JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

Reconstructing three-dimensional protein crystal intensities from sparse unoriented two-axis X-ray diffraction patterns

Ti-Yen Lan,[a] Jennifer L. Wierman,[b,c] Mark W. Tate,[a] Hugh T. Philipp,[a] Veit Elser[a] and Sol M. Gruner[a,b,c,d]*

[a]Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853, USA, [b]Cornell High Energy Synchrotron Source (CHESS), Cornell University, Ithaca, NY 14853, USA, [c]Macromolecular Diffraction Facility at CHESS (MacCHESS), Cornell University, Ithaca, NY 14853, USA, and [d]Kavli Institute for Nanoscale Science, Cornell University, Ithaca, NY 14853, USA. *Correspondence e-mail: smg26@cornell.edu

Received 6 February 2017
Accepted 1 May 2017

Edited by A. Barty, DESY, Hamburg, Germany

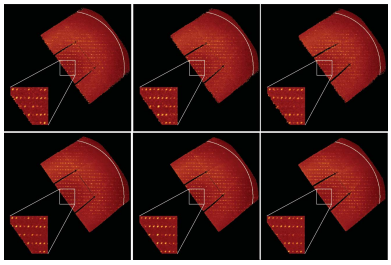
Keywords: X-ray serial microcrystallography; sparse data; EMC algorithm; protein microcrystallography; synchrotron radiation sources.

Recently, there has been a growing interest in adapting serial microcrystallography (SMX) experiments to existing storage ring (SR) sources. For very small crystals, however, radiation damage occurs before sufficient numbers of photons are diffracted to determine the orientation of the crystal. The challenge is to merge data from a large number of such 'sparse' frames in order to measure the full reciprocal space intensity. To simulate sparse frames, a dataset was collected from a large lysozyme crystal illuminated by a dim X-ray source. The crystal was continuously rotated about two orthogonal axes to sample a subset of the rotation space. With the EMC algorithm [expand–maximize–compress; Loh & Elser (2009). *Phys. Rev. E*, **80**, 026705], it is shown that the diffracted intensity of the crystal can still be reconstructed even without knowledge of the orientation of the crystal in any sparse frame. Moreover, parallel computation implementations were designed to considerably improve the time and memory scaling of the algorithm. The results show that EMC-based SMX experiments should be feasible at SR sources.

1. Introduction

The advance of serial femtosecond microcrystallography (SFX) at X-ray free-electron lasers (XFELs) (Chapman *et al.*, 2011; Boutet *et al.*, 2012) allows structure determination with protein crystals whose sizes are too small for conventional crystallography experiments. SFX is based on injecting a sequence of randomly oriented microcrystals to intercept a train of X-ray pulses. The tens of femtoseconds long pulse width enables the photon scattering process to outrun the radiation damage of the crystals, while the ultra-high brightness of the pulses results in a sufficient number of resolvable Bragg peaks collected by a fast-framing detector (Philipp *et al.*, 2008) for indexing. Using this concept of 'diffract before destroy' (Neutze *et al.*, 2000), a complete dataset can be obtained given enough indexed data frames. SFX has the advantage of rapid data collection owing to the high repetition rate of the X-ray pulses and provides a promising means to study proteins that do not readily form large single crystals.

Despite the success of SFX, the paucity of XFEL beamtime has inspired interest in adapting serial microcrystallography (SMX) experiments to storage ring (SR) sources (Gati *et al.*, 2014; Stellato *et al.*, 2014; Heymann *et al.*, 2014; Nogly *et al.*, 2015; Botha *et al.*, 2015; Gruner & Lattman, 2015; Schubert *et al.*, 2016). Radiation damage cannot be outrun in the same way at SR sources, thereby much more strongly limiting the dose per protein that can be tolerated. The limit to the



smallest usable crystal size in SMX is the ability to index each data frame, since frames collected from very small crystals will have so few diffracted photons that Bragg peaks will not be obviously identifiable. Because successfully indexing a frame usually requires at least 20–30 resolvable Bragg peaks, under conventional processing schemes data frames that are too weak to identify this number of Bragg peaks would be discarded.

Instead of determining the orientation of each data frame individually, the expand–maximize–compress (EMC) algorithm (Loh & Elser, 2009) seeks to reconstruct a consistent three-dimensional intensity model using all the data frames simultaneously. The EMC algorithm treats the orientation of each data frame as a probability distribution conditional on the current model and iteratively updates the model by maximizing the associated likelihood function. The validation of its probabilistic modeling of orientations has been demonstrated in many proof-of-concept experiments (Loh *et al.*, 2010; Philipp *et al.*, 2012; Ayyer *et al.*, 2014, 2015; Ekeberg *et al.*, 2015; Wierman *et al.*, 2016), even in some cases where the number of collected photons per frame is extremely low. This success has motivated us to apply the EMC algorithm to SMX to push the limit of usable crystal sizes.

This study is the latest of a series of proof-of-concept table-top experiments with increasing complexity to test the applicability of the EMC algorithm to SMX. We simulated the data frames collected in an SMX experiment by taking a large volume of data frames of very brief exposures from a large hen egg white lysozyme (HEWL) crystal with a dim laboratory X-ray source and a fast-framing mixed-mode pixel array detector (MM-PAD) (Tate *et al.*, 2013). In contrast to our previous work (Wierman *et al.*, 2016), where the crystal was rotated about a single axis, the data frames used in this study were collected from a crystal rotated about two orthogonal axes continuously to sample a greater portion of the rotation space. The crystal intensities were reconstructed using the discrete three-dimensional rotation samples lying in this rotation subset, so our method applies to randomly oriented frames by replacing the rotation subset with the whole three-dimensional rotation space. However, like the full exploration of the rotation space, this rotation sampling results in a cubic growth with resolution in the memory and time scaling of the EMC algorithm, which makes the two-axis problem more difficult than its single-axis counterpart. To remedy this problem, we developed computing schemes that greatly reduce the memory usage and computation time. With no input information on the orientation of each data frame, the EMC algorithm successfully reconstructed the Bragg reflections to 2.27 Å resolution. This result further paves the way for EMC-based SMX experiments.

This paper is organized as follows. In §2, we present the details of the experiment, an overview of the EMC algorithm and other aspects of the data processing. In particular, we introduce a local update scheme to speed up the EMC algorithm at high resolution. In §3, we examine the sparsity of the data frames input to the EMC algorithm and the results of the intensity reconstruction. In the appendices, we describe a memory-efficient parallel implementation of the EMC algorithm and quantify the speed-up of the local update scheme in practice.

2. Materials and methods

2.1. Setup of rotation axes

In order to sample a greater portion of the rotation space than a single rotation axis does (Wierman *et al.*, 2016), two orthogonal rotation axes were built by fixing the φ rotation stage (Newport URS100) perpendicularly to the ψ rotation stage (Newport UE17CC), as schematically shown in Fig. 1. In the present experiment rotations about the ψ axis bring the φ rotation stage into alignment with the X-ray beam, so the ψ rotation was limited to 18° to avoid blocking the beam and collision with the detector. This limited angular range of ψ is acceptable because, as described in §2.4.2, the solution to the two-axis problem is readily generalized to full three-dimensional rotations even over this limited angular range. Using a microscope, the rotation axes were adjusted to intersect perpendicularly and their intersection was centered within the X-ray beam using a fluorescent fiber, whose position was recorded for the subsequent sample centering.

2.2. Sample preparation

The protein crystallization technique followed is similar to that described by Wierman *et al.* (2016). Lyophilized hen egg white lysozyme powder (Sigma, Saint Louis, MO, USA) was dissolved in deionized water to 25 mg ml^{-1} without further purification. Crystals were grown at room temperature (293 K) by the hanging-drop diffusion method by mixing 2 µl of protein solution with 2 µl of reservoir solution containing 1.0 *M* sodium chloride and 0.1 *M* sodium acetate pH 4.5. Crystals appeared after a few days of growth with dimensions 0.4 × 0.4 × 0.6 mm. Using a large-opening pipette fixed on end with a small length of poly(ethylene terephthalate)

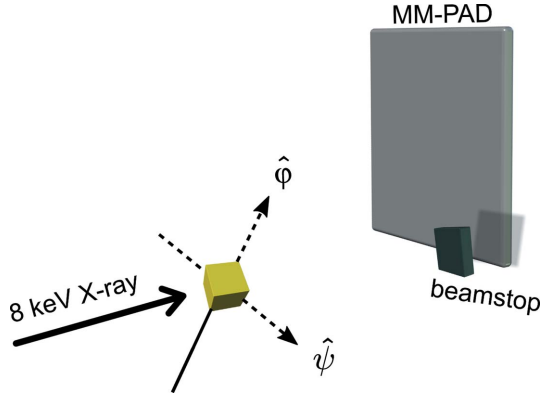


Figure 1
A simplified schematic of the experimental setup with two orthogonal rotation axes. The beam incidence is perpendicular to the ψ axis and the MM-PAD, and the main beam is blocked by the beamstop. The crystal is rotated in increments of 0.1° about the ψ axis, with the data frames recorded by the MM-PAD when φ traverses 360° continuously at each value of ψ. The figure is not drawn to scale.

capillary (outside diameter = 864 µm, wall thickness = 25.4 µm; Advanced Polymers, Salem, NH, USA), a crystal was then retrieved from the crystallization droplets and positioned approximately 22 mm from one end of the capillary. This provided enough remaining length of the capillary on the opposite end to add a small amount of mother liquor solution near the crystal to maintain hydration *via* vapor diffusion during data collection. Mother liquor directly surrounding the crystal was carefully removed from the 22 mm end of the capillary with a paper wick (Hampton Research, Aliso Veijo, CA, USA), and the ends of the capillary were sealed with vacuum grease. The capillary containing the crystal was then mounted on a Hampton Research pin base and attached to a goniometer for data collection.

2.3. Data collection

The X-ray diffraction patterns were collected from the capillary containing a single HEWL protein crystal centered at the intersection of the two orthogonal rotation axes and illuminated by a 0.5×0.5 mm Cu $K\alpha$ X-ray beam (1.54 Å wavelength). The X-rays were generated from a rotating anode machine set to 36 kV and 50 mA (Rigaku RU-H3R) and focused using Ni-coated Franks mirrors placed 1 m from the sample with a divergence of 1 mrad and a flux of 10^7 photons per second. The beam incidence was perpendicular to the ψ axis and the MM-PAD, and the sample-to-detector distance was 60 mm. The center of the beam was placed in one corner of the active area of the MM-PAD, giving a resolution of 2.0 Å in the opposite corner. A pin-diode beamstop was used to prevent the direct beam from striking the MM-PAD during data collection.

The capillary and crystal were rotated about the ψ axis from 0 to 17.9° and then from −18.0 to −0.1° in increments of 0.1°. At each value of ψ, the capillary and crystal were rotated by 360° about the φ axis continuously at a constant angular velocity of 0.5° per second. The MM-PAD collected images at a framing rate of 4 ms per frame in each revolution of φ, which gave an oscillation angle of 0.002° per frame. After the average dark signal had been subtracted and the chip-to-chip global response adjusted, pixel counts were thresholded to avoid false positives. The thresholded pixel counts were then quantized to photon counts by dividing with a known gain and rounding to the nearest integer. Only counts from pixels with at least one photon hit were recorded during data collection to reduce the file size for storage and allow more images to be recorded with the available disk space.

Owing to radiation damage and possible dehydration of the crystal, we only kept the data frames recorded at ψ ranging from 0 to 15.9° to pass on to processing. We also discarded frames that did not record any photons, which was possibly caused by glitches of the rotating anode. To simulate the signal level of an SMX experiment, we further collapsed every 100 successive frames that did not contain any discarded frames, since they were recorded when the crystal was rotated continuously in φ at a fixed value of ψ. We note that an intensity reconstruction was attempted by collapsing every 30 successive frames, but the Bragg reflections could not be reconstructed beyond 3 Å. The collapse of every 100 successive frames gave us 2.7×10^5 frames with an average of 3000 photons per collapsed frame. These collapsed frames were then passed to the EMC algorithm for intensity reconstruction, though their relative orientations were unknown to the algorithm.

It was discovered after data had been collected and the apparatus disassembled that the crystal was of poor quality. The actual Bragg spot intensities obtained by summing adjacent frames with their known relative orientations cannot be phased to produce a high-resolution structure even though the Bragg peaks do extend to high resolutions. The goal of the experiment, however, was not to solve the well known lysozyme structure but rather to demonstrate that the EMC approach can reconstruct the intensity map in the two-axis case. Because the quality of the reconstructed intensities can be assessed by comparing with the actual intensities, the goal of the experiment could be met even though the crystal was of poor quality for solving a structure.

2.4. Intensity reconstruction

2.4.1. EMC algorithm. The unoriented data frames were merged into a three-dimensional intensity map iteratively with the EMC algorithm (Loh & Elser, 2009). Each iteration of the algorithm consists of three steps: expand (E), maximize (M) and compress (C). Consider a reconstruction problem with $M_{\rm pix}$ detector pixels, $M_{\rm rot}$ rotation samples, $M_{\rm data}$ data frames and N average photons per frame. Starting with an initial intensity model $W(\mathbf{q})$, where $\mathbf{q}$ denotes the spatial frequency, the E step calculates the average photon number W_{ij}, measured at pixel i from $W(\mathbf{q})$ when the crystal has orientation Ω_j. With the data frames represented by K_{ik}, the photon count recorded at pixel i in frame k, the matrices W_{ij} and K_{ik} are cross correlated in the M step to evaluate the conditional probability $P_{jk}(W)$ that frame k was measured at crystal orientation Ω_j based on the current intensity model W. Assuming Poisson statistics, $P_{jk}(W)$ is given by

$$P_{jk}(W) = \frac{w_j \prod_{i=1}^{M_{\rm pix}} W_{ij}^{K_{ik}} \exp(-W_{ij})}{\sum_{j=1}^{M_{\rm rot}} [w_j \prod_{i=1}^{M_{\rm pix}} W_{ij}^{K_{ik}} \exp(-W_{ij})]}, \qquad (1)$$

where w_j is the fraction of the continuous rotation group assigned to sample Ω_j. The algorithm subsequently maximizes the expectation value of the log-likelihood function over $P_{jk}(W)$ by updating the model according to the rule

$$W_{ij} \rightarrow W'_{ij} = \frac{\sum_{k=1}^{M_{\rm data}} P_{jk}(W) K_{ik}}{\sum_{k=1}^{M_{\rm data}} P_{jk}(W)}. \qquad (2)$$

The C step maps W'_{ij} back to the reciprocal space to form a new intensity map $W'(\mathbf{q})$ to ensure consistency among all the tomograms W_{ij} calculated in the next iteration. The algorithm then takes $W'(\mathbf{q})$ as the initial intensity model $W(\mathbf{q})$ of the next iteration and repeats the iterations until $W(\mathbf{q}) \simeq W'(\mathbf{q})$.

2.4.2. Reference intensity map, rotation sampling and initial seeding. Although the relative orientations of the data

frames were not passed to the EMC algorithm, we can use them to construct a 'reference' intensity map to compare with the reconstructed intensity map. The data frames were mapped to the reciprocal space to form a three-dimensional intensity map according to their relative orientations when recorded. The reciprocal lattice of the crystal is embedded in the intensity map and differs from the laboratory frame by a global rotation R_g. We determined R_g by segmenting out the Bragg peaks (Wierman *et al.*, 2016) and then applying indexing (Steller *et al.*, 1997) to the peaks. The intensity map was subsequently rotated by R_g to align with the laboratory frame, and this aligned intensity map is what we call the reference intensity map.

We generated the discrete rotation samples using quaternions (Loh & Elser, 2009), where the angular resolution $\delta\theta \simeq 0.944/n$ is specified by the order $n = 1, 2, \ldots$. In this study, we confined the rotation samples to those in the subset of rotation space explored by the rotated crystal. The range of the subset in the laboratory frame was found by applying the global rotation R_g obtained above to the relative orientations between the data frames, though we need to stress that the orientation of each data frame within the subset was unknown to the EMC algorithm. This choice of rotation samples makes the solution to the two-axis problem directly applicable to the randomly oriented frames in real SMX experiments, where the rotation subset is replaced with the whole three-dimensional rotation space.

We seeded the initial intensity map with small three-dimensional Gaussian peaks of random height at each predicted Bragg position, with the lattice constants given by the indexing process mentioned above. In real SMX experiments with a small beam, this information can be obtained from indexing the pseudo-powder patterns. No symmetry was imposed in either the seeding or the reconstruction process.

2.4.3. Local update scheme. Owing to the exhaustive search in rotations, an EMC reconstruction is usually challenged by its poor time and memory scaling, which are both proportional to $M_{\rm rot}$. Resolving peaks at high resolution becomes especially difficult, since

$$M_{\rm rot} \propto n^3 \propto q_{\rm max}^3, \tag{3}$$

where $q_{\rm max}$ denotes the highest resolvable spatial frequency. Here we propose an update scheme to speed up the EMC algorithm at high resolution, and a parallel implementation that alleviates the memory burden is discussed in Appendix *A*.

To understand how to speed up the EMC algorithm, we first review how an EMC reconstruction converges in qualitative terms. The peaks at low resolution of the intensity map are reconstructed first owing to the strong diffraction signal at low q. These low-resolution peaks hence give each data frame a great preference for certain orientations, and the intensity map is refined about these probable orientations to resolve peaks at higher resolution. With improved signal-to-noise ratio in the intensity map, the convergence gradually proceeds from low q to high q. This observation shows that the intensity reconstruction has a special feature of locality in orientations: each data frame has high probabilities only at a handful of orientations favored by the low-resolution peaks, while the other orientations with negligible probabilities actually do not contribute to the refinement of the intensity map. Restricting the search to the vicinity of the probable orientations on a per-frame basis can therefore significantly reduce the computation time.

The computing scheme that we call the local update scheme takes advantage of the locality in orientations to speed up the convergence of the intensity reconstruction, and we hereafter refer to the scheme discussed in §2.4.1 as the standard update scheme. The local update scheme consists of two major parts: the calculation of the probable orientation list and the refinement of the intensity map. Starting with a converged low-resolution intensity model $W(\mathbf{q})$ and a coarse rotation sampling $\{\Omega_{j_c}\}$ of order n_c, the local update scheme calculates the probabilities $P_{j_c k}(W)$ according to equation (1). The probable orientation list is represented by the binary matrix $B_{j_c k}$ with

$$B_{j_c k} = \begin{cases} 1, & \text{if } P_{j_c k}(W) > \varepsilon_p, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where ε_p is a pre-defined threshold.

In the second part, the intensity map is refined using a fine rotation sampling $\{\Omega_{j_f}\}$ of order n_f without calculating all the elements of $P_{j_f k}(W)$. For each coarse rotation sample Ω_{j_c}, we define its neighborhood as the subset of rotation space that is closer to Ω_{j_c} than to any other samples, and assign the fine rotation samples Ω_{j_f} that lie in this subset as the neighbors of Ω_{j_c}. This mapping is stored as a matrix $C_{j_c j_f}$, where

$$C_{j_c j_f} = \begin{cases} 1, & \text{if } \Omega_{j_f} \text{ is a neighbor of } \Omega_{j_c}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The intensity map is then refined in the same way as in the standard update scheme, with the exception that only the entries of $P_{j_f k}(W)$ that satisfy the conditions $B_{j_c k} = 1$ and $C_{j_c j_f} = 1$ are calculated while the others are set as zero. We hence restrict the calculation of $P_{j_f k}(W)$ to the neighbors of the probable coarse rotation samples in each data frame. The probable orientation list, or equivalently the binary matrix $B_{j_c k}$, is only recalculated after the intensity map converges to allow a global search over all the coarse rotation samples. The refinement then continues with the updated matrix $B_{j_c k}$. The whole process terminates when the update of the probable orientation list stops changing the intensity map.

Restricting the search in orientations saves a great amount of computation because calculating the probability matrix is the most time-consuming part of the EMC algorithm. A simple estimate (see Appendix *B*) shows that the local update scheme can achieve a speed-up by tens to hundreds of times in practice. In addition, the matrices $B_{j_c k}$ and $C_{j_c j_f}$ are both sparse, so they barely add any burden to the memory usage. Since the local update scheme places no special focus on the Bragg peaks, it is also applicable to single-particle imaging.

The idea of our local update scheme is similar to the sparse update scheme proposed by Neal & Hinton (1998), which speeds up the expectation maximization algorithm by freezing the probabilities of improbable values in most of the iterations

and only updating them once every many iterations. The only difference is the specific property of locality in our intensity reconstruction application, which allows us to search in a finer grid about the probable coarse rotation samples to refine the intensity map at high resolution. Nonetheless, we need to stress that the only reason to adopt the local update scheme is to speed up the reconstruction at high resolution. The likelihood function maximized in each local update iteration cannot exceed its counterpart when the whole rotation group is explored.

2.5. Integration

After a converged intensity map had been obtained, the reflections were summed over ellipsoidal windows centered at each Bragg position and aligned with the reciprocal lattice. We used the average of the neighboring voxels outside each ellipsoidal window to estimate the background level of each reflection. Reflections with their ellipsoidal windows intersecting the detector gaps or the boundary of the intensity map were considered as partial peaks and rejected.

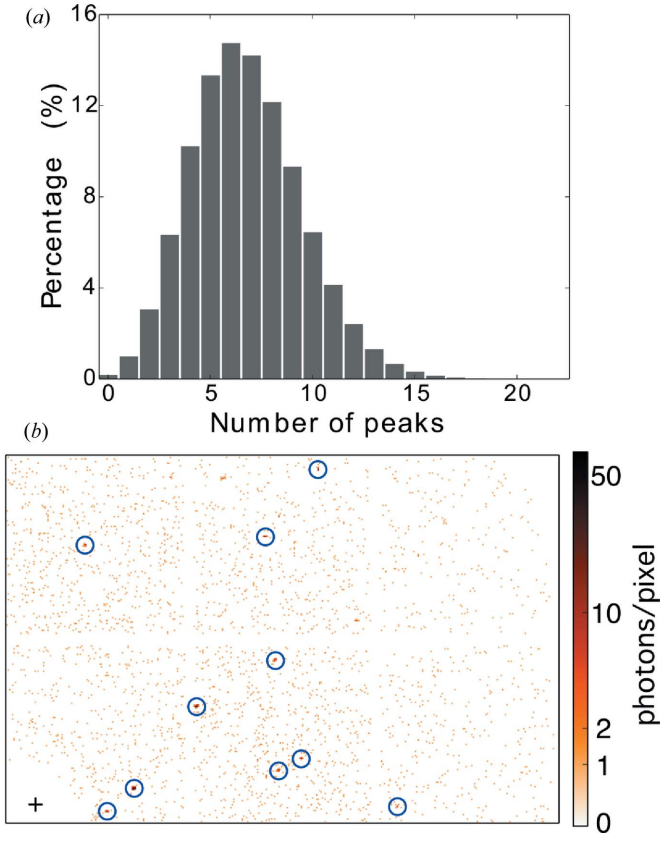


Figure 2
(*a*) Histogram of the number of peaks per collapsed frame, which is the sum of 100 successive frames in the raw data. A patch with more than two connected pixels and an average of no less than two photons per pixel is identified as a peak. (*b*) A random selection of the collapsed frames, with identified peaks marked with blue circles. The cross denotes the beam center, and the resolution at the upper right corner is about 2 Å.

3. Results

3.1. Sparsity of data frames

To show the sparsity of the collapsed data frames described in §2.3, we counted the number of peaks per frame with the criterion that a peak has more than two connected pixels and an average of no less than two photons per pixel. As shown in Fig. 2, most of the frames do not have enough peaks to meet the requirements of conventional indexing methods (at least 20–30), even with this generous criterion for peak finding.

Following the calculation by Holton (2009), we also estimated the energy absorbed by the crystal over the exposure of one collapsed frame, assuming that protein crystals have the same mass energy absorption cross section as water. Our calculation showed that an 8 μm^3 crystal would have endured a 0.2 MGy radiation dose if it had scattered the same number of photons as our large HEWL crystal during this period. This dose is within the lifetime of protein crystals at room temperature if the radiation is delivered quickly (Owen *et al.*, 2012), so the signal level in our study should be comparable to that in a real SMX experiment.

3.2. Intensity reconstruction

Given the 2.7×10^5 collapsed frames, we started an EMC reconstruction from the randomly seeded model described in §2.4.2 using the standard update scheme and a rotation sampling of order $n = 40$. Only data up to 3 Å were used at this stage because our goal was to quickly obtain a converged intensity map at low resolution. After the intensity map had converged, we took its probability distribution and assembled all the data frames to form an intensity map using equation (2) to include data up to 2 Å resolution. This intensity map was then used as the initial model of the local update scheme using rotation samplings of orders $(n_c, n_f) = (40, 60)$ for refinement.

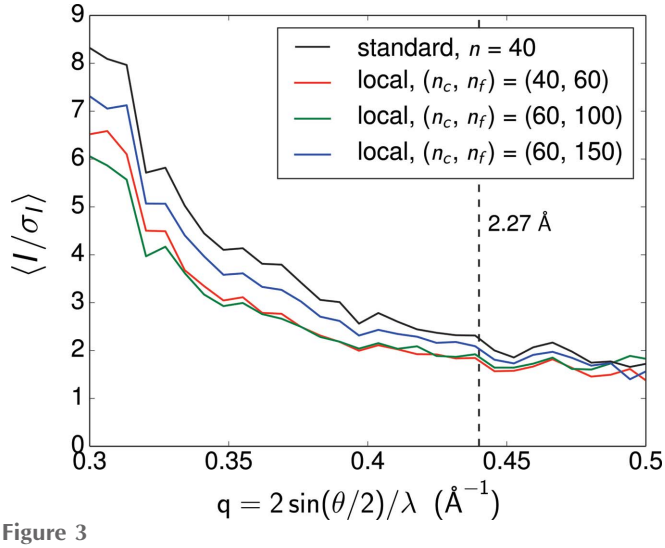


Figure 3
The average signal-to-noise ratio of the integrated reflections from the converged intensity maps at different stages of the reconstruction. The increase of $\langle I/\sigma_I \rangle$ at high q indicates the reconstruction of high-resolution peaks. The 2.27 Å resolution determined by CC* is marked by the black dashed line.

Different pairs of orders (n_c, n_f) with increasing angular resolutions were sequentially used in the local update scheme to extend the peak convergence to high resolution.

Fig. 3 shows the average signal-to-noise ratio $\langle I/\sigma_I \rangle$ of the integrated reflections from the converged intensity maps at different stages of the reconstruction. We first see that $\langle I/\sigma_I \rangle$ dropped at low resolution while it remained at similar levels at high resolution when moving from the standard update scheme of $n = 40$ to the local update scheme of $(n_c, n_f) = (40, 60)$. The lack of improvement at high resolution indicates that the current angular resolution of the local update scheme still cannot resolve high-resolution peaks. On the other hand, the inclusion of data beyond 3 Å slightly disrupted the original probability distribution, which in turn reduced $\langle I/\sigma_I \rangle$ at low resolution. The improvement of $\langle I/\sigma_I \rangle$ when increasing the angular resolutions signals the reconstruction of high-resolution peaks and justifies the local update scheme.

With the converged intensity map from the local update scheme of $(n_c, n_f) = (60, 150)$ as our final intensity reconstruction, Fig. 4 compares the slices of the reconstructed and reference intensity maps perpendicular to the k axis of the reciprocal lattice. Although we did not impose any symmetry in the process of seeding or reconstruction, the converged intensity map still follows the reflection condition $0k0 : k = 2n$ required by the space-group symmetry $P4_32_12$ of the HEWL crystal (Hahn, 2006), which demonstrates the success of the EMC reconstruction. We note that the discrepancy between the two intensity maps in high-resolution peaks is consistent with the low signal-to-noise ratio at high resolutions (see Fig. 3). Because the photons contributing to the high-resolution shells were mostly collected by the upper left corner of the MM-PAD (Fig. 1), the resulting lower signal-to-noise ratio made the orientation reconstruction more challenging in this region.

A further comparison is shown in the scatter plot of the integrated reflections from the reconstructed and reference intensity maps (Fig. 5), which excludes the reflections with the signal-to-noise ratio $I/\sigma_I < 2$. The linear correlation of the reflections shows the consistency of the two intensity maps. Using

$$R = \sum_{hkl} |F_{\rm ref} - F_{\rm reconst}| / \sum_{hkl} F_{\rm ref}, \qquad (6)$$

where $F_{\rm ref}$ and $F_{\rm reconst}$ are the structure factors calculated from the reference and reconstructed intensity maps, respectively, we quantify the discrepancy between the two sets of integrated reflections as $R = 21\%$, where the reflections with $I/\sigma_I < 2$ are also excluded. This larger discrepancy than $R = 4.7\%$ obtained by Wierman *et al.* (2016) could be caused by the adverse influence of the background scatter and the exploration of a much larger rotation subset than a single rotation axis. By summing the total photon counts of both the integrated and the partial peaks, we estimated the fraction of photons coming from the background and diffuse scatter as ∼90%. We expect to improve the quality of our reconstruction and push the limit to sparser data frames by reducing the background scatter and using a larger detector to gain more information to assist orientation reconstruction.

Another way to assess the quality of the reconstruction is through the calculation of CC*, the correlation coefficient of the observed reflections with the underlying true signal (Karplus & Diederichs, 2012). We first randomly separated the symmetry-related reflections of each unique reflection into two subsets and then calculated the correlation coefficient $CC_{1/2}$ between the average intensities of the two subsets in different resolution shells. Under the assumption that the errors of the two subsets are independent, identically

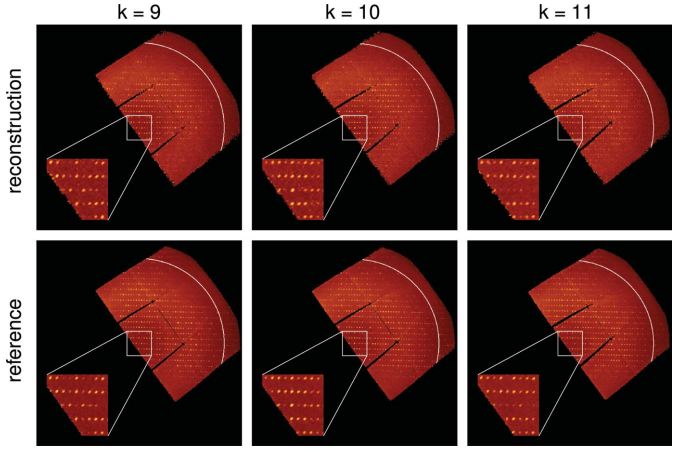


Figure 4
Slices of the reconstructed and reference intensity maps in the *hl* plane at constant values of k of the reciprocal lattice. Even without imposing any symmetry in the process of seeding or reconstruction, the converged intensity map still follows the reflection condition $0k0 : k = 2n$ required by the space-group symmetry $P4_32_12$ of the HEWL crystal (see insets). The 2.27 Å resolution determined by CC* is marked by the arcs in white. The mapping into reciprocal space transforms the detector gaps (Tate *et al.*, 2013) into curves.

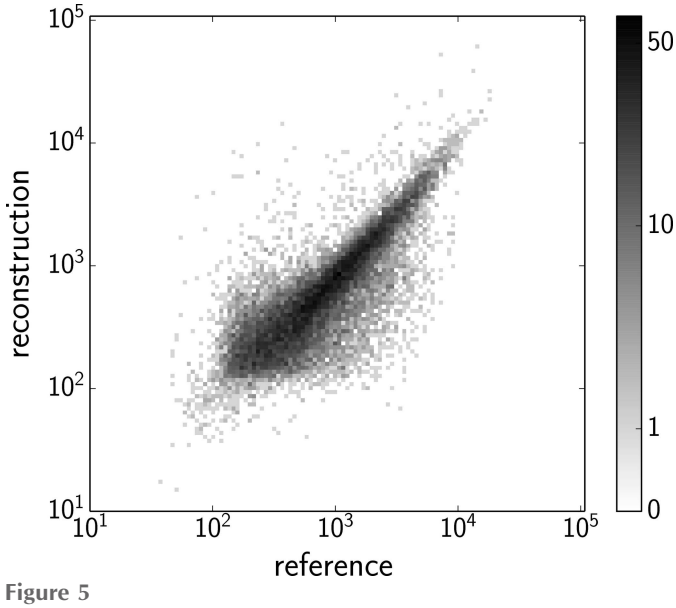


Figure 5
Scatter plot comparing the integrated reflections from the reconstructed and reference intensity maps. Reflections with the signal-to-noise ratio $I/\sigma_I < 2$ are excluded from the plot. The linear correlation shows the agreement between the two intensity maps.

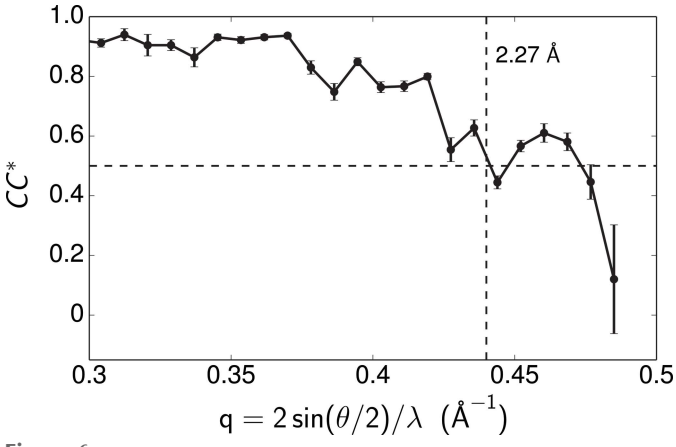


Figure 6
The distribution of CC* as a function of spatial frequencies. The resolution of the reflections is determined as 2.27 Å by a threshold CC* = 0.5. The error bars are estimated by repeating the random separation of reflections 1000 times, while the ups and downs in CC* result from the binning in resolution shells.

distributed and free from the errors of the true signal, the value of CC* can be estimated from

$$\mathrm{CC}^* = \left(\frac{2\mathrm{CC}_{1/2}}{1+\mathrm{CC}_{1/2}}\right)^{1/2}. \qquad (7)$$

The distribution of CC* as a function of spatial frequencies is shown in Fig. 6, with the error bars estimated by repeating the random separation of reflections 1000 times. The large error bar in the highest-resolution shell shows the low correlation between the intensities of the two subsets, which is consistent with the low signal-to-noise ratio at high resolution. We determine the resolution of the reconstructed reflections as 2.27 Å by a threshold CC* = 0.5. This choice is consistent with the resolution where the average signal-to-noise ratio of our final reconstructed intensity (the black curve in Fig. 3) drops to 2. We note that the value of the correlation coefficient is dominated by the stronger peaks in each resolution shell. Therefore, CC* can still have moderate values at high resolutions even if some low-signal peaks are not resolvable, as indicated by the discrepancy between the two intensity maps in high-resolution peaks in Fig. 4.

4. Conclusion

The results of this study show that the limit to the usable crystal sizes in current SMX experiments could be relaxed by employing the EMC algorithm. Because the algorithm leverages the data redundancy arising from the common arcs between pairs of diffraction patterns, the intensity reconstruction is feasible even though each frame may not contain sufficient information to be oriented individually. The computing schemes we have developed in this article further alleviate the computational requirements of the EMC reconstruction, which makes EMC-based SMX experiments more practical.

The fact that the EMC algorithm was able to reconstruct the actual X-ray intensity incident on the detector, irrespective of crystal quality, illustrates the generality of the algorithm. The algorithm has no 'knowledge' of what is being reconstructed. Everything in the detected X-ray field – background, diffraction spots, diffuse scatter *etc.* – is reconstructed.

Several issues remain to be addressed to put EMC-based SMX experiments into practice. In contrast to the randomly sampled crystal orientations in real SMX experiments, the data frames in this study were taken from a crystal rotated continuously. From our past experience, this difference in orientation sampling should not affect the EMC reconstruction as long as the random orientation sampling size is large enough. We estimate that 10^5–10^6 data frames are required, which amounts to a data collection time of within a day when a 10% single-crystal hit rate and an exposure time of 10 ms per frame are assumed.

In this paper we used a large single crystal in various orientations to emulate the data expected from multiple small crystals. The obvious next step towards practical application of the method is to try the EMC algorithm on data from multiple small crystals. It will be necessary to experimentally determine the severity of difficulties arising from sources including varying crystal diffraction quality and occasional multiple crystals in the beam. We expect to incorporate metrics such as the normalized surprise function (Munke *et al.*, 2016) into the EMC algorithm to estimate the reliability of each frame based on the current intensity model and reject frames containing multiple lattices. To tackle the frame-to-frame crystal size variation, the EMC algorithm also needs to calculate the relative contribution of each frame to the intensity model iteratively. As indicated by Loh *et al.* (2010), the intensity model could be updated by maximizing the likelihood function with respect to the relative crystal sizes and the crystal orientations alternately, with the cost of doubling the computation time.

The last issue is the background scatter. In principle, the EMC algorithm is able to deal with stable background by modifying the conditional probability calculation. However, background reduction by improving the experiment becomes necessary when the frame-to-frame variation in background is significant. With the above issues considered, the analysis of SMX data would involve first reconstructing a low-resolution intensity map using the standard EMC update scheme, and then refining the high-resolution peaks using the local update scheme because of its computational efficiency.

APPENDIX *A*
Memory-efficient parallel implementation

Here we describe a memory-efficient parallel implementation of the EMC algorithm, which is applicable to both the standard and local update schemes. The memory usage of the EMC algorithm is dominated by the matrices $P_{jk}(W)$ and W_{ij}/W'_{ij}, with sizes of $M_{\mathrm{rot}} \times M_{\mathrm{data}}$ and $M_{\mathrm{pix}} \times M_{\mathrm{rot}}$, respectively. Since $M_{\mathrm{rot}} \propto q_{\mathrm{max}}^3$, the required memory rapidly

becomes intractable, for example, hundreds of terabytes (TB), even to resolve peaks at moderate resolutions (Ayyer *et al.*, 2016).

However, only a small portion of the entries of $P_{jk}(W)$ are significant according to our discussion in §2.4.3, so we can treat $P_{jk}(W)$ as a sparse matrix to reduce the required memory. This amounts to the fact that a given data frame only has non-negligible probabilities at a small fraction of the orientations, unless the signal level is as weak as only several photons per frame. In our implementation, we distribute blocks of data frames (ranges in k index) to different processors, each of which holds the same copy of the intensity map $W(\mathbf{q})$. The algorithm strides through the $M_{\rm rot}$ rotations in steps of size $M_{\rm step}$ and calculates W_{ij} and $R_{jk}(W)$ in each step. Here we define $R_{jk}(W)$ as

$$R_{jk}(W) = w_j \prod_{i=1}^{M_{\rm pix}} W_{ij}^{K_{ik}} \exp(-W_{ij}), \tag{8}$$

and each processor dynamically updates the value of $\max_j R_{jk}(W)$ when walking through all the orientations. From the inequality

$$P_{jk}(W) = \frac{R_{jk}(W)}{\sum_{j=1}^{M_{\rm rot}} R_{jk}(W)} \leq \frac{R_{jk}(W)}{\max_j R_{jk}(W)}, \tag{9}$$

the entries of $R_{jk}(W)$ are saved only when the ratio $R_{jk}(W)/\max_j R_{jk}(W)$ exceeds the pre-defined threshold $\varepsilon_{\rm p}$, and this condition is checked for all the saved entries every time the value of $\max_j R_{jk}(W)$ is updated. After going through all the rotation samples, the algorithm calculates the significant values of $P_{jk}(W)$ by normalizing the saved entries of $R_{jk}(W)$ over orientations. Subsequently, we update the tomograms W'_{ij} also in steps of size $M_{\rm step}$ over all the orientations, and map W'_{ij} back to the copy of the updated intensity map $W'(\mathbf{q})$ held by each processor after each step. Finally, $W'(\mathbf{q})$ is reduced among all the processors to complete the iteration.

As a result, the memory scaling of $P_{jk}(W)$ and W_{ij}/W'_{ij} is reduced to $N_{\rm p} M_{\rm data}$ and $M_{\rm pix} M_{\rm step}$, respectively, where $N_{\rm p}$ denotes the average number of probable orientations per frame and is governed by the threshold $\varepsilon_{\rm p}$ and the signal level. The memory usage can be limited to only tens of gigabytes even when using an extremely fine rotation sampling, since the dominant memory scaling is independent of $q_{\rm max}$.

APPENDIX *B*
Speed-up of the local update scheme

The most time-intensive part of a standard EMC update is to calculate the probability matrix $P_{jk}(W)$, which is proportional to $M_{\rm rot}$. In a local update scheme with $M_{\rm coarse}$ and $M_{\rm fine}$ rotation samples, this proportionality becomes $N_{\rm p} M_{\rm fine}/M_{\rm coarse}$, where $N_{\rm p}$ is the average number of the probable coarse rotation samples per frame and $M_{\rm fine}/M_{\rm coarse}$ is the average number of neighbors each coarse rotation sample has. Consider a local update scheme that recalculates the most probable orientation lists $B_{j_{\rm c}k}$ every $N_{\rm iter}$ iterations. The speed-up is given by

$$\frac{N_{\rm iter} M_{\rm fine}}{M_{\rm coarse} + N_{\rm iter} N_{\rm p} M_{\rm fine}/M_{\rm coarse}}, \tag{10}$$

where the numerator is the time proportionality for $N_{\rm iter}$ iterations of standard updates using the fine rotation sampling, and $M_{\rm coarse}$ in the first term of the denominator denotes the proportionality for the recalculation of $B_{j_{\rm c}k}$.

The number of rotations that samples the rotation space with order n is given by $M_{\rm rot} = 10(5n^3 + n)$. Consider a local update scheme that uses $N_{\rm iter} = 10$ and rotation samplings of orders $(n_{\rm c}, n_{\rm f}) = (60, 150)$. Assume that a chosen value of $\varepsilon_{\rm p}$ leads to $N_{\rm p} \simeq 100$. Equation (7) gives a speed-up of 156 times.

Acknowledgements

We thank Marian Szebenyi, the Macromolecular Diffraction at CHESS (MacCHESS) team and other members of the Gruner Group for their support, and Robert E. Thorne and Hakan Atakisi for discussion. This work is based upon research conducted at the Gruner Laboratory, which is supported by the Department of Energy (DOE) grants DE-FG02-10ER46693 and DE-SC0016035, by the Elser group, which is supported by DOE grant DE-FG02-11ER16210, and at the Cornell High Energy Synchrotron Source (CHESS), which is supported by the National Science Foundation (NSF) and the National Institutes of Health/National Institute of General Medical Sciences under NSF award DMR-1332208, using the MacCHESS facility, which is supported by award GM-103485 from the National Institute of General Medical Sciences, National Institutes of Health. TYL is grateful for support from the Taiwan Government Scholarship to Study Abroad.

References

Ayyer, K., Lan, T.-Y., Elser, V. & Loh, N. D. (2016). *J. Appl. Cryst.* **49**, 1320–1335.

Ayyer, K., Philipp, H. T., Tate, M. W., Elser, V. & Gruner, S. M. (2014). *Opt. Express*, **22**, 2403–2413.

Ayyer, K., Philipp, H. T., Tate, M. W., Wierman, J. L., Elser, V. & Gruner, S. M. (2015). *IUCrJ*, **2**, 29–34.

Botha, S., Nass, K., Barends, T. R. M., Kabsch, W., Latz, B., Dworkowski, F., Foucar, L., Panepucci, E., Wang, M., Shoeman, R. L., Schlichting, I. & Doak, R. B. (2015). *Acta Cryst.* D**71**, 387–397.

Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.

Chapman, H. N. *et al.* (2011). *Nature*, **470**, 73–77.

Ekeberg, T. *et al.* (2015). *Phys. Rev. Lett.* **114**, 098102.

Gati, C., Bourenkov, G., Klinge, M., Rehders, D., Stellato, F., Oberthür, D., Yefanov, O., Sommer, B. P., Mogk, S., Duszenko, M., Betzel, C., Schneider, T. R., Chapman, H. N. & Redecke, L. (2014). *IUCrJ*, **1**, 87–94.

Gruner, S. M. & Lattman, E. E. (2015). *Annu. Rev. Biophys.* **44**, 33–51.

Hahn, T. (2006). Editor. *International Tables for Crystallography*, Vol. A, *Space-Group Symmetry*, 1st online ed. Chester: International Union of Crystallography.

Heymann, M., Opthalage, A., Wierman, J. L., Akella, S., Szebenyi, D. M. E., Gruner, S. M. & Fraden, S. (2014). *IUCrJ*, **1**, 349–360.

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

Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.

Loh, N. D., Bogan, M. J. *et al.* (2010). *Phys. Rev. Lett.* **104**, 225501.
Loh, N. D. & Elser, V. (2009). *Phys. Rev. E*, **80**, 026705.
Munke, A. *et al.* (2016). *Sci. Data*, **3**, 160064.
Neal, R. M. & Hinton, G. E. (1998). *Learning in Graphical Models*, edited by M. I. Jordan, pp. 355–368. Dordrecht: Springer.
Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature*, **406**, 752–757.
Nogly, P. *et al.* (2015). *IUCrJ*, **2**, 168–176.
Owen, R. L., Axford, D., Nettleship, J. E., Owens, R. J., Robinson, J. I., Morgan, A. W., Doré, A. S., Lebon, G., Tate, C. G., Fry, E. E., Ren, J., Stuart, D. I. & Evans, G. (2012). *Acta Cryst.* D**68**, 810–818.
Philipp, H. T., Ayyer, K., Tate, M. W., Elser, V. & Gruner, S. M. (2012). *Opt. Express*, **20**, 13129–13137.
Philipp, H. T., Koerner, L. J., Hromalik, M. S., Tate, M. W. & Gruner, S. M. (2008). *Nuclear Science Symposium Conference. Record 2008 NSS '08*, pp. 1567–1571. IEEE.
Schubert, R., Kapis, S., Gicquel, Y., Bourenkov, G., Schneider, T. R., Heymann, M., Betzel, C. & Perbandt, M. (2016). *IUCrJ*, **3**, 393–401.
Stellato, F. *et al.* (2014). *IUCrJ*, **1**, 204–212.
Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.
Tate, M. W., Chamberlain, D., Green, K. S., Philipp, H. T., Purohit, P., Strohman, C. & Gruner, S. M. (2013). *J. Phys. Conf. Ser.* **425**, 062004.
Wierman, J. L., Lan, T.-Y., Tate, M. W., Philipp, H. T., Elser, V. & Gruner, S. M. (2016). *IUCrJ*, **3**, 43–50.