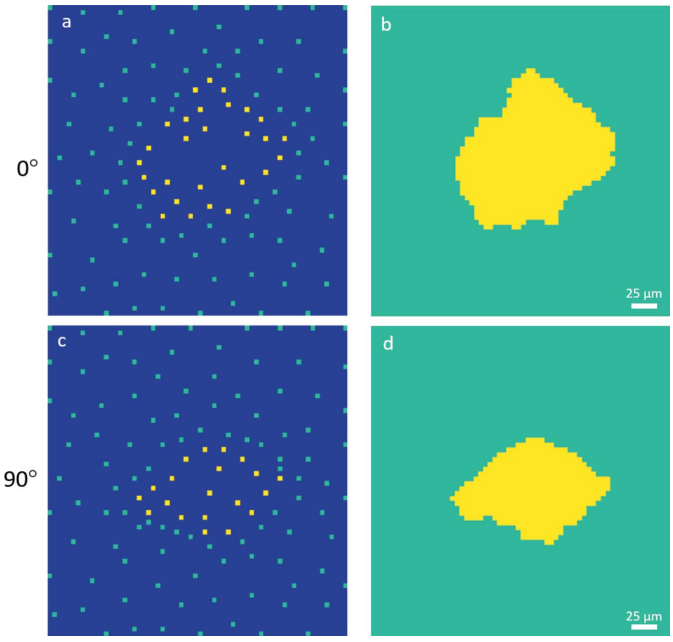


Figure 6
SLADS implementation at Argonne National Laboratory with lysozyme. (*a*) Measured locations for the 0° rotation acquired; green = background; yellow = crystal; 3% of the image was sampled and approximately 5% of the interior of crystal was measured. (*b*) Generated reconstruction from (*a*). (*c*) Measured locations for the 90° rotation acquired; green = background; yellow = crystal; 3% of the image was sampled and approximately 5% of the interior of crystal was measured. (*d*) Generated reconstruction from (*c*).

crystal and green for no crystal, and Figs. 6(*b*) and 6(*d*) show the reconstructed images. From these figures it is clear that the positions of the crystals can be found by sampling just 3% of the total image and approximately 5% of the crystal.

4.4. Comparison of SLADS and conventional approaches

In terms of X-ray exposure prior to data collection, SLADS significantly decreases dosage when compared with raster scanning; where raster scanning exposes 100% of the sample to X-rays, SLADS resulted in only 9.0% of the crystal being sampled using the 5 µm-diameter beam in simulations and only approximately 5% of the crystal sampled in real systems. Furthermore, those pixels were predominately located on the edges leaving the central areas of the crystal pristine and available for data acquisition. Additionally, as the resolution increases, the potential advantages of SLADS correspondingly remains. In the simulations corresponding to a 1 µm-diameter beam, the fraction of sampled pixels required to reliably locate a crystal was only 11.5%.

From the preceding analysis, it should be evident that there is clearly an interplay between the diameter of the beam relative to the size and the aspect ratio of the crystal. For the 5 µm-diameter beam measurements, the widths of the crystal were on average ∼5× larger than the beam width. In contrast, the short axes of the crystals were 25-fold greater than the beam width assuming a 1 µm-diameter beam. When the crystal size approaches the dimensions of the X-ray beam, the advantages of dynamic sampling are reduced. In the limit of crystals comparable or smaller than the X-ray beam, dynamic sampling provides no benefit as each crystal would comprise a single pixel. Provided crystals are significantly larger than the beam, it is clear that the smallest dose to the central portions of the crystals are achieved with the smallest diameter X-ray beams. It should also be clear from simple geometric arguments that the exposed fraction of the crystals would be higher for crystals with high aspect ratios.

In practice, the potential benefits of dynamic sampling may be offset in part by technical challenges associated with rapid, random access sampling. With improvements in detectors, data collection can proceed in as little as a few milliseconds per frame, whereas goniometers may require 10 ms to 100 ms to perform random access positioning. In addition, the timeframe for image transfer and analysis for diffraction-like peaks, along with the calculations to perform the dynamic sampling, can increase the overall positioning time relative to the theoretical limit. However, many of these same constraints still hold for conventional raster scanning at the majority of beamlines integrating raster scanning for crystal positioning. In this context, the advantages associated with reduction in X-ray exposure prior to data collection may often offset any increases in measurement and/or analysis time for crystal positioning, particularly for X-ray labile samples and/or room temperature data collection.

Several figures of merit are worth noting when comparing the proposed approach with alternative strategies for crystal positioning. Whereas SHG and TPE-UVF imaging methods are faster, have high resolution and do not expose the sample to any X-ray damage, the number of beamlines currently equipped with such capabilities is limited to one. Furthermore, there is a large diversity in SHG activity depending on the symmetry of the protein crystal, as well as variability in intrinsic TPE-UVF from the requirement of aromatic amino acids.

5. Conclusions and future work

The incorporation of the SLADS approach for XRD raster scanning was demonstrated for protein crystal positioning with significant decreases in the X-ray dosage experienced by the crystal. SLADS was found to reduce the exposure experienced by model protein crystals by a factor of 11 for the interior fraction when using a 5 µm-diameter X-ray beam in simulations to position ∼25 µm-diameter crystals. Good agreement in crystal position was observed between reconstructions and ground-truth X-ray diffraction measurements. Implementation of SLADS at GM/CA allowed automatic identification of crystal position with an approximately 20-fold reduction in dose relative to a conventional raster scan. The crystal positions recovered by SLADS were also in excellent agreement with locations determined independently by SHG and TPEF microscopy measurements performed prior to X-ray exposure. For beamlines not equipped with such nonlinear optical imaging capabilities, SLADS provides an alternative that is directly compatible with end-stations

currently capable of performing raster-based crystal positioning.

Future extensions of the SLADS approach may further expand upon the capabilities of the measurements. All the data acquired in the present studies were collected under cryogenic conditions, in which the extent of X-ray radiation damage is typically limited to distances only a few micrometers from the exposed locations. For room-temperature data collection, diffusion allows for substantially greater distances over which exposure results in damage. The observed reduction in exposure to protein-free regions of SLADS may be even more advantageous at room temperature for this reason. In addition, the SLADS approach, as demonstrated herein, based on binary image reconstruction provides the location of the crystal, but the quality of diffraction may vary depending on location within each single-crystal. In future implementations, grayscale reconstruction could be performed following crystal positioning in order to identify likely locations of quality diffractions.

Acknowledgements

NMS, SZ, JAN, AUC, MJS, CD and GJS gratefully acknowledge support from the NIH grant Nos. R01GM-103910 and R01GM-103410. DG, DHY and CB gratefully acknowledge support from AFOSR/MURI grant No. FA9550-12-1-0458 and AFRL/RX Contract Number FA8650-10-D-5201-0038. GM/CA@APS has been funded in whole or in part with Federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006). This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357.

References

Aishima, J., Owen, R. L., Axford, D., Shepherd, E., Winter, G., Levik, K., Gibbons, P., Ashton, A. & Evans, G. (2010). *Acta Cryst.* D**66**, 1032–1035.

Andrey, P., Lavault, B., Cipriani, F. & Maurin, Y. (2004). *J. Appl. Cryst.* **37**, 265–269.

Broennimann, Ch., Eikenberry, E. F., Henrich, B., Horisberger, R., Huelsen, G., Pohl, E., Schmitt, B., Schulze-Briese, C., Suzuki, M., Tomizaki, T., Toyokawa, H. & Wagner, A. (2006). *J. Synchrotron Rad.* **13**, 120–130.

Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.

Cherezov, V., Hanson, M. A., Griffith, M. T., Hilgart, M. C., Sanishvili, R., Nagarajan, V., Stepanov, S., Fischetti, R. F., Kuhn, P. & Stevens, R. C. (2009). *J. R. Soc. Interface*, **6**, S587–S597.

Dettmar, C. M., Newman, J. A., Toth, S. J., Becker, M., Fischetti, R. F. & Simpson, G. J. (2015). *Proc. Natl Acad. Sci. USA*, **112**, 696–701.

Diederichs, K. (2006). *Acta Cryst.* D**62**, 96–101.

Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.

Godaliyadda, G. M. D., Ye, D. H., Uchic, M. D., Groeber, M. A., Buzzard, G. T. & Bouman, C. A. (2016). *Electron. Imaging*, **2016**, 1–8.

Hilgart, M. C., Sanishvili, R., Ogata, C. M., Becker, M., Venugopalan, N., Stepanov, S., Makarov, O., Smith, J. L. & Fischetti, R. F. (2011). *J. Synchrotron Rad.* **18**, 717–722.

Holton, J. M. (2009). *J. Synchrotron Rad.* **16**, 133–142.

Jain, A. & Stojanoff, V. (2007). *J. Synchrotron Rad.* **14**, 355–360.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Kissick, D. J., Gualtieri, E. J., Simpson, G. J. & Cherezov, V. (2010). *Anal. Chem.* **82**, 491–497.

Kissick, D. J., Wanapun, D. & Simpson, G. J. (2011). *Annu. Rev. Anal. Chem.* **4**, 419–437.

Leslie, A. G. & Powell, H. R. (2007). *Evolving Methods for Macromolecular Crystallography*, pp. 41–51. Berlin: Springer.

Lukk, T., Gillilan, R. E., Szebenyi, D. M. E. & Zipfel, W. R. (2016). *J. Appl. Cryst.* **49**, 234–240.

Madden, J. T., DeWalt, E. L. & Simpson, G. J. (2011). *Acta Cryst.* D**67**, 839–846.

Madden, J. T., Toth, S. J., Dettmar, C. M., Newman, J. A., Oglesbee, R. A., Hedderich, H. G., Everly, R. M., Becker, M., Ronau, J. A., Buchanan, S. K., Cherezov, V., Morrow, M. E., Xu, S., Ferguson, D., Makarov, O., Das, C., Fischetti, R. & Simpson, G. J. (2013). *J. Synchrotron Rad.* **20**, 531–540.

Martin-Garcia, J. M., Conrad, C. E., Coe, J., Roy-Chowdhury, S. & Fromme, P. (2016). *Arch. Biochem. Biophys.* **602**, 32–47.

Minor, W., Tomchick, D. & Otwinowski, Z. (2000). *Structure*, **8**, R105–R110.

Moukhametzianov, R., Burghammer, M., Edwards, P. C., Petit-demange, S., Popov, D., Fransen, M., McMullan, G., Schertler, G. F. X. & Riekel, C. (2008). *Acta Cryst.* D**64**, 158–166.

Nave, C. & Garman, E. F. (2005). *J. Synchrotron Rad.* **12**, 257–260.

Newman, J. A., Zhang, S., Sullivan, S. Z., Dow, X. Y., Becker, M., Sheedlo, M. J., Stepanov, S., Carlsen, M. S., Everly, R. M., Das, C., Fischetti, R. F. & Simpson, G. J. (2016). *J. Synchrotron Rad.* **23**, 959–965.

Padayatti, P., Palczewska, G., Sun, W., Palczewski, K. & Salom, D. (2012). *Biochemistry*, **51**, 1625–1637.

Pohl, E., Ristau, U., Gehrmann, T., Jahn, D., Robrahn, B., Malthan, D., Dobler, H. & Hermes, C. (2004). *J. Synchrotron Rad.* **11**, 372–377.

Pothineni, S. B., Strutz, T. & Lamzin, V. S. (2006). *Acta Cryst.* D**62**, 1358–1368.

Sanishvili, R., Yoder, D. W., Pothineni, S. B., Rosenbaum, G., Xu, S., Vogt, S., Stepanov, S., Makarov, O. A., Corcoran, S., Benn, R., Nagarajan, V., Smith, J. L. & Fischetti, R. F. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 6127–6132.

Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.

Schlichting, I. (2015). *IUCrJ*, **2**, 246–255.

Sezgin, M. (2004). *J. Electron. Imaging*, **13**, 146–168.

Shu, X., Shaner, N. C., Yarbrough, C. A., Tsien, R. Y. & Remington, S. J. (2006). *Biochemistry*, **45**, 9639–9647.

Song, J., Mathew, D., Jacob, S. A., Corbett, L., Moorhead, P. & Soltis, S. M. (2007). *J. Synchrotron Rad.* **14**, 191–195.

Stepanov, S., Hilgart, M., Yoder, D. W., Makarov, O., Becker, M., Sanishvili, R., Ogata, C. M., Venugopalan, N., Aragão, D., Caffrey, M., Smith, J. L. & Fischetti, R. F. (2011). *J. Appl. Cryst.* **44**, 772–778.

Stepanov, S., Makarov, O., Hilgart, M., Pothineni, S. B., Urakhchin, A., Devarapalli, S., Yoder, D., Becker, M., Ogata, C., Sanishvili, R., Venugopalan, N., Smith, J. L. & Fischetti, R. F. (2011). *Acta Cryst.* D**67**, 176–188.

Vernede, X., Lavault, B., Ohana, J., Nurizzo, D., Joly, J., Jacquamet, L., Felisaz, F., Cipriani, F. & Bourgeois, D. (2006). *Acta Cryst.* D**62**, 253–261.

Warkentin, M., Hopkins, J. B., Badeau, R., Mulichak, A. M., Keefe, L. J. & Thorne, R. E. (2013). *J. Synchrotron Rad.* **20**, 7–13.

Zhang, Z., Sauter, N. K., van den Bedem, H., Snell, G. & Deacon, A. M. (2006). *J. Appl. Cryst.* **39**, 112–119.