STRUCTURAL BIOLOGY
ISSN 2059-7983

Challenges in solving structures from radiation-damaged tomograms of protein nanocrystals assessed by simulation

Ariana Peck,[a]* Qing Yao,[a] Aaron S. Brewster,[b] Petrus H. Zwart,[b,c] John M. Heumann,[d] Nicholas K. Sauter[b] and Grant J. Jensen[a]*

[a]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA, [b]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [c]Center for Advanced Mathematics in Energy Research Applications, Lawrence Berkeley National Laboratory, Berkeley CA 94720, USA, and [d]Department of Molecular, Cellular and Developmental Biology, University of Colorado Boulder, Boulder, CO 80309, USA. *Correspondence e-mail: apeck@caltech.edu, jensen@caltech.edu

Received 19 September 2020
Accepted 2 March 2021

Edited by Q. Hao, University of Hong Kong

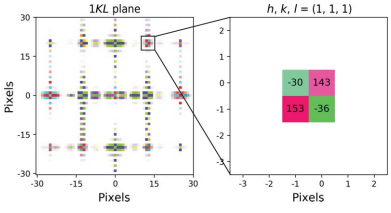
Keywords: cryo-electron tomography; diffraction methods; nanocrystals; radiation damage.

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

Structure-determination methods are needed to resolve the atomic details that underlie protein function. X-ray crystallography has provided most of our knowledge of protein structure, but is constrained by the need for large, well ordered crystals and the loss of phase information. The rapidly developing methods of serial femtosecond crystallography, micro-electron diffraction and single-particle reconstruction circumvent the first of these limitations by enabling data collection from nanocrystals or purified proteins. However, the first two methods also suffer from the phase problem, while many proteins fall below the molecular-weight threshold required for single-particle reconstruction. Cryo-electron tomography of protein nanocrystals has the potential to overcome these obstacles of mainstream structure-determination methods. Here, a data-processing scheme is presented that combines routines from X-ray crystallography and new algorithms that have been developed to solve structures from tomograms of nanocrystals. This pipeline handles image-processing challenges specific to tomographic sampling of periodic specimens and is validated using simulated crystals. The tolerance of this workflow to the effects of radiation damage is also assessed. The simulations indicate a trade-off between a wider tilt range to facilitate merging data from multiple tomograms and a smaller tilt increment to improve phase accuracy. Since phase errors, but not merging errors, can be overcome with additional data sets, these results recommend distributing the dose over a wide angular range rather than using a finer sampling interval to solve the protein structure.

1. Introduction

Protein structure determination critically advances our understanding of biochemical mechanisms and the molecular basis of disease. X-ray crystallography has been the principal method of structure determination for decades, accounting for 90% of the atomic models deposited in the Protein Data Bank (Powell, 2017). However, two limitations constrain the broader applicability of this method. Firstly, some proteins do not readily form sufficiently large (>10 µm) and well ordered crystals for characterization at synchrotron sources (Smith *et al.*, 2012; Holton & Frankel, 2010; Holton, 2009). Secondly, the phase information needed to solve the structure of a protein cannot be experimentally measured. Estimating this lost information requires experimental perturbations that some crystals are not amenable to, very high resolution diffraction data or the availability of a homologous structure (Taylor, 2003). Such prior information is often limited for proteins that are difficult to crystallize, compounding the challenge of structurally characterizing new protein families.



The need for large crystals has been overcome in part by the development of techniques suitable for submicrometre-sized nanocrystals. One of these methods is serial femtosecond crystallography (SFX), which relies on femtosecond-length X-ray pulses that are orders of magnitude brighter than synchrotron radiation (Schlichting, 2015). SFX has significantly advanced structural studies of difficult-to-crystallize membrane proteins (Liu *et al.*, 2013; Zhu *et al.*, 2016) and crystals that are sensitive to radiation damage (Young *et al.*, 2016; Ebrahim *et al.*, 2019), but suffers from low throughput. A second method is micro-electron diffraction (microED). This modality of cryo-electron microscopy (cryo-EM) takes advantage of the strong interaction between electrons and matter to enable data collection from crystals of less than 1 µm in thickness (Rodriguez *et al.*, 2017). Similar to SFX, microED has enabled structural studies of amyloid proteins that do not readily form large, well ordered crystals (Rodriguez *et al.*, 2015; de la Cruz *et al.*, 2017). In addition, selected microfocus beamlines at synchrotron sources have been designed for crystals as small as 500 nm (Beale *et al.*, 2020). However, none of these methods can measure the crystallographic phases, so this information must be inferred indirectly or supplied by other techniques (Taylor, 2003).

Single-particle reconstruction (SPR) is another cryo-EM modality that is increasingly effective for high-resolution structure determination. In SPR, projection images of purified proteins are recorded, aligned and merged to determine the structure of the protein (Cheng, 2015). This method thus overcomes the main limitations of X-ray crystallography, bypassing the need for crystallization and retaining the phase information. Further, recent technological advances have enabled SPR to achieve reconstructions with high-resolution limits comparable to those in X-ray crystallography (Cheng, 2015). However, without the signal amplification that results from coherent scattering by a crystal, molecular weight is a limiting factor. As the scattering mass of the object decreases, errors in aligning projection images increase and attenuate the high-resolution signal (Henderson, 1995; Jensen, 2001). To date, the smallest macromolecule solved by SPR to better than 4 Å resolution is a 40 kDa riboswitch, and only seven proteins of <100 kDa have been solved to similar resolution (Wu & Lander, 2020). By contrast, the median mass of proteins in the human proteome is 41 kDa (Brocchieri & Karlin, 2005), rendering SPR unsuitable for many proteins.

Cryo-electron tomography (cryo-ET) is a third modality of cryo-EM, and its application to nanocrystals could overcome the principal obstacles of mainstream structure-determination methods (Oikonomou & Jensen, 2017). In cryo-ET, a tilt series of projection images is collected from a sample embedded in vitreous ice and reconstructed into a tomogram, a volume that contains 3D structural information about the specimen. This method has traditionally been used to study cellular ultrastructure, and the reconstruction approach of subtomogram averaging has recently provided near-atomic resolution (<5 Å) structures of selected purified proteins and ribosomes *in situ* (Schur, 2019; Himes & Zhang, 2018; Tegunov *et al.*, 2021). Like microED, cryo-ET is suitable for protein nanocrystals, with an expected optimal sample thickness of 50–300 nm (Martynowycz *et al.*, 2017; Lučić *et al.*, 2013). Like SPR, imaging rather than diffraction data are collected, so the phase information is retained during the experiment (Unwin & Henderson, 1975). Further, collecting data in imaging mode provides a unique opportunity to spatially characterize and computationally correct for disorder (Nederlof *et al.*, 2013; Henderson *et al.*, 1990). In 2D electron crystallography, this was achieved by a procedure called lattice 'unbending' (Schenk *et al.*, 2010; Henderson *et al.*, 1990). Alternatively, disorder could be corrected in real space with subtomogram averaging algorithms (Nicastro *et al.*, 2006; Heumann *et al.*, 2011; Castaño-Díez *et al.*, 2012; Bharat & Scheres, 2016; Himes & Zhang, 2018; Chen *et al.*, 2019). Another benefit of applying cryo-ET to nanocrystals is the potential to develop a hybrid method that combines high-resolution diffraction intensities from microED and low-resolution phases from imaging for structure solution. Even phases to intermediate resolution (∼7 Å) should suffice to resolve α-helices and provide a robust starting model for phase extension (Jackson *et al.*, 2015; Stuart & Abrescia, 2013). Development of this hybrid method should be straightforward, as sample preparation and the microscope are shared between the two techniques. Experimental phases from cryo-ET could similarly be used to phase diffraction intensities obtained by X-ray crystallography, and could prove particularly valuable in the absence of a molecular-replacement model.

While electron imaging of 3D crystals has revealed lattice structure at nanoscale lengths (Nederlof *et al.*, 2013; Gallagher-Jones *et al.*, 2019; van Genderen *et al.*, 2016), to date structure determination has not been achieved from tomograms of such samples. In our first attempts to solve structures from experimental tomograms of nanocrystals, we found that low completeness was particularly severe at high tilt angles and was prohibitive to merging data sets. These observations prompted us to use simulations to determine how to collect more complete tilt series and the data characteristics required to merge tomograms. We focused on factors that predicted the ability to merge multiple data sets, as merging is necessary to overcome the missing wedge of information inherent to tomography and the loss of completeness due to radiation damage. Our results anticipate that only a merged data set will be sufficiently complete to position the experimental phases on a valid crystallographic phase origin, which is a prerequisite for quantifying the phase errors introduced by other effects such as multiple scattering and the contrast transfer function (Henderson *et al.*, 1990; Subramanian *et al.*, 2015). Although more challenges remain, establishing a data-processing workflow and optimizing data collection are the first critical steps towards evaluating the effectiveness of this structure-determination method. Here, we describe a data-processing scheme to solve structures from tomograms of nanocrystals that leverages software from X-ray crystallography and algorithms that we have developed to handle challenges unique to tomography data. We also assess the tolerance of this workflow to the effects of radiation damage using simulated crystals. Our results recommend a data-collection strategy that

maximizes the angular spread of the reflections recorded in each tomogram to increase the likelihood of successfully merging data sets.

2. Design of a data-processing pipeline

Structure determination from tomograms of nanocrystals could leverage algorithms from crystallography or, alternatively, rely on subtomogram averaging (Wan & Briggs, 2016; see Section 5). Here, we focus on the former approach. Since the retention of phase information is unique to imaging methods, we emphasize the steps required to recover crystallographic phases from the Fourier transform of the tomogram. Our data-processing pipeline leverages functions available in the *SPARX* (Tang *et al.*, 2007; Hohn *et al.*, 2007), *DIALS* (Waterman *et al.*, 2013; Winter *et al.*, 2018) and *cctbx* (Grosse-Kunstleve *et al.*, 2002) software packages in addition to providing new algorithms developed specifically for tomographic data of crystal specimens. The steps are presented schematically in Supplementary Fig. S1(*a*) and are described in the following sections.

To develop this workflow and explore different ways of processing these data, we simulated tomograms of protein nanocrystals. We selected lysozyme in *P*1 symmetry (PDB entry 6d6g; Juers *et al.*, 2018) and a peptide inhibitor in space group $P2_12_12_1$ (PDB entry 4bfh; Nguyen *et al.*, 2014) as model crystal systems to test distinct space-group symmetries and protein folds. For each crystal system, the protein coordinates from the Protein Data Bank were tiled in *UCSF Chimera* (Pettersen *et al.*, 2004) to generate 3D nanocrystals containing ten unit cells along each dimension. Density maps were simulated from the atomic coordinates of the nanocrystal to 2.5 Å resolution using electron scattering factors with the *cctbx* software package and were rotated to randomly sampled orientations (Grosse-Kunstleve *et al.*, 2002). These intact volumes were projected into tilt series spanning either a ±60° tilt range with 3° increments or a ±40° tilt range with 2° increments between projection images. No optical aberrations from the microscope, including the contrast transfer function, were simulated. The resulting tilt series were reconstructed into tomograms using *IMOD* (Kremer *et al.*, 1996). The workflow used to generate these simulated data sets is shown in Supplementary Fig. S1(*b*).

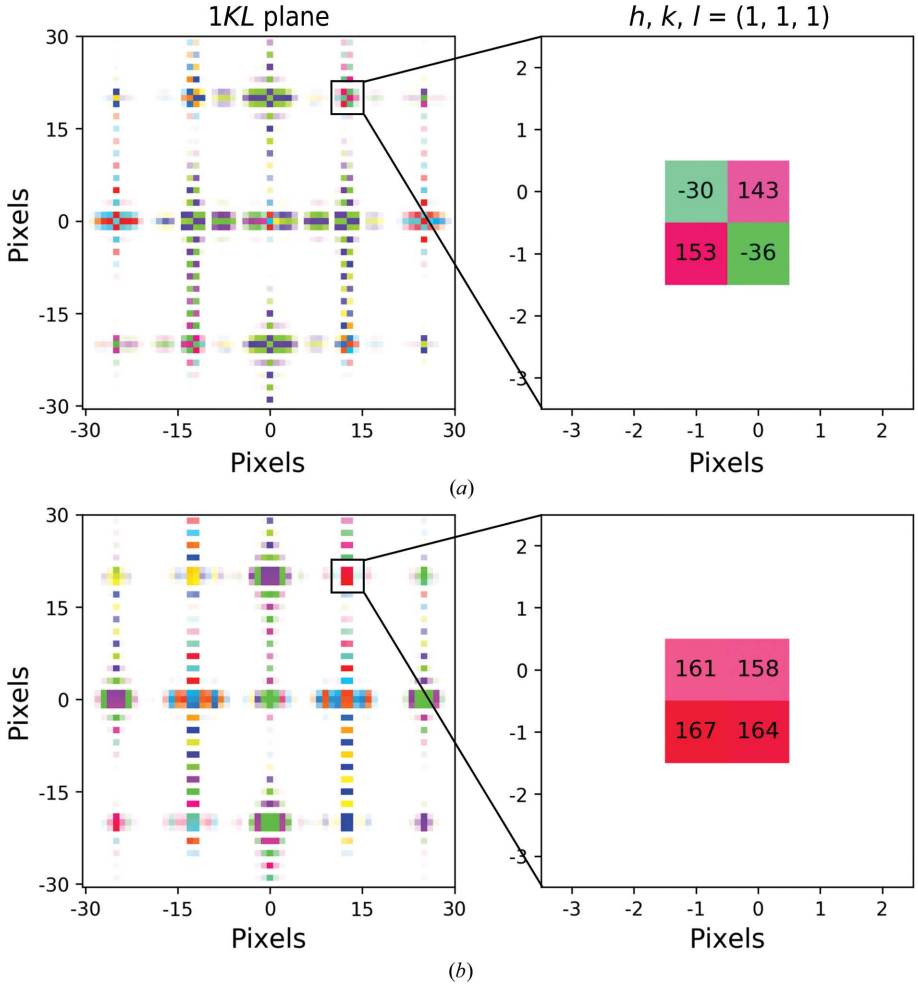


Figure 1
Elimination of phase splitting at Bragg peaks. (*a*) Left: the 1*KL* plane is visualized for a simulated crystal, which is imperfectly periodic because the unit-cell dimensions do not span an integer number of pixels. The brightness and color of each pixel are determined by its intensity and phase values, respectively. Right: inset showing the (1, 1, 1) reflection. Pixels with intensity values within threefold of the maximum intensity are visualized, and the phase values of these high-intensity pixels are noted in degrees. In (*b*), a tapering function was applied to the crystal. The density was centered by auto-convolution and shifted to the origin of the volume before computing the Fourier transform.

2.1. Eliminating phase splitting

Real crystals are characterized by imperfections that result in a loss of exact periodicity. Imaging introduces further non-idealities, such as interpolation errors from discrete sampling and truncation of crystal edges due to a finite field of view. These deviations from perfect crystallinity in real space result in spatially smeared intensities and phase splitting at Bragg peak positions in reciprocal space (Supplementary Fig. S2). Specifically, the phase values of pixels immediately surrounding peak centers shift by 180° between adjacent octants (Fig. 1*a*, Supplementary Fig. S1). The split phases result from the presence of a circular discontinuity at the point considered to be the origin by the discrete Fourier transform, causing the phase to alternate in sign between frequency bins in Fourier space. There is no such discontinuity for signals that are exactly periodic in the window of the

Fourier transform, and hence no phase splitting for ideal (albeit finite) crystals.

For imperfect crystals, we found that phase splitting could be eliminated by applying a symmetric tapering function in real space, followed by centering the density within the volume. Shifting the center of the resulting volume to the origin removes the circular discontinuity at the origin of the Fourier transform calculation. Here, we chose a tapered cosine (Tukey) window with a tapering fraction of 0.5 and computed the translation needed to center the crystal density using the auto-convolution function in the *SPARX* library (Tang *et al.*, 2007; Hohn *et al.*, 2007). Applying these pre-processing steps to crystals with imperfect periodicity eliminates phase splitting and results in a consistent phase value in the immediate vicinity of each Bragg peak (Fig. 1*b*, Supplementary Fig. S2).

2.2. Spot-finding and indexing

We adapted algorithms from the *DIALS* crystallographic software package to index the Bragg peaks in the 3D Fourier transform of the tomogram (Waterman *et al.*, 2013). Peaks were identified by scanning the Fourier transform along the x axis (parallel to the axis of rotation), and high-intensity pixels in each slice were identified using the pixel-thresholding algorithm described in Winter *et al.* (2018). Sets of contiguous bright pixels in adjacent slices were assembled into an initial list of spots, which was then filtered according to the specified gain (the ratio of electrons per detector pixel to reported counts), resolution limits and minimum and maximum number of included pixels (Winter *et al.*, 2018).

Spot centroids were then mapped to reciprocal space. We used a 1D fast Fourier transform algorithm to index the spots and estimate the orientation and unit-cell constants of the crystal to within a magnification scale factor (Sauter *et al.*, 2004; Winter *et al.*, 2018). The provisional *P*1 unit-cell constants were refined by enforcing constraints of the known space-group symmetry. For experimental data, it is unclear whether the intensity information from imaging will be sufficiently accurate to determine the Laue class. Further, systematically forbidden reflections may be present due to multiple scattering or lie in an unsampled region of reciprocal space, preventing the identification of screw axes (Hovmöller, 1992). However, symmetry could readily be determined by microED instead (see Section 5). Space-group determination is critical for locating the crystallographic phase origin, as the phases do not reflect the space-group symmetry until a crystallographic phase origin has been found (Hovmöller, 1992).

2.3. Bragg peak fitting

Peak fitting involves assigning pixels to Bragg peaks, followed by integrating the intensities and averaging the phases of the assigned pixels. In X-ray crystallography and microED, profile-fitting algorithms are used to integrate diffraction peaks (Pflugrath, 1999; Kabsch, 2010; Winter *et al.*, 2018; Hattne *et al.*, 2015). Profile fitting assumes a standard spot shape and models how each reflection is sampled based on its orientation with respect to the rotation axis, crystal mosaicity and beam divergence (Kabsch, 2010). In the case of tomographic data, however, a generic reflection profile cannot be assumed. Each reflection is only partially measured due to the large spacing between tilt increments and the lack of continuous rotation (Supplementary Fig. S3*a*), and at present we do not have adequate models to account for these effects. Developing a model to compensate for reflection partiality would improve the estimated intensities, but it would be challenging to devise a corresponding correction for the estimated phases.

Given these challenges, we implemented the following heuristic approach to determine reflection profiles from tomographic data. Firstly, the pixel coordinates of each lattice point in the Fourier transform of the tomogram were estimated from the indexing matrix of the crystal. The reflection was discarded if it was predicted to lie in the missing wedge or in the volume outside the $\pm 60^\circ$ or $\pm 40^\circ$ region spanned by the tilt series. A spherical subvolume centered on each retained reflection within a specified radius was then considered; for the simulated crystals analyzed here, we chose a radius of 7 Å^{-1}. Pixels were initially assigned to the peak if their intensities exceeded four standard deviations above the mean intensity of the subvolume (Figs. 2*a* and 2*b*). If multiple sets of noncontiguous pixels were found, the set with the centroid nearest to the predicted peak center was retained. The reflection was discarded if the observed peak centroid exceeded a distance of two reciprocal pixels from the predicted peak center. Such discrepancies between the ideal and observed Bragg positions resulted from peak centers being poorly sampled due to the large angular spacing between tilt increments. These partially measured reflections were rejected because the observed phase values were frequently shifted from the expected values by 180° (Supplementary Figs. S4 and S5).

To estimate the background intensity, the same subvolume used to assign peak pixels was considered. Both the pixels assigned to the peak and any contiguous pixels two standard deviations above the mean intensity of the subvolume were masked, with unmasked pixels considered as background. This threshold was less conservative than that used for assigning pixels to the peak to reduce the likelihood of including high-intensity outliers in the background region. We then estimated the values of the masked pixels by trilinear interpolation to approximate the background intensity beneath the peak. We used interpolation rather than assuming a constant background to better account for the anisotropic structure of the non-Bragg intensity, which results from the measured data being oriented along skew slices of finite thickness that are commonly separated by an angular increment of 2° or 3°. As a result, the orientation and magnitude of the background vary among peaks, depending on the position of the peak relative to the planes of sampled signal in Fourier space (Supplementary Fig. S3). The estimated background values were then subtracted from the original intensities of the corresponding pixels.

The assignment of pixels to Bragg peaks was then refined using the phase values (Fig. 2*c*). Specifically, the intensity-

weighted standard deviation of the phases of the assigned pixels was computed. If this metric exceeded a specified threshold, outlier pixels were iteratively removed until this threshold was reached. Empirically, we found that a threshold of 15° enabled the recovery of accurate phase values. Ensuring consistent phase values within the peak is critical, as the phase can change sharply outside the Bragg peak (Fig. 2*c*). Finally, the intensity and phase of the reflection were estimated from the sum of the intensity values and the intensity-weighted mean of the phase values of the retained pixels, respectively.

Both the intensity and phase components of the Bragg peaks were fitted using tomograms pre-processed as described in Section 2.1, which eliminated phase splitting. Although the application of a tapered cosine function in real space reduced the intensities, it also prevented sharp edges in the window of the Fourier transform from being convolved with the shape of the Bragg peaks in the frequency domain. As a result, the reflection profiles were more spherical and a larger number of Bragg peaks were retained by the peak-fitting algorithm described above, which benefited from this rounder shape.

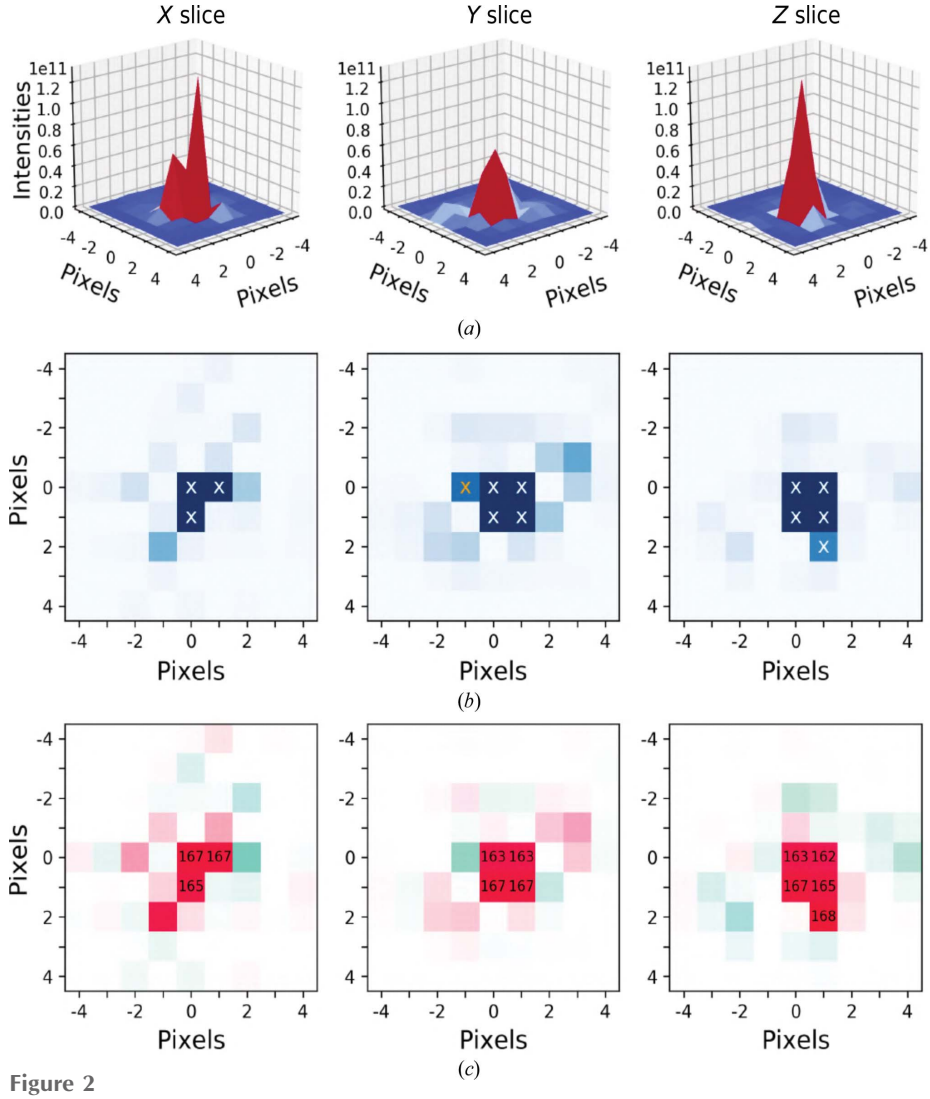


Figure 2
Peak fitting for a representative reflection. (*a*) The intensity profiles of slices are visualized along the indicated direction and centered on the (−2, 3, −8) reflection of a simulated crystal tomogram. In (*b*), the intensities are shown in 2D, with high-intensity pixels assigned to the peak marked by an X. In (*c*), the phases in the vicinity of the peak are visualized using the same color scheme as in Fig. 1. The phase values of pixels retained as part of the peak are noted in degrees. The high-intensity pixel marked by a yellow X in (*b*) had a phase value of 17° and was discarded as an outlier.

2.4. Merging data sets

In cryo-ET, the experimental geometry limits the accessible tilt range to ±70° relative to the untilted orientation of the sample (Wan & Briggs, 2016; Mastronarde, 1997). In principle, this rotation range should suffice to collect a complete data set from a single crystal for most point groups (Dauter, 1999). In practice, however, 140° is an overestimate of the useful rotation range because intermediate- to high-resolution information is lost both at high tilt angles due to increased specimen thickness and in images recorded late in the tilt series due to accumulated dose (Hagen *et al.*, 2017). Further, a popular data-collection strategy in cryo-ET uses a 3° tilt increment for a tilt range of ±60° (Hagen *et al.*, 2017),

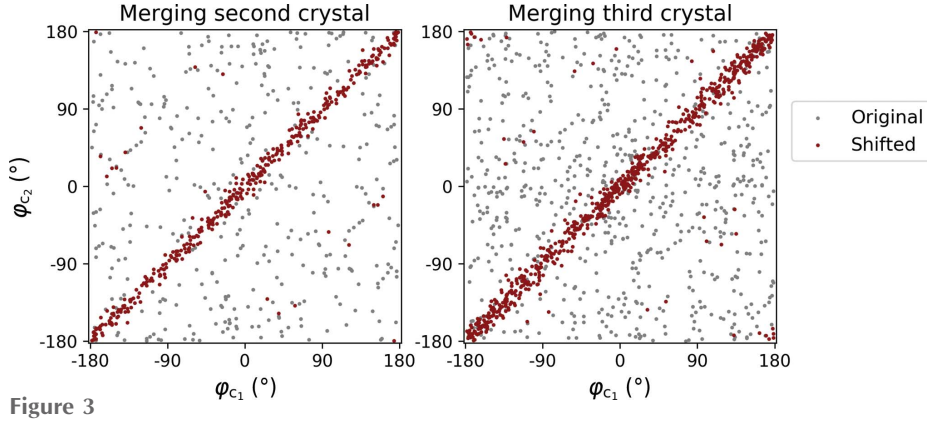


Figure 3
Identifying a common phase origin. The phases extracted from tomograms of three randomly oriented simulated crystals were merged. Each panel shows the relationship between the phases of reflections shared between the reference (φ_{c_1}) and added (φ_{c_2}) data sets before (gray) and after (red) shifting the latter to the reference origin determined by the first crystal. With each additional data set the number of reference reflections increases, while the origin remains fixed.

so a fraction of Bragg peaks in the nominal tilt range will fall between recorded images and will not be sampled. The size of this fraction is anticipated to vary between samples and to depend on factors such as the mosaicity and the resolution range. Achieving high completeness thus requires merging data from multiple crystals in different orientations.

Merging phases from different crystals requires positioning the data sets on a common phase origin. Unless this is also a crystallographic origin, the phases of symmetry-equivalent reflections are not related and the phase data effectively have $P1$ symmetry. Here, we established a common phase origin by treating one crystal as a reference and shifting the phase origins of the remaining data sets to this reference origin (Fig. 3). The fractional unit-cell row vector, $\mathbf{u}$, that shifts the phases of a second crystal to the reference origin was determined by minimizing the intensity-weighted mean residual between the reference phases and the shifted phases of the second crystal,

$$\text{argmin}_{\mathbf{u}} \frac{\sum_{\mathbf{h}} \bar{I}_c |[\varphi_{c_2}(\mathbf{h}) - 2\pi\mathbf{hu}] - \varphi_{c_1}(\mathbf{h})|}{\sum_{\mathbf{h}} \bar{I}_c}, \quad (1)$$

where the minimization is performed over all discrete values of $\mathbf{u}$ that fractionally sample the unit cell in real space according to a specified interval along each cell dimension. The sum is over the shared reflections between data sets c_1 and c_2, $\mathbf{h}$ is a column vector of Miller indices and $\bar{I}_c$ is the mean intensity for reflection $\mathbf{h}$. Equation (1) is similar to the phased translation function used in molecular-replacement searches, although here we compute intensity-weighted phase residuals rather than calculating complex structure factors (Read & Schierbeek, 1988). The term in square brackets, $[\varphi_{c_2}(\mathbf{h}) - 2\pi\mathbf{hu}]$, describes how the phases of c_2 change when its phase origin is shifted by $\mathbf{u}$. This expression was then used to position the phases of all reflections of c_2 on the reference origin. To merge the next crystal, the reference phase set was updated to include all reflections recorded in the original c_1 and c_2 data sets. The reference phase set thus expanded with each merge while the reference origin remained roughly fixed, with only minor adjustments after averaging the phases from different crystals. Data sets were merged in the order that maximized the number of shared reflections between the reference data and the data set being merged at each step. Once the final data set had been merged, the phase of each reflection was estimated as the intensity-weighted mean phase for all observations of that reflection.

Intensity information is unaffected by the choice of phase origin due to translational invariance. For consistency, however, we also treated the intensities as having $P1$ symmetry during merging. Intensities from a second data set, I_{c_2}, were uniformly scaled to the first data set, I_{c_1}, by minimizing the sum of squared residuals,

$$\text{argmin}_{m,b,\sigma} \sum_{\mathbf{h}\in(c_1\cap c_2)} \left\{\log[m\exp(-q^2\sigma^2)I_{c_2}(\mathbf{h}) + b] - \log[I_{c_1}(\mathbf{h})]\right\}^2, \quad (2)$$

where q is the magnitude of the reciprocal-space vector associated with reflection $\mathbf{h}$, the exponential term is the Debye–Waller factor (Blessing *et al.*, 1996) and least-squares optimization is used to determine the scaling parameters m, b and σ. We found that the use of logarithmic residuals improved the stability of the optimization algorithm. The next data set was then scaled to the averaged intensities of the original c_1 and c_2 data sets. Once all data sets had been merged, the intensity of each reflection was computed as the mean of all observations of that reflection.

2.5. Locating a crystallographic phase origin

As noted above, it is unlikely that the phase origin of the merged data set will coincide with a crystallographic origin consistent with the space group of the crystal. In 2D electron crystallography, the crystallographic origin was found by testing fractional cell positions in the plane of the repeating unit for the fulfillment of phase constraints imposed by symmetry (Hovmöller, 1992; Amos *et al.*, 1982). If the plane group was not known in advance, this origin-refinement procedure was performed for every possible plane group, and the plane group that yielded the lowest phase residual was selected. Crystal symmetry determination is easier for 2D crystals, however, as there are only 17 plane groups in contrast to 230 space groups (Hovmöller, 1992). Here, we assumed that the space group was already known and extended the method of origin refinement to 3D crystals as follows.

The unit cell was sampled using a real-space grid with equally spaced nodes along each dimension, and each node or fractional cell position was considered a candidate origin. The phases of the merged data, φ_0, were shifted by the fractional cell vector, $\mathbf{u}$, to this origin,

$$\varphi_{\mathbf{u}}(\mathbf{h}) = \varphi_0(\mathbf{h}) - 2\pi\mathbf{hu}. \quad (3)$$

We then evaluated the following three metrics. The first metric corresponded to the phase residual of symmetry-equivalent reflections. The mean phase value for a given reflection $\mathbf{h}$ can be estimated from its symmetry-equivalent reflections as follows (Hovmöller, 1981),

$$\langle\varphi_{\mathbf{u}}(\mathbf{h})\rangle = \langle\varphi_{\mathbf{u}}(\mathbf{h}_s) - 2\pi\mathbf{h}_s\mathbf{t}_s\rangle, \quad (4)$$

where the symmetry-equivalent reflections $\mathbf{h}$ and $\mathbf{h}_s$ are related by the translation vector $\mathbf{t}_s$, and the phase origin is $\mathbf{u}$. We then computed the symmetry phase residual as the sum of the phase differences between this mean value and independent observations from the symmetry-equivalent reflections,

$$R_{\text{sym},\,\varphi_{\mathbf{u}}} = \frac{\sum_{\mathbf{h}}\sum_{\mathbf{h}_s} I(\mathbf{h}_s)|[\varphi_{\mathbf{u}}(\mathbf{h}_s) - 2\pi\mathbf{h}_s\mathbf{t}_s] - \langle\varphi_{\mathbf{u}}(\mathbf{h})\rangle|}{\sum_{\mathbf{h}}\sum_{\mathbf{h}_s} I(\mathbf{h}_s)}, \quad (5)$$

where the double sum enumerates the set of unique reflections and, for each, the set of symmetry-equivalent reflections. The residual was intensity-weighted to favor stronger reflections.

The second metric considered the difference between centric reflections and their expected phase values. Centric reflections satisfy the condition $\mathbf{hR} = -\mathbf{h}$, where $\mathbf{R}$ is the rotation matrix associated with reflection $\mathbf{h}$ (Hovmöller, 1981). When reflection data are positioned on a crystallographic phase origin, the phases of centric reflections are restricted to

a limited set of possible values. We computed the intensity-weighted mean residual between the observed phase values at each candidate origin **u** and the expected phase values for these reflections as follows,

$$R_{\mathrm{cen},\,\varphi_{\mathbf{u}}} = \frac{\sum_{\mathbf{h}} I_{\mathbf{h}} \min\{|\varphi_{\mathbf{u}}(\mathbf{h}) - \pi\mathbf{ht}|, |\varphi_{\mathbf{u}}(\mathbf{h}) - \pi(\mathbf{ht}+1)|\}}{\sum_{\mathbf{h}} I_{\mathbf{h}}}, \quad (6)$$

where **t** is the translation vector associated with reflection **h** and the sum is restricted to the observed centric reflections (Rupp, 2010). This metric was omitted for data sets that did not contain any centric reflections.

For the third metric, an electron-density map of the unit cell was computed from the intensities of the merged data set and the shifted phases, $\varphi_{\mathbf{u}}$, using the *cctbx* library (Grosse-Kunstleve *et al.*, 2002). The skew of the density values was evaluated, as the density distribution of macromolecular crystals exhibits a positive skew, in contrast to the Gaussian distribution characteristic of random maps. This distinction is used to judge map quality during automated structure solution after experimental phasing (Terwilliger *et al.*, 2009). The advantage of this metric is that it uses all available reflections, rather than just the subset of either centric reflections or reflections with high multiplicity. For consistency with the phase residual metrics, the negative of the skew was computed such that lower values indicated more probable origins.

These three metrics were normalized and summed with equal weighting to score each candidate origin given by the fractional cell vector, **u**. The fractional cell position with the lowest combined score was selected, and the phases of the merged data set were shifted to this origin. The intensities and shifted phases were then reduced to the asymmetric unit and symmetry constraints were imposed. Specifically, the phases of symmetry-equivalent reflections were mapped to their expected values in the asymmetric unit (see equation 4) and the intensity-weighted mean phase value was computed. The intensities of symmetry-equivalent reflections were averaged.

As an alternative approach, individual data sets could be positioned on a crystallographic origin prior to merging. In this case, the merging procedure would require a search over a limited number of positions rather than the entire unit cell, as space-group symmetry restricts the number of crystallographic origins. However, we found that the process of merging two data sets was slightly more robust than positioning individual data sets on a crystallographic origin. We compared each strategy using simulated data sets of a $P2_12_12_1$ crystal with unit-cell dimensions a = 16.2, b = 29.1, c = 47.7 Å generated with a range of mean phase errors (0–40°), completeness (10–40% prior to accounting for space-group symmetry) and relative B factors (described in more detail in Section 3; see equation 7). The searches for the common origin between data sets and the crystallographic origin were both performed using a sampling interval of 0.2 Å; the phase origin was considered to be correctly found if the distance between the correct origin and the origin found by the algorithm was less than twice the sampling interval. We observed a 4% higher success rate for finding a common phase origin between two data sets than for locating a crystallographic origin for each individual data set (Supplementary Fig. S6*a*). Further, identifying a crystallographic origin was successful in 8% more cases when the procedure was performed on the merged data from two tomograms than on a single tomogram (Supplementary Fig. S6*b*). In contrast, the success rate for merging two and three data sets was similar (Supplementary Fig. S6*c*). These trends suggest a modest advantage to merging data sets to a common phase origin prior to finding a crystallographic origin, as the latter procedure appeared to be more sensitive to the number of available reflections.

3. Validation of the workflow using simulated crystals

We validated the full workflow presented in Section 2 (Supplementary Fig. S1*a*) by evaluating the accuracy of the reflection data recovered from simulated tomograms. The processed data were compared with phases and intensities computed directly from the atomic model, in addition to reflection data recovered from 'intact volumes' processed in the same manner as the simulated tomograms. Intact volumes refer to rotated crystal densities generated in the same manner as the simulated tomograms (see Section 2), except without projection into tilt series and reconstruction into tomograms. This comparison between intact volumes and simulated tomograms enabled an assessment of the inaccuracy and loss of completeness introduced by tomographic sampling beyond the baseline interpolation errors from simulating and rotating the crystal densities onto a discrete grid. For each structure (PDB entries 6d6g and 4bfh) and type of volume (intact volume/tomogram), performance was judged based on merging five data sets. This merging procedure was repeated ten times with unique sets of five data sets.

Metrics evaluating the quality of the recovered reflection data are shown in Table 1. The information from the intact volumes was consistent with the reference phases ($R_{\mathrm{ref},\varphi} < 2.5°$) and intensities (CC > 0.97) computed from the initial atomic models. For the simulated tomograms, merging multiple data sets was required to achieve high completeness. When the tilt range spanned 120°, or 67% of reciprocal space, the data recovered from each tomogram were only 40% complete in the absence of internal symmetry (Table 1). The unexpectedly low completeness stems from the large angular increment of the tilt series: many Bragg peaks lie between tilt images and were either not observed or discarded as poorly sampled. Generating tomograms using a ±40° tilt range with 2° increments rather than a ±60° tilt range with 3° increments resulted in a slight decrease in the overall completeness (∼5%) but a modest improvement in the accuracy of the reflection data due to the finer sampling (Table 1). Merging tomograms and the presence of internal symmetry also improved the accuracy of the recovered phases, which showed good agreement with the reference ($R_{\mathrm{ref},\varphi}$ of 6.2° and 2.8° for the $P1$ and $P2_12_12_1$ crystal systems, respectively, when the tilt range spanned ±60°).

In real space, loss of information due to the missing wedge leads to elongation of density in the direction of the beam (Radermacher, 2007). These anisotropic effects were apparent

Table 1
Quality of merged data.

	Intact volumes		Undamaged tomograms				Mildly damaged tomograms, c = 0.3				Moderately damaged tomograms, c = 0.2				Severely damaged tomograms, c = 0.1			
			±60°		±40°		±60°		±40°		±60°		±40°		±60°		±40°	
No. of crystals merged	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5	1	5
PDB entry 6d6g ($P1$)																		
Completeness c†	0.96	0.99	0.39	0.89	0.34	0.85	0.30	0.72	0.30	0.78	0.20	0.55	0.20	0.58	0.10	0.31	0.10	0.33
$R_{\mathrm{ref},\varphi}$‡ (°)	2.4	2.6	9.2	6.2	7.4	5.5	18.9	15.0	10.4	8.8	20.2	20.9	13.6	13.6	19.5	22.1	12.3	13.6
CC(I_{sim}, I_{ref})§	1.00	1.00	0.71	0.74	0.79	0.81	0.59	0.57	0.65	0.65	0.51	0.44	0.53	0.48	0.47	0.42	0.54	0.48
CC(map, model)¶	0.98	0.99	0.58	0.85	0.56	0.88	0.50	0.72	0.50	0.79	0.42	0.60	0.41	0.64	0.34	0.52	0.34	0.58
PDB entry 4bfh ($P2_12_12_1$)																		
Completeness c	0.95	1.00	0.73	0.99	0.67	0.99	0.61	0.96	0.62	0.97	0.46	0.87	0.45	0.87	0.25	0.63	0.25	0.63
$R_{\mathrm{ref},\varphi}$ (°)	1.0	1.0	6.1	2.8	5.2	2.2	11.0	4.9	8.4	3.5	16.8	12.6	12.2	8.1	24.6	27.0	11.9	15.3
CC(I_{sim}, I_{ref})	0.98	0.98	0.81	0.87	0.86	0.90	0.72	0.76	0.73	0.76	0.59	0.55	0.59	0.57	0.56	0.46	0.58	0.51
CC(map, model)	0.97	0.99	0.79	0.92	0.78	0.95	0.69	0.87	0.72	0.88	0.58	0.74	0.59	0.74	0.46	0.60	0.44	0.61

† To a high-resolution limit of 3.3 Å. Entries show the mean value from ten runs of merging the indicated number of data sets consisting of intact volumes, undamaged tomograms or tomograms simulated with radiation damage to achieve the specified fractional $P1$ completeness (c). ‡ The intensity-weighted mean residual between phases from the data sets reduced to the asymmetric unit and reference phases computed directly from the atomic model. § The correlation coefficient between the logs of reflection intensities computed from the reduced data set and the atomic model. ¶ The correlation coefficient between the real-space maps computed from the reduced data set and the atomic model.

in density maps computed from single tomograms of the $P1$ crystal, with continuous density along parallel bands and gaps along the protein backbone between these streaks (Fig. 4*a*). Incorrect connectivity was less pronounced in density maps computed from single tomograms of the orthorhombic crystal, as the presence of internal symmetry mitigated the directionality of the missing-wedge effect (Fig. 4*b*). In both cases, merging multiple crystals improved the correlation with the reference map (Table 1).

4. Tests of robustness to radiation damage

Radiation damage is a limiting factor in cryo-ET and results in a loss of information, particularly at high resolution, as data collection progresses (Baker & Rubinstein, 2010). Application of our data-processing workflow to experimental tomograms collected from proteinase K, lysozyme and ferritin nanocrystals using a conventional data-acquisition scheme revealed characteristics consistent with radiation damage (Supplementary Section S1 and Fig. S7). Although the Fourier transforms of the tomograms could be consistently indexed and yielded cell constants that matched those of X-ray crystal structures, we were unable to locate a common phase origin between data sets (Supplementary Figs. S7*b* and S7*d*). We suspected that the failure to merge data sets stemmed from the low overall completeness (∼10–20% to 7 Å resolution in $P1$) of the individual tomograms. However, we also observed that the recovered reflections were not uniformly spread across the tilt range but instead concentrated at low tilt angles acquired early in data collection (Supplementary Fig. S7*c*). We thus sought to use simulations to determine the strategies needed for successful data collection and merging.

In our simulations, damage events were assumed to occur at random sites in the crystal volume, and each 'hit' was modeled as a blurring of the local density (Atakisi *et al.*, 2019). The blurring was performed by replacing a cubic subvolume of length 5 Å around each selected site with its Gaussian-filtered copy, using a standard deviation of 1 Å along each axis for the 3D Gaussian kernel. Following a dose-symmetric tilt scheme (Hagen *et al.*, 2017), an equal number of hits was applied to the randomly oriented crystal before computing each projection image to mimic a linear increase in absorbed dose across the tilt series. Tomograms were reconstructed from these damaged tilt series, and those that could not be indexed in the correct space group were discarded. For both crystal systems, we calibrated the total number of hits, or effective dose, that on average yielded tomograms with a completeness of 30%, 20% or 10% before accounting for space-group symmetry, compared with the ∼40% $P1$ completeness observed for undamaged tomograms (Table 1). To focus on the tolerance of data processing to low completeness, ten unique sets of five tomograms were merged if the completeness of each tomogram and the average of the set was within 2% and 1%, respectively, of the target completeness. Metrics evaluating the average quality of the individual tomograms and merged data sets are reported in Table 1.

For both crystal systems, the reference structure was consistently recovered by merging tomograms that were each 30% complete prior to accounting for space-group symmetry. Merging improved the accuracy of the phases and resolved the connectivity errors observed in real-space maps computed from a single damaged tomogram (Table 1, Fig. 5). While an initial completeness of 20% could also be tolerated by the orthorhombic system, results for the triclinic crystal were variable, with the phase accuracy frequently decreasing during the course of merging despite the increase in multiplicity (Supplementary Fig. S8). For both crystal systems, an initial completeness of 10% per tomogram was prohibitive to recovering the reference structure. Despite the modest increase in cross-correlation with the reference map, merging reduced the phase accuracy under this starting condition, indicating that a common phase origin was not correctly found (Supplementary Fig. S8). The density of the merged tomograms showed the wrong connectivity, including smearing of the density in the solvent region between neighboring chains and gaps along the backbone (Fig. 5, right). In several cases, the merging

procedure visibly aligned the missing wedges of the tomograms being merged (Supplementary Fig. S9), suggesting that the shape of the missing wedge rather than the reflection data dominated the signal. Similar trends in the relationship between initial completeness and failure to merge were observed when damaged tomograms were reconstructed from tilt series spanning either a ±60° tilt range with 3° increments or a ±40° tilt range with 2° increments, despite the lower initial phase errors of the latter (Table 1).

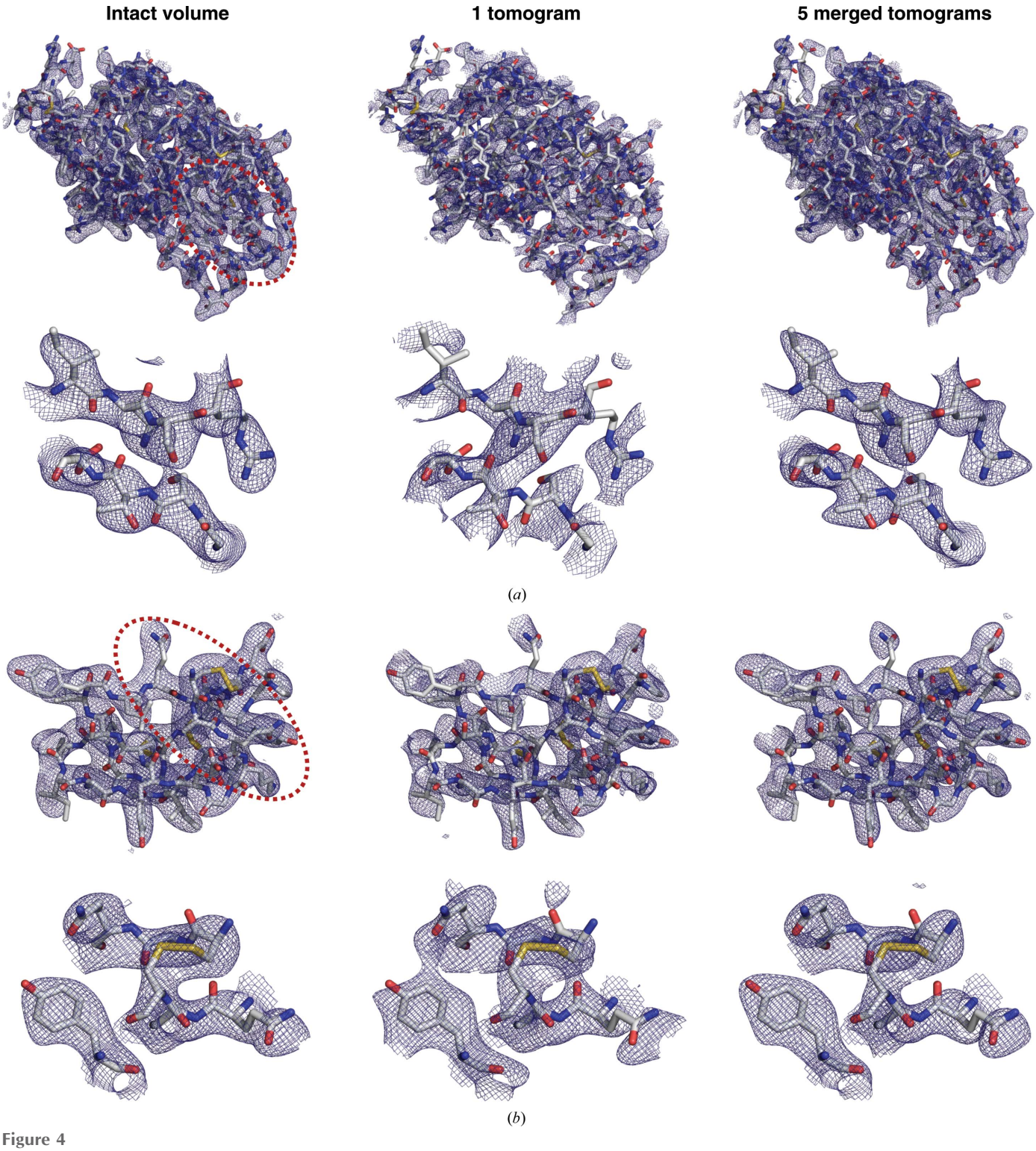


Figure 4
Merging undamaged tomograms recovers the reference density. Density maps were computed from reflection data for a representative intact volume (left), one tomogram (center) or five merged tomograms (right). The simulated crystal system had either (*a*) triclinic or (*b*) orthorhombic symmetry. Each panel visualizes the map for the entire protein (top) and the subset of residues circled in red (bottom).

Although this analysis suggests that >10% completeness per data set is required for recovery of the correct structure, loss of overall completeness is not the only hallmark of radiation damage. The progressive accumulation of damage alters the spatial distribution of the reflection data by reducing the number of reflections recorded late in the tilt series, or at high tilt angles for a dose-symmetric scheme. In addition, the simulated damage introduced phase errors, although the loss of phase accuracy varied for the same amount of simulated dose (Table 1). To disentangle the impact of incompleteness, the angular spread of reflections across the tilt range and phase error on the ability to merge tomograms, we simulated radiation damage in reciprocal space according to the following model. For each simulated data set, the crystal lattice was subjected to a random rotation in 3D. Structure factors were then computed to 3.3 Å resolution, and the positions of Bragg peaks were predicted as a function of tilt angle; reflections in the missing-wedge region were excluded. Reflection intensities were modeled as decaying with a B factor that increased linearly with dose (Atakisi *et al.*, 2019),

$$I_{\rm d}(\mathbf{h}) = I_0(\mathbf{h}) \exp\left(-\frac{1}{16\pi^2} n \tilde{B}_f q^2\right), \tag{7}$$

where I_0 is the intensity of the undamaged reflection, q is the magnitude of the reciprocal-space vector associated with reflection $\mathbf{h}$, $\tilde{B}_f$ is a relative B factor and n is the image number in the tilt series, with n = 1, 2, 3, ... corresponding to tilt angles 0°, −3°, +3°, −6°, ... or 0°, −2°, +2°, −4°, ... for a dose-symmetric scheme spanning ±60° or ±40°, respectively. The highest-intensity reflections in the simulated tilt range were retained to achieve the specified initial completeness. The phase origin was subjected to a random fractional shift (see equation 3), and phase errors were drawn from a normal distribution with the standard deviation chosen to achieve the

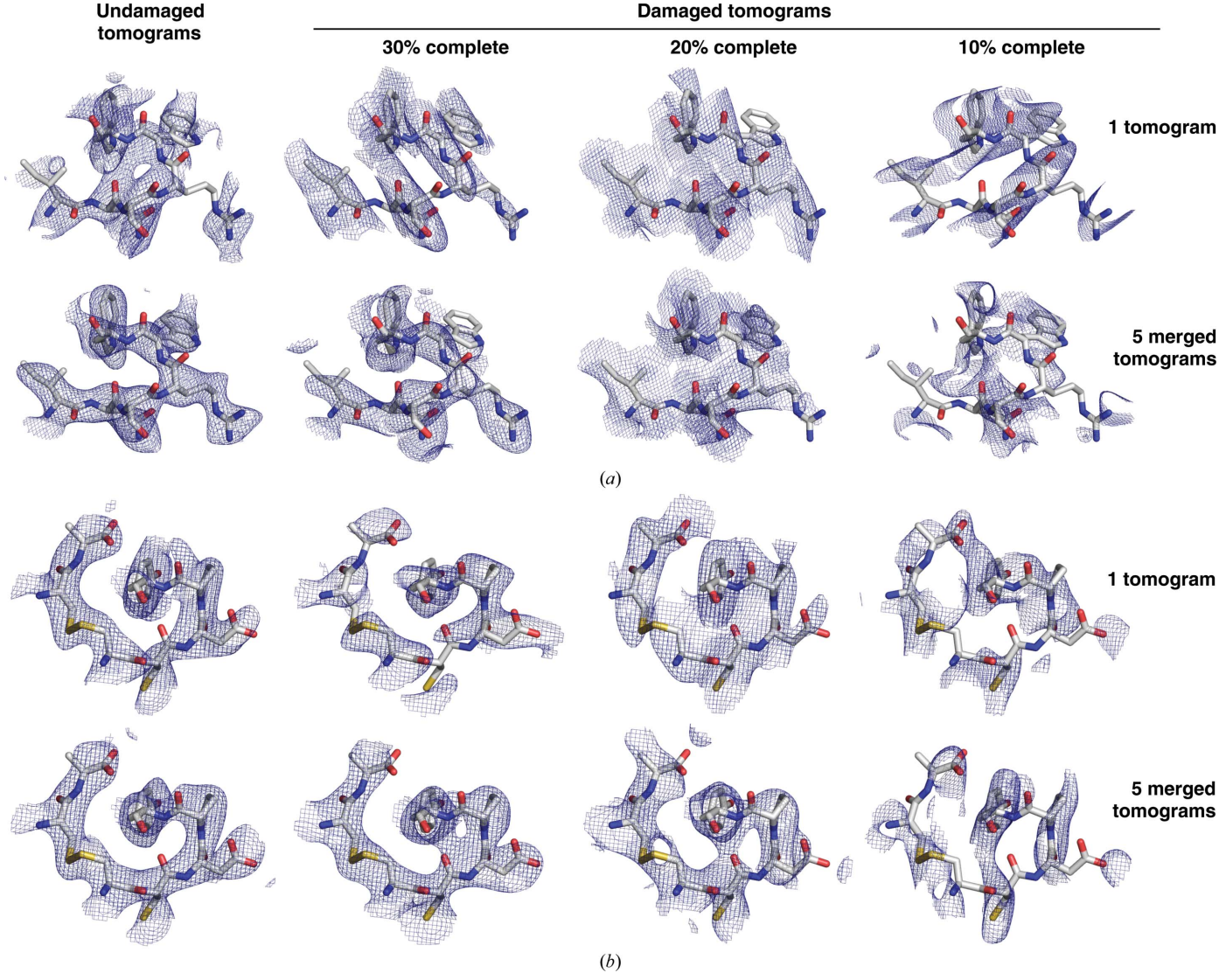


Figure 5
Merging tomograms with damage-induced loss of completeness. Density maps were computed from reflection data recovered from a representative undamaged tomogram or a damaged tomogram with the specified $P1$ completeness (top) and after merging five tomograms of the indicated type (bottom). A subset of residues is visualized for the (*a*) triclinic and (*b*) orthorhombic crystal systems.

target mean phase error. We assumed that the phase errors were dose-independent, since to our knowledge there are no models of how phase accuracy decays as a function of dose. This error model also assumes that the effects of tilt and the contrast transfer function, which modulate reflection phases with predictable behavior, can be corrected for (Henderson *et al.*, 1990). The advantage of this damage model is that it allowed us to independently tune the phase accuracy, overall completeness and angular spread of the reflection data across the tilt range. Data were generated in this manner for the triclinic and orthorhombic systems used above in addition to a tetragonal crystal of proteinase K (PDB entry 2id8; Wang *et al.*, 2006), which has an ∼20-fold larger unit cell by volume and higher symmetry. Reflection data were then merged as described in Section 2.4.

Structure solution from tomograms of nanocrystals depends on finding the origin shift that correctly aligns the phases of multiple data sets. We assessed the accuracy of this merging procedure based on the merge error: the magnitude of the vector difference between the fractional origin shift estimated by our merging algorithm and the true shift required to align two data sets each subjected to a random phase shift. We considered the correct origin to be found when the merge error was less than twice the sampling interval used by the merging algorithm, corresponding to a fractional merge error of <0.03 for each crystal system. We then examined the dependence of this merge error separately on the phase accuracy, overall completeness and the spatial distribution of the reflections of the data sets being merged. For the latter, we measured how unevenly the reflection data were distributed across the tilt range using the Jensen–Shannon (JS) distance (Lin, 1991). This statistical metric measures the difference between a probability distribution of interest, p, and a reference probability distribution, q,

$$\mathrm{JS} = \left\{\frac{1}{2}[D(p||m) + D(q||m)]\right\}^{1/2}, \tag{8}$$

where p is the normalized distribution of tilt angles for the observed reflections, q is the normalized uniform distribution spanning ±60° and $m = \frac{1}{2}(p + q)$. D is the Kullback–Leibler divergence given by

$$D(x||y) = \sum_{i,x(i)\neq 0} x(i) \log \frac{x(i)}{y(i)}, \tag{9}$$

where the width of the angular bin i was set to 1°. Using our chosen reference distribution, the Jensen–Shannon distance measures how unevenly reflections are distributed across a full tilt range of ± 60°, and the score is independent of the completeness of the data set. Data sets that span a narrower tilt range than ±60° are penalized even when the reflection data are uniformly distributed since the normalized counts differ from the expected frequencies (Supplementary Fig. S10). Radiation damage also increases the Jensen–Shannon distance due to the loss of Bragg peaks at higher tilt angles as data collection progresses (Supplementary Fig. S10).

The results from the three simulated crystal systems were pooled to examine trends that held across different space-group symmetries and a range of unit-cell volumes. We found that the initial phase error and the angular spread of the reflections across the tilt series, as measured by the Jensen–Shannon distance, better predicted the likelihood of merging success compared with overall completeness (Fig. 6*a*). Provided that the reflections from individual tomograms were approximately evenly spread across the tilt range (JS < 0.18), the correct origin shift could be determined even in the presence of an average phase error of up to 40°. By contrast, even relatively high completeness (>50%, without accounting for space-group symmetry) did not guarantee finding the correct phase origin in the presence of moderate phase errors when the reflections were not uniformly distributed in reciprocal space (Figs. 6*b* and 6*c*). Consistent with these findings, we observed that our experimental data sets were characterized by high JS scores of ∼0.5 (Supplementary Fig. S7*c*), which predict a low chance of merging success. These trends argue for distributing the dose across as wide a tilt range as possible to maximize the angular spread of reflections available from each tomogram. Although the real-space simulations indicated that a data-collection scheme spanning ±40° with 2° increments would reduce phase errors relative to ±60° with 3° increments due to the narrower tilt increment (Table 1), these results predict that the modest increase in phase accuracy will be outweighed by the increased difficulty in correctly merging data sets due to the narrower angular distribution of the reflection data across the tilt range.

5. Discussion

Cryo-ET of protein nanocrystals could deliver a method of high-resolution structure determination that both retains experimental phases and circumvents the need for large crystals in conventional X-ray crystallography and high-molecular-weight proteins in SPR. Here, we describe a data-processing pipeline to solve structures from tomograms of nanocrystals. This workflow both leverages software from X-ray crystallography and provides new algorithms to handle challenges unique to tomographic data from crystalline specimens. These challenges are related to correctly extracting the reflection phases, including eliminating phase splitting due to imperfect periodicity, excluding partial reflections that result in phase errors of 180° and extending the phase-origin search procedure from 2D plane groups to 3D space groups. We validated this data-processing scheme using simulated crystals and found that the recovered reflection information was accurate, yielding maps with a correlation coefficient to the reference of ∼0.9 after merging five tomograms without any structure refinement.

We also assessed the robustness of this pipeline to radiation damage, which limits the effectiveness of structure-determination methods and was evident in our initial analysis of experimental data. Merging tomograms of crystals in different orientations is required to compensate for loss of completeness, which results from both radiation damage and the

missing wedge. We found that this merging procedure was especially sensitive to the angular spread of the reflection data across the tilt range and phase errors, but less so to the completeness of the individual tomograms. These simulations indicate a trade-off between a wider tilt range to facilitate merging data sets and a finer tilt increment to reduce phase errors. Since including more data sets can overcome phase errors to some extent but not incorrect origin shifts, these results recommend data-collection strategies that maintain a wide tilt range rather than decreasing the sampling interval. We also predict that a wider tilt range would be favored over finer sampling for high-resolution subtomogram averaging to similarly avoid aligning the missing wedge experienced by individual particles.

Once experimental data can be collected using an acquisition scheme that facilitates merging tomograms, additional challenges will remain. One critical issue is to determine the optimal specimen thickness: thicker crystals increase the signal to noise, but the accuracy of the signal suffers due to a larger defocus gradient and increased multiple scattering (Henderson *et al.*, 1990; Subramanian *et al.*, 2015). We anticipate that the defocus gradient can largely be accounted for by tailoring existing software to perform a 3D correction of the contrast transfer function for nanocrystal specimens (Jensen & Kornberg, 2000; Turoňová *et al.*, 2017). Another concern is the impact of multiple or dynamical scattering on data quality. Although multi-slice simulations have predicted that multiple scattering would prevent structure solution from crystals thicker than 0.1 µm (Subramanian *et al.*, 2015), microED has shown empirically that multiple-scattering effects introduce only marginal (∼5%) errors in the measured intensities for crystals with a thickness of 0.5–1 µm (Shi *et al.*, 2013; Nannenga, Shi, Leslie *et al.*, 2014; Martynowycz *et al.*, 2017). This conflict between theory and experiment is likely to arise from the simulations failing to account for lattice disorder, bulk solvent, crystal orientation along non-zone axes and continuous rotation. Although high-resolution cryo-ET data cannot yet be collected using continuous rotation, these other mitigating factors suppress multiple scattering. Whether the phase errors introduced by multiple scattering will be as low as the intensity errors observed by microED remains unknown. However, the impact of multiple scattering and the contrast transfer function on phase accuracy can only be quantified once a merged data set is sufficiently complete to enable positioning of the phases on a valid crystallographic phase origin.

Another common tomography challenge is the poorer image quality at high tilt angles. Given the sensitivity of the

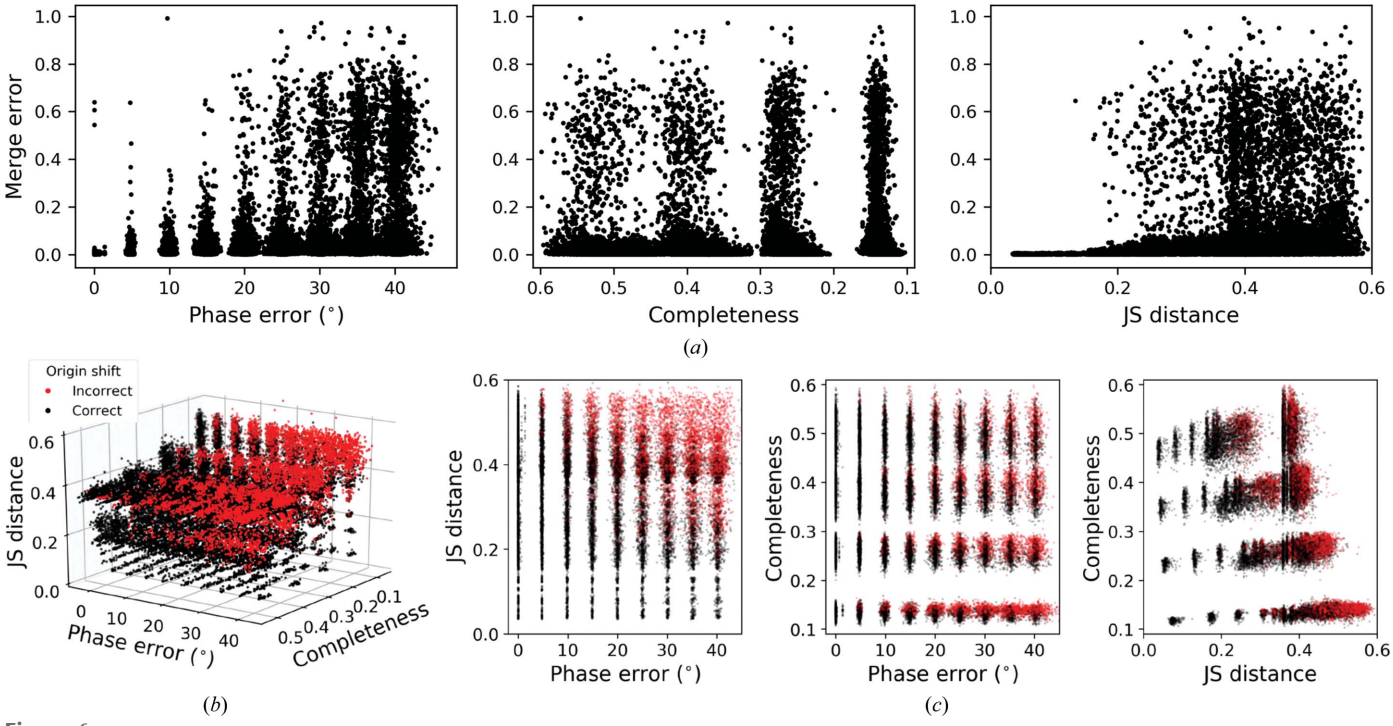


Figure 6
The dependence of merging success on phase accuracy, completeness and the angular spread of reflection data. A total of 25 920 simulated data sets were generated across three crystal systems, spanning a range of *P*1 completeness, phase errors and relative *B* factors. Pairs of data sets with similar starting characteristics and in random orientations were merged. (*a*) The fractional merge error was computed as the magnitude of the vector error between the true and estimated fractional shifts required to position data sets on a common phase origin. This metric is shown as a function of the mean phase error (left), *P*1 completeness (center) and how unevenly spread the reflections were across the tilt range (right) for the data sets being merged. The latter was estimated as the Jensen–Shannon (JS) distance to a uniform angular distribution spanning ±60° (see equation 8). Although 2880 pairs of data sets with a phase error of 0° were merged, the results are visually superimposed in the leftmost panel. Merge errors are shown as a function of (*b*) these three parameters or (*c*) the two indicated parameters of the data sets being merged. In (*b*) and (*c*), red and black indicate data sets for which the incorrect and correct origin shift, respectively, was determined by the merging algorithm. The error threshold was set to be twice the sampling interval used by the algorithm, such that the correct phase origin was found only when the fractional merge error was <0.03.

merging procedure to the distribution of the reflections in reciprocal space, simulations that additionally account for the loss of image quality at high tilt angles may be useful for further refining data-collection strategies from nanocrystals. Finally, stage limitations that preclude data collection beyond $\pm 60^{\circ}$ combined with preferred crystal orientations on the grid could prevent high completeness being achieved for low-symmetry space groups even after merging multiple tomograms. However, to our knowledge only one crystal system (catalase) has shown a preferred orientation (Nannenga, Shi, Hattne *et al.*, 2014), so we do not anticipate this to be a general problem. When such cases are encountered, different grid-surface chemistries could be explored to randomize specimen orientation (Drulyte *et al.*, 2018). While our simulations advise pursuing a nontraditional data-collection strategy that enables recording the most uniform distribution of Bragg peaks in reciprocal space, it will still be a significant undertaking to fully explore the optimal dose per tilt angle, crystal size and defocus that achieve this aim in light of these experimental and technical considerations.

Despite these challenges, recent successes in cryo-ET of noncrystalline specimens support the application of this technique to solve structures from nanocrystals to high resolution. In particular, subtomogram averages of purified immature HIV-1 Gag particles to 3.1 Å resolution and of *in situ* ribosomes to 3.7 Å resolution demonstrate that high-resolution signal is available in tilt-series data (Himes & Zhang, 2018; Tegunov *et al.*, 2021). As with SPR, only particles with high molecular weight can be aligned with sufficient accuracy for this high-resolution information to be retained during subtomogram averaging. Alignment requires both determining the relative orientations of particles and positioning them on a common phase origin.

In comparison to SPR, the relative orientations of constituent proteins in a nanocrystal can be readily determined by indexing the coherent signal from the entire crystal. Even tiny nanocrystals will have a higher molecular weight than an individual HIV-1 Gag hexamer or ribosome, thereby extending the reach of cryo-ET to smaller proteins. In comparison to X-ray crystallography, both microED and cryo-ET of nanocrystals require a wider angular range of reflection data for indexing due to the negligible curvature of the Ewald sphere using electron radiation. Although successful indexing of individual electron diffraction stills has been reported (Gevorkov *et al.*, 2020), in general a tilt range of $\pm 20^{\circ}$ is considered to be necessary to index microED data (Nannenga & Gonen, 2019). Cryo-ET data pose additional challenges, as imaging but not diffraction data are negatively impacted by several factors including the defocus gradient, translational drift and loss of coherence due to inelastic scattering. These effects have been predicted to cumulatively reduce the signal-to-noise ratio of projection images by over 90% compared with diffraction stills collected from the same crystal, with increasing loss at higher resolution and for thicker crystals (Clabbers & Abrahams, 2018). The diminished signal to noise will result in fewer observed reflections when collecting data in imaging versus diffraction mode for a fixed dose or will exacerbate radiation damage if the dose is increased to compensate. Either scenario could make indexing more difficult for cryo-ET data compared with microED data. Here, we adapted algorithms from the *DIALS* crystallographic software package to index the Fourier transforms of simulated nanocrystal tomograms (Waterman *et al.*, 2013; Winter *et al.*, 2018). We found that indexing was robust even for simulated tomograms with low completeness.

While the use of nanocrystals facilitates the determination of relative particle orientations, the low molecular weight of the constituent proteins makes finding a common phase origin, the other step in particle alignment, more challenging. High-resolution subtomogram averaging has exclusively targeted high-molecular-weight proteins, and with two exceptions, isolated particles in solvent (Schur, 2019; Himes & Zhang, 2018; Tegunov *et al.*, 2021; Bharat *et al.*, 2017). These characteristics facilitate finding a common phase origin from the center of mass of each particle. For nanocrystals composed of small and densely packed proteins, a more extensive search over the unit-cell volume is required to position data sets on a consistent phase origin. Finding a common phase origin to merge data sets is critical for improving the accuracy of the recovered reflection data and overcoming low completeness.

In the future, we anticipate that cryo-ET of nanocrystals could enable structure determination from disordered crystals, which are typically discarded during diffraction experiments. In crystals, disorder attenuates the high-resolution signal and has been observed on length scales relevant for nanocrystals (Nederlof *et al.*, 2013; Gallagher-Jones *et al.*, 2019). In contrast to diffraction methods, imaging permits spatial characterization and computational correction of such disorder. One approach is to denoise the tomogram in Fourier space, followed by computationally 'unbending' lattice distortions using cross-correlation analysis between the denoised and original tomograms. Lattice unbending overcame the resolution-limiting effects of disorder in 2D electron crystallography, in which specimens were frequently bent or wrinkled (Schenk *et al.*, 2010; Henderson *et al.*, 1990). Alternatively, subtomogram averaging could be performed in real space (Wan & Briggs, 2016). This technique has provided subnanometre reconstructions of pleomorphic viruses with imperfect helical symmetry and could prove similarly useful for solving structures from tomograms of disordered nanocrystals (Wan *et al.*, 2017). In addition to extending the high-resolution limit of structure determination, the ability to spatially characterize disorder could provide insights into both the organization of proteins that form ordered arrays *in vivo* and defects in these biological crystals (Lange, 1974; Wolf *et al.*, 1999; Dadinova *et al.*, 2019).

In addition, we anticipate that the development of a hybrid method involving microED and cryo-ET of nanocrystals could become routine. This approach would combine diffraction intensities with low-resolution phases from experimental images to solve an initial structure, followed by phase extension to the high-resolution limit of the microED data (Taylor, 2003; Cowtan, 2010). Combining these techniques will be facilitated by the fact that sample preparation is shared

between the techniques and data can be collected using the same microscope and detector (Rodriguez *et al.*, 2017; Hattne *et al.*, 2019). Further, a single microED data set can be collected in as little as one minute (de la Cruz *et al.*, 2019), so the additional time burden would be negligible. Although the collection of a cryo-ET tilt series generally takes 20 min, faster acquisition schemes are under active development with the goal of reducing collection time to rival that of microED (Chreifi *et al.*, 2019, 2021). Implementation of both disorder corrections and a hybrid approach with microED will be critical to realize the full potential of cryo-ET of nanocrystals for high-resolution structure determination.

6. Code availability

The code developed to process tomograms of nanocrystals is available at https://github.com/apeck12/cryoetX.

7. Related literature

The following references are cited in the supporting information for this article: Tivol *et al.* (2008), Mastronarde (2005), Zheng *et al.* (2017) and Xiong *et al.* (2009).

Acknowledgements

We thank Lauren Ann Metskas and Florian Schur for valuable discussions, in addition to David Stokes and Steven Ludtke for advice on the phase-splitting phenomenon.

Funding information

AP is The Mark Foundation for Cancer Research Fellow of the Damon Runyon Cancer Research Foundation (DRG 2361-19). This work was supported by NIH grants R35 GM122588 (to GJJ), AI150464 (to GJJ) and GM117126 (to NKS).

References

Amos, L. A., Henderson, R. & Unwin, P. N. (1982). *Prog. Biophys. Mol. Biol.* **39**, 183–231.
Atakisi, H., Conger, L., Moreau, D. W. & Thorne, R. E. (2019). *IUCrJ*, **6**, 1040–1053.
Baker, L. A. & Rubinstein, J. L. (2010). *Methods Enzymol.* **481**, 371–388.
Beale, E. V., Warren, A. J., Trincão, J., Beilsten-Edmands, J., Crawshaw, A. D., Sutton, G., Stuart, D. & Evans, G. (2020). *IUCrJ*, **7**, 500–508.
Bharat, T. A. M. & Scheres, S. H. W. (2016). *Nat. Protoc.* **11**, 2054–2065.
Bharat, T. A. M., Kureisaite-Ciziene, D., Hardy, G. G., Yu, E. W., Devant, J. M., Hagen, W. J. H., Brun, Y. V., Briggs, J. A. G. & Löwe, J. (2017). *Nat. Microbiol.* **2**, 17059.
Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.
Brocchieri, L. & Karlin, S. (2005). *Nucleic Acids Res.* **33**, 3390–3400.
Castaño-Díez, D., Kudryashev, M., Arheit, M. & Stahlberg, H. (2012). *J. Struct. Biol.* **178**, 139–151.
Chen, M., Bell, J. M., Shi, X., Sun, S. Y., Wang, Z. & Ludtke, S. J. (2019). *Nat. Methods*, **16**, 1161–1168.
Cheng, Y. (2015). *Cell*, **161**, 450–457.
Chreifi, G., Chen, S. & Jensen, G. J. (2021). *J. Struct. Biol.* **213**, 107716.
Chreifi, G., Chen, S., Metskas, L. A., Kaplan, M. & Jensen, G. J. (2019). *J. Struct. Biol.* **205**, 163–169.
Clabbers, M. T. B. & Abrahams, J. P. (2018). *Crystallogr. Rev.* **24**, 176–204.
Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.
Cruz, M. J. de la, Hattne, J., Shi, D., Seidler, P., Rodriguez, J., Reyes, F. E., Sawaya, M. R., Cascio, D., Weiss, S. C., Kim, S. K., Hinck, C. S., Hinck, A. P., Calero, G., Eisenberg, D. & Gonen, T. (2017). *Nat. Methods*, **14**, 399–402.
Cruz, M. J. de la, Martynowycz, M. W., Hattne, J. & Gonen, T. (2019). *Ultramicroscopy*, **201**, 77–80.
Dadinova, L. A., Chesnokov, Y. M., Kamyshinsky, R. A., Orlov, I. A., Petoukhov, M. V., Mozhaev, A. A., Soshinskaya, E. Y., Lazarev, V. N., Manuvera, V. A., Orekhov, A. S., Vasiliev, A. L. & Shtykova, E. V. (2019). *FEBS Lett.* **593**, 1360–1371.
Dauter, Z. (1999). *Acta Cryst.* D**55**, 1703–1717.
Drulyte, I., Johnson, R. M., Hesketh, E. L., Hurdiss, D. L., Scarff, C. A., Porav, S. A., Ranson, N. A., Muench, S. P. & Thompson, R. F. (2018). *Acta Cryst.* D**74**, 560–571.
Ebrahim, A., Moreno-Chicano, T., Appleby, M. V., Chaplin, A. K., Beale, J. H., Sherrell, D. A., Duyvesteyn, H. M. E., Owada, S., Tono, K., Sugimoto, H., Strange, R. W., Worrall, J. A. R., Axford, D., Owen, R. L. & Hough, M. A. (2019). *IUCrJ*, **6**, 543–551.
Gallagher-Jones, M., Ophus, C., Bustillo, K. C., Boyer, D. R., Panova, O., Glynn, C., Zee, C. T., Ciston, J., Mancia, K. C., Minor, A. M. & Rodriguez, J. A. (2019). *Commun. Biol.* **2**, 26.
Genderen, E. van, Li, Y., Nederlof, I. & Abrahams, J. P. (2016). *Acta Cryst.* D**72**, 34–39.
Gevorkov, Y., Barty, A., Brehm, W., White, T. A., Tolstikova, A., Wiedorn, M. O., Meents, A., Grigat, R.-R., Chapman, H. N. & Yefanov, O. (2020). *Acta Cryst.* A**76**, 121–131.
Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.
Hagen, W. J. H., Wan, W. & Briggs, J. A. G. (2017). *J. Struct. Biol.* **197**, 191–198.
Hattne, J., Martynowycz, M. W., Penczek, P. A. & Gonen, T. (2019). *IUCrJ*, **6**, 921–926.
Hattne, J., Reyes, F. E., Nannenga, B. L., Shi, D., de la Cruz, M. J., Leslie, A. G. W. & Gonen, T. (2015). *Acta Cryst.* A**71**, 353–360.
Henderson, R. (1995). *Q. Rev. Biophys.* **28**, 171–193.
Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. (1990). *J. Mol. Biol.* **213**, 899–929.
Heumann, J. M., Hoenger, A. & Mastronarde, D. N. (2011). *J. Struct. Biol.* **175**, 288–299.
Himes, B. A. & Zhang, P. (2018). *Nat. Methods*, **15**, 955–961.
Hohn, M., Tang, G., Goodyear, G., Baldwin, P., Huang, Z., Penczek, P. A., Yang, C., Glaeser, R. M., Adams, P. D. & Ludtke, S. J. (2007). *J. Struct. Biol.* **157**, 47–55.
Holton, J. M. (2009). *J. Synchrotron Rad.* **16**, 133–142.
Holton, J. M. & Frankel, K. A. (2010). *Acta Cryst.* D**66**, 393–408.
Hovmöller, S. (1981). *Rotation Matrices and Translation Vectors in Crystallography*. Chester: International Union of Crystallography.
Hovmöller, S. (1992). *Ultramicroscopy*, **41**, 121–135.
Jackson, R. N., McCoy, A. J., Terwilliger, T. C., Read, R. J. & Wiedenheft, B. (2015). *Nat. Protoc.* **10**, 1275–1284.
Jensen, G. J. (2001). *J. Struct. Biol.* **133**, 143–155.
Jensen, G. J. & Kornberg, R. D. (2000). *Ultramicroscopy*, **84**, 57–64.
Juers, D. H., Farley, C. A., Saxby, C. P., Cotter, R. A., Cahn, J. K. B., Holton-Burke, R. C., Harrison, K. & Wu, Z. (2018). *Acta Cryst.* D**74**, 922–938.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 133–144.
Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. (1996). *J. Struct. Biol.* **116**, 71–76.
Lange, R. H. (1974). *J. Ultrastruct. Res.* **46**, 301–307.
Lin, J. (1991). *IEEE*, **37**, 145–151.
Liu, W., Wacker, D., Gati, C., Han, G. W., James, D., Wang, D., Nelson, G., Weierstall, U., Katritch, V., Barty, A., Zatsepin, N. A., Li, D.,

Messerschmidt, M., Boutet, S., Williams, G. J., Koglin, J. E., Seibert, M. M., Wang, C., Shah, S. T., Basu, S., Fromme, R., Kupitz, C., Rendek, K. N., Grotjohann, I., Fromme, P., Kirian, R. A., Beyerlein, K. R., White, T. A., Chapman, H. N., Caffrey, M., Spence, J. C. H., Stevens, R. C. & Cherezov, V. (2013). *Science*, **342**, 1521–1524.
Lučić, V., Rigort, A. & Baumeister, W. (2013). *J. Cell Biol.* **202**, 407–419.
Martynowycz, M. W., Glynn, C., Miao, J., de la Cruz, J., Hattne, J., Shi, D., Cascio, D., Rodriguez, J. & Gonen, T. (2017). *bioRxiv*, 152504.
Mastronarde, D. N. (1997). *J. Struct. Biol.* **120**, 343–352.
Mastronarde, D. N. (2005). *J. Struct. Biol.* **152**, 36–51.
Nannenga, B. L. & Gonen, T. (2019). *Nat. Methods*, **16**, 369–379.
Nannenga, B. L., Shi, D., Hattne, J., Reyes, F. E. & Gonen, T. (2014). *eLife*, **3**, e03600.
Nannenga, B. L., Shi, D., Leslie, A. G. W. & Gonen, T. (2014). *Nat. Methods*, **11**, 927–930.
Nederlof, I., Li, Y. W., van Heel, M. & Abrahams, J. P. (2013). *Acta Cryst.* D**69**, 852–859.
Nguyen, P. Q. T., Wang, S., Kumar, A., Yap, L. J., Luu, T. T., Lescar, J. & Tam, J. P. (2014). *FEBS J.* **281**, 4351–4366.
Nicastro, D., Schwartz, C., Pierson, J., Gaudette, R., Porter, M. E. & McIntosh, J. R. (2006). *Science*, **313**, 944–948.
Oikonomou, C. M. & Jensen, G. J. (2017). *Annu. Rev. Biochem.* **86**, 873–896.
Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.
Powell, H. R. (2017). *Biosci. Rep.* **37**, BSR20170227.
Radermacher, M. (2007). *Electron Tomography*, edited by J. Frank, pp. 245–273. New York: Springer.
Read, R. J. & Schierbeek, A. J. (1988). *J. Appl. Cryst.* **21**, 490–495.
Rodriguez, J. A., Eisenberg, D. S. & Gonen, T. (2017). *Curr. Opin. Struct. Biol.* **46**, 79–86.
Rodriguez, J. A., Ivanova, M. I., Sawaya, M. R., Cascio, D., Reyes, F. E., Shi, D., Sangwan, S., Guenther, E. L., Johnson, L. M., Zhang, M., Jiang, L., Arbing, M. A., Nannenga, B. L., Hattne, J., Whitelegge, J., Brewster, A. S., Messerschmidt, M., Boutet, S., Sauter, N. K., Gonen, T. & Eisenberg, D. S. (2015). *Nature*, **525**, 486–490.
Rupp, B. (2010). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology.* New York: Garland Science.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.
Schenk, A. D., Castaño-Díez, D., Gipson, B., Arheit, M., Zeng, X. & Stahlberg, H. (2010). *Methods Enzymol.* **482**, 101–129.
Schlichting, I. (2015). *IUCrJ*, **2**, 246–255.
Schur, F. K. M. (2019). *Curr. Opin. Struct. Biol.* **58**, 1–9.
Shi, D., Nannenga, B. L., Iadanza, M. G. & Gonen, T. (2013). *eLife*, **2**, e01345.
Smith, J. L., Fischetti, R. F. & Yamamoto, M. (2012). *Curr. Opin. Struct. Biol.* **22**, 602–612.
Stuart, D. I. & Abrescia, N. G. A. (2013). *Acta Cryst.* D**69**, 2257–2265.
Subramanian, G., Basu, S., Liu, H., Zuo, J.-M. & Spence, J. C. H. (2015). *Ultramicroscopy*, **148**, 87–93.
Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I. & Ludtke, S. J. (2007). *J. Struct. Biol.* **157**, 38–46.
Taylor, G. (2003). *Acta Cryst.* D**59**, 1881–1890.
Tegunov, D., Xue, L., Dienemann, C., Cramer, P. & Mahamid, J. (2021). *Nat. Methods*, **18**, 186–193.
Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Acta Cryst.* D**65**, 582–601.
Tivol, W. F., Briegel, A. & Jensen, G. J. (2008). *Microsc. Microanal.* **14**, 375–379.
Turoňová, B., Schur, F. K. M., Wan, W. & Briggs, J. A. G. (2017). *J. Struct. Biol.* **199**, 187–195.
Unwin, P. N. & Henderson, R. (1975). *J. Mol. Biol.* **94**, 425–440.
Wan, W. & Briggs, J. A. G. (2016). *Methods Enzymol.* **579**, 329–367.
Wan, W., Kolesnikova, L., Clarke, M., Koehler, A., Noda, T., Becker, S. & Briggs, J. A. G. (2017). *Nature*, **551**, 394–397.
Wang, J., Dauter, M. & Dauter, Z. (2006). *Acta Cryst.* D**62**, 1475–1483.
Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A. S., Sauter, N. K. & Evans, G. (2013). *CCP4 Newsl. Protein Crystallogr.* **49**, 16–19.
Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst.* D**74**, 85–97.
Wolf, S. G., Frenkiel, D., Arad, T., Finkel, S. E., Kolter, R. & Minsky, A. (1999). *Nature*, **400**, 83–85.
Wu, M. & Lander, G. C. (2020). *Curr. Opin. Struct. Biol.* **64**, 9–16.
Xiong, Q., Morphew, M. K., Schwartz, C. L., Hoenger, A. H. & Mastronarde, D. N. (2009). *J. Struct. Biol.* **168**, 378–387.
Young, I. D., Ibrahim, M., Chatterjee, R., Gul, S., Fuller, F., Koroidov, S., Brewster, A. S., Tran, R., Alonso-Mori, R., Kroll, T., Michels-Clark, T., Laksmono, H., Sierra, R. G., Stan, C. A., Hussein, R., Zhang, M., Douthit, L., Kubin, M., de Lichtenberg, C., Pham, L. V., Nilsson, H., Cheah, M. H., Shevela, D., Saracini, C., Bean, M. A., Seuffert, I., Sokaras, D., Weng, T.-C., Pastor, E., Weninger, C., Fransson, T., Lassalle, L., Bräuer, P., Aller, P., Docker, P. T., Andi, B., Orville, A. M., Glownia, J. M., Nelson, S., Sikorski, M., Zhu, D., Hunter, M. S., Lane, T. J., Aquila, A., Koglin, J. E., Robinson, J., Liang, M., Boutet, S., Lyubimov, A. Y., Uervirojnangkoorn, M., Moriarty, N. W., Liebschner, D., Afonine, P. V., Waterman, D. G., Evans, G., Wernet, P., Dobbek, H., Weis, W. I., Brunger, A. T., Zwart, P. H., Adams, P. D., Zouni, A., Messinger, J., Bergmann, U., Sauter, N. K., Kern, J., Yachandra, V. K. & Yano, J. (2016). *Nature*, **540**, 453–457.
Zheng, S. Q., Palovcak, E., Armache, J.-P., Verba, K. A., Cheng, Y. & Agard, D. A. (2017). *Nat. Methods*, **14**, 331–332.
Zhu, L., Weierstall, U., Cherezov, V. & Liu, W. (2016). *Adv. Exp. Med. Biol.* **922**, 151–160.