Acta Crystallographica Section D
**Biological Crystallography**
ISSN 1399-0047

# Breaking the indexing ambiguity in serial crystallography

**Wolfgang Brehm and Kay Diederichs***

Department of Biology, Universität Konstanz, Box 647, 78457 Konstanz, Germany

Correspondence e-mail: kay.diederichs@uni-konstanz.de

Received 7 August 2013
Accepted 13 September 2013

In serial crystallography, a very incomplete partial data set is obtained from each diffraction experiment (a 'snapshot'). In some space groups, an indexing ambiguity exists which requires that the indexing mode of each snapshot needs to be established with respect to a reference data set. In the absence of such re-indexing information, crystallographers have thus far resorted to a straight merging of all snapshots, yielding a perfectly twinned data set of higher symmetry which is poorly suited for structure solution and refinement. Here, two algorithms have been designed for assembling complete data sets by clustering those snapshots that are indexed in the same way, and they have been tested using 15 445 snapshots from photosystem I [Chapman *et al.* (2011), *Nature (London)*, **470**, 73–77] and with noisy model data. The results of the clustering are unambiguous and enabled the construction of complete data sets in the correct space group $P6_3$ instead of (twinned) $P6_322$ that researchers have been forced to use previously in such cases of indexing ambiguity. The algorithms thus extend the applicability and reach of serial crystallography.

## 1. Introduction

In 38 of the 65 space groups in which proteins consisting of L-amino acids crystallize, data sets can be merged without re-indexing. However, if the symmetry of the Bravais lattice is higher than the symmetry of the space group, an indexing ambiguity results. This is the case in 27 space groups, where a decision between two (in 24 space groups) or four (in $P3$, $P3_1$ and $P3_2$) possible indexing modes has to be made for each data set that is to be merged or compared with other data sets. A survey of the PDB (Berman *et al.*, 2000) reveals that one out of every six crystal structures is obtained in these 'merohedral' space groups (Table 1).

The indexing-mode decision can easily be made in conventional crystallography, since data sets usually have a large number of common reflections: the possible modes can be tried in turn and the one that gives the best agreement between the data sets is chosen as the correct one.

However, the decision is difficult in the emerging method of serial crystallography (Kirian *et al.*, 2011; White *et al.*, 2012, 2013). For example, when the snapshots from microcrystals of photosystem I (PSI; Chapman *et al.*, 2011) were obtained, it was recognized that a single snapshot cannot serve as a reference for the indexing of the other snapshots because the number of reflections common to two snapshots is small and the experimental error is too large to reliably make a decision. As a consequence, the 15 445 snapshots were merged without

**Table 1**
As of 26 July 2013, there are 29 418 non-racemic X-ray structures in the nonredundant (<90% sequence identity) PDB.

5002 (17%) of these crystallize in a merohedral space group with an indexing ambiguity.

| Point group | Space group | Frequency in PDB |
|---|---|---|
| 3 | $P3$, $P3_1$, $P3_2$ | 47, 204, 199 |
| 6 | $P6_x$, $x = 0 \ldots 5$ | 42, 365, 75, 197, 85, 336 |
| 312 | $P312$, $P3_112$, $P3_212$ | 5, 35, 57 |
| 321 | $P321$, $P3_121$, $P3_221$ | 124, 935, 826 |
| 4 | $P4_x$, $x = 0 \ldots 3$ | 34, 165, 26, 217 |
| 4 | $I4$, $I4_1$ | 159, 107 |
| 3 | $R3$ ($H3$) | 406 |
| 23 | $P23$, $F23$, $I23$, $P2_13$, $I2_13$ | 24, 24, 113, 135, 60 |

establishing the indexing mode. Since, statistically, experimental snapshots represent the possible modes with equal frequency, a perfectly twinned data set was obtained.

Solving a structure from perfectly twinned data is not straightforward: firstly, experimental phasing needs a very elaborate treatment (Yeates & Rees, 1987) and very accurate data; secondly, perfect twinning reduces the number of unique data by a factor of (usually) two, thus lowering the data-to-parameter ratio in refinement; thirdly, it leads to strong model bias in maps; and fourthly, it invalidates the usual model $R$-value statistics (see, for example, Evans & Murshudov, 2013). For the PSI data, an additional problem is that the data are low resolution (8.7 Å), which prevents structure solution. In principle, structure solution from twinned data would allow a reference data set based on model intensities to be obtained; in practice, this is very difficult and is unlikely to succeed for data obtained by serial crystallography: working with (almost) perfectly twinned data is already difficult in conventional crystallography (see, for example, Jameson *et al.*, 2002; Ramadan *et al.*, 2002; Heffron *et al.*, 2006; Xu *et al.*, 2003).

The indexing-ambiguity conundrum thus poses a chicken-and-egg problem: only if a reference is available can the indexing mode of a snapshot be established, but the construction of the reference requires the correct indexing mode of each snapshot.

To the knowledge of the authors, attempts by various groups to solve the alternate indexing problem did not yield a solution or could not be applied because the algorithm requires rather ideal data (Zhou *et al.*, 2013). Bootstrapping procedures, in which the reference data set is gradually built up from snapshots that seem to agree in their indexing modes, apparently suffer from the high level of random and systematic error in the snapshots. This makes it difficult to unambiguously choose the next snapshot to merge with the existing partial reference. Thus far, bootstrap methods have therefore not resulted in suitable reference data sets.

As serial crystallography is still emerging, there is a tendency to avoid projects with indexing ambiguity. However, as is the case for photosystem I, there are undoubtedly many biologically relevant projects that would benefit from progress in breaking it. Since the indexing ambiguity may seemingly only be broken by using a reference data set obtained by conventional crystallography, this limits the applicability of this promising new technique.

In this paper, we demonstrate that the indexing ambiguity can be broken without bootstrapping by taking all of the available information into account in a consistent and conceptually simple fashion. Our algorithm has modest computing requirements and has successfully been applied to 15 445 snapshots from $P6_3$ crystals of photosystem I.

## 2. Materials and methods

Since the relative scale of the intensities of each snapshot is unknown, we use Pearson's correlation coefficient

$$r_{ij} = \frac{\sum_h \left[I_i(h) - \overline{I}_i\right]\left[I_j(h) - \overline{I}_j\right]}{\left\{\sum_h \left[I_i(h) - \overline{I}_i\right]^2 \sum_h \left[I_j(h) - \overline{I}_j\right]^2\right\}^{1/2}}, \quad (1)$$

as a scale-invariant indicator of the similarity of two snapshots $i$ and $j$ with intensities $I_i$ and $I_j$ of reflections $h$.

As a preparation for clustering of snapshots, we determine $r_{ij}$ for all pairs of snapshots $i$ and $j$ ($j > i$) with $l_{ij} > 2$ common reflections. This step has a computational complexity of order $n \times \langle m \rangle \times \langle l \rangle$, where $n$ is the number of snapshots, $m_i$ is the number of snapshots that snapshot $i$ has common reflections with and $\langle l \rangle$ is the average $l_{ij}$.

In the following, we use the words 'vector' and 'point' interchangeably. In particular, we show the end points of vectors $\mathbf{x}_i$ referring to a common origin as dots in figures, while in formulae we use vectors.

### 2.1. Algorithm 1

In the first algorithm, we consider a vector space in which each shot $i$ is represented as a vector $\mathbf{x}_i$. $k$, the dimension of this vector space, has to be chosen according to the target function (see below). This $k$-dimensional space, in which distances are defined in the usual Euclidean way, has no crystallographic meaning, but is needed to embed and then cluster the snapshots. To this end, we set negative $r_{ij}$ to 0 and minimize

$$\Psi = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[(1 - r_{ij}) - |\mathbf{x}_i - \mathbf{x}_j|\right]^2 \quad (2)$$

as a function of the coordinates of each pair of snapshots $\mathbf{x}_i$ and $\mathbf{x}_j$ using the L-BFGS minimization algorithm (Liu & Nocedal, 1989) with analytical derivatives. This seeks to minimize the difference between $1 - r_{ij}$ and the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in a $k$-dimensional space. The underlying assumption is that $\mathbf{x}_i$ and $\mathbf{x}_j$ should be close if their 'Pearson distance' $1 - r_{ij}$ is small ($r_{ij} \lesssim 1$) and distant if $r_{ij} \simeq 0$. Evidently, short distances exist between snapshots that belong to the same cluster. The minimization is overdetermined since there are $n \times (n - 1)$ $r_{ij}$ values but only $k \times n$ vector components.

The starting coordinates $\mathbf{x}$ are chosen randomly in the interval $(0 \ldots 1)$ for each of the $k$ dimensions. The

minimization is terminated when the norm of the gradient vanishes or after 200 iterations, whichever comes first. The computational complexity of each iteration is $n \times k \times \langle m \rangle$.

The result of the minimization is inspected graphically; in the important case of $k = 2$ we expect two clouds of points corresponding to the two possible indexing modes. For point group 3 (space groups $P3$, $P3_1$ and $P3_2$) we choose $k = 3$ and expect four clouds of points to be centred at the edges of a regular tetrahedron; this geometry has the property that the distances between any two of the four edges are the same, which is impossible for $k = 2$.

Since distances are invariant against rotation, translation and inversion of all points, the resulting arrangement of points in space is not strictly unique. Rather, the centres of the two (four) clouds have to be identified. Knowing these, the line (six planes if $k = 3$) dividing the indexing modes can be established. The division is midway between the centres of gravity of the clouds, perpendicular to their connection(s), because the number of points in each of the two (four) clouds is expected to be the same.

After separating the clouds by a line ($k = 2$) or by six planes ($k = 3$), each cloud represents one indexing mode and its snapshots may be merged to obtain a data set. Obviously, the other clouds may be treated in the same way and the resulting data sets can be re-indexed accordingly and merged.

### 2.2. Algorithm 2

As in algorithm 1, we consider a $k$-dimensional space, but instead minimize the function

$$\Phi = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (r_{ij} - \mathbf{x}_i \cdot \mathbf{x}_j)^2 \qquad (3)$$

using the L-BFGS minimization algorithm. Our choice of the function is motivated by the formal similarity of the definition of the correlation coefficient to that of a scalar product between two normalized zero-centred vectors. Approximating a correlation coefficient with a scalar product of $\mathbf{x}_i$ and $\mathbf{x}_j$ is thus going to optimize the lengths of $\mathbf{x}_i$ and $\mathbf{x}_j$ and the angle between them. It is mainly the angle that we are interested in; if $i$ and $j$ belong to two different indexing modes, we expect $r_{ij}$ to be close to 0 and an angle between $\mathbf{x}_i$ and $\mathbf{x}_j$ of around 90°. For the 24 merohedral chiral space groups, this can be accomplished in two-dimensional space with two clouds of points, one on or near the $x$ axis and the other on or near the $y$ axis. For space groups $P3$, $P3_1$ and $P3_2$ we choose $k = 4$ and expect four clouds of points, one on or near each of the four axes. $k = 3$ is not fully suitable because the angle between the edges of a regular tetrahedron is 120° instead of the expected value of 90°.

In each case, we start the minimization of the target function from random coordinates in the range $(0 \ldots 1)$. The optimization is terminated after 50 iterations or upon convergence. As for algorithm 1, the resulting pattern of points is not unique as the angle is invariant against rotation and inversion, but the line separating the clouds in the $k = 2$ case is easy to find since it is close to the diagonal of the first quadrant and divides the cloud of points evenly into two equal-sized halves. The deviation from the diagonal is small because the minimization has a tendency to keep the centre of gravity of the coordinates in the first quadrant.

### 2.3. Experimental and model data

To test our algorithms, we used the same data set as Chapman *et al.* (2011). Each of the 15 445 individual snapshots has between 98 and 247 reflections (with an average of 157) indexed in one of the two possible modes ($h, k, l$ and $\overline{h} - k, k, \overline{l}$) of $P6_3$. Altogether, there are 2 425 556 observations, resulting in 6151 unique reflections and an average multiplicity of 394 in $P6_3$ and an average multiplicity of 705 in $P6_322$.

For our assignment of indexing mode, those reflections that are invariant to the re-indexing between the two modes were not used in the calculation of $r_{ij}$ because they artificially increase $r_{ij}$ if $i$ and $j$ belong to different modes. Furthermore, if two or more reflections of a snapshot referred to the same unique reflection, they were averaged before the calculation of correlation coefficients.

Model data in space group $P6_3$ were generated using the same set of 15 445 snapshots and their indices. The intensities were calculated from a model (PDB entry 1jb0; Jordan *et al.*, 2001) using *phenix.f_model* (Adams *et al.*, 2010), taking bulk solvent into account. For every second snapshot, the intensity was taken from a re-indexed (using the $\overline{h} - k, k, \overline{l}$ transformation) model data set.

Noise was added to the model intensities $I_i$ in the following way. Firstly, a Gaussian variate with zero mean and a sigma corresponding to twice the average intensity of the data set was added to the model intensities. Secondly, the resulting value was multiplied by a random number evenly distributed in $(0 \ldots 1)$ to simulate the effect that the unknown partiality of reflections has, yielding the noisy intensities $N_i$. This further increases the error, resulting in an $R_{\rm scale}$ of 47%, where

$$R_{\rm scale} = \frac{\sum |I_i - N_i|}{\frac{1}{2}\sum I_i + N_i}, \qquad (5)$$

of the noisy against the unperturbed model data, a correlation of 0.70 and a $\langle \text{signal} \rangle / \langle \text{error} \rangle$ of only 0.54. This amount of error was chosen, after some experimentation, such that the plots of the experimental and model quantities (see below) appeared to be approximately similar.

To obtain model data for point group 3, we changed PDB entry 1jb0 to space group $P3$ but kept the unit-cell parameters, and used the re-indexing transformations $(\overline{h} - k, k, \overline{l})$, $(\overline{h}, \overline{k}, l)$ and $(h + k, \overline{k}, \overline{l})$ to obtain model data with a fourfold indexing ambiguity. Each possible indexing mode was applied in turn to establish a relation between the snapshot number and the indexing mode which could later be used to evaluate the results. Noise was added as described above.

### 2.4. Evaluation of algorithms

It should be noted that it is only necessary to *break* the indexing ambiguity, *i.e.* to *bias* the resulting assignment towards the correct one, rather than to completely resolve it

without error, because it is possible to construct a reference data set from the non-ideally twinned data resulting from the inclusion of mis-assigned shots. Therefore, even if application of our methods does not completely resolve the problem in one step, it may be completely resolved in an iterative way using a bootstrap procedure with one of our algorithms producing the seed.

The results of our algorithms can be visualized graphically for both the experimental and the model data (see §3). Numerically, there is no unambiguous way to verify the correctness of the assignment for the experimental data. The closest surrogate of the true data are the model intensities of PDB entry 1jb0. These are a rather poor reference because they correspond to a model refined against a different experimental data set which is not perfectly isomorphous with the FEL data set. Use of these data as a reference can only demonstrate that the indexing ambiguity is broken by our methods; any quantitative assessment is inaccurate.

On the other hand, for our model data we can count the number of mis-assigned shots after application of one of our algorithms, because the indexing mode of each shot is known. Using $w$ for the number of wrong assignments and $c$ for the number of correct assignments, we can calculate

$$\alpha_{\rm est} = \frac{w}{w + c} \tag{5}$$

as an estimate of the twinning fraction of the merged data. It is likely that those shots that are particularly weak or noisy will more often be mis-assigned than strong, less noisy shots. Since weak or noisy mis-assigned shots should be expected to have less influence on the resulting twinning fraction than an average shot, $\alpha_{\rm est}$ is a conservative estimate.

The border(s) between the two (four) clusters can be more easily delineated in the results of algorithm 2, mainly because the possible arrangements resulting from algorithm 2 have lower symmetry than those from algorithm 1, but also because algorithm 2 produces clusters which are more densely populated in the centre. Thus, we calculated $\alpha_{\rm est}$ only for the results of algorithm 2 acting on the noisy model data.

For counting wrong and correct assignments, the border was defined, for $k = 2$, by a line passing through the origin and separating the two clusters such that these have equal counts of shots.

For $k = 4$, we used an iterative procedure to find the centres of the shot clusters. The initial centre positions corresponded to the coordinates of four randomly chosen shots. We then iterated the following two steps 50 times: (i) shots were assigned to the closest centre so that each centre received one

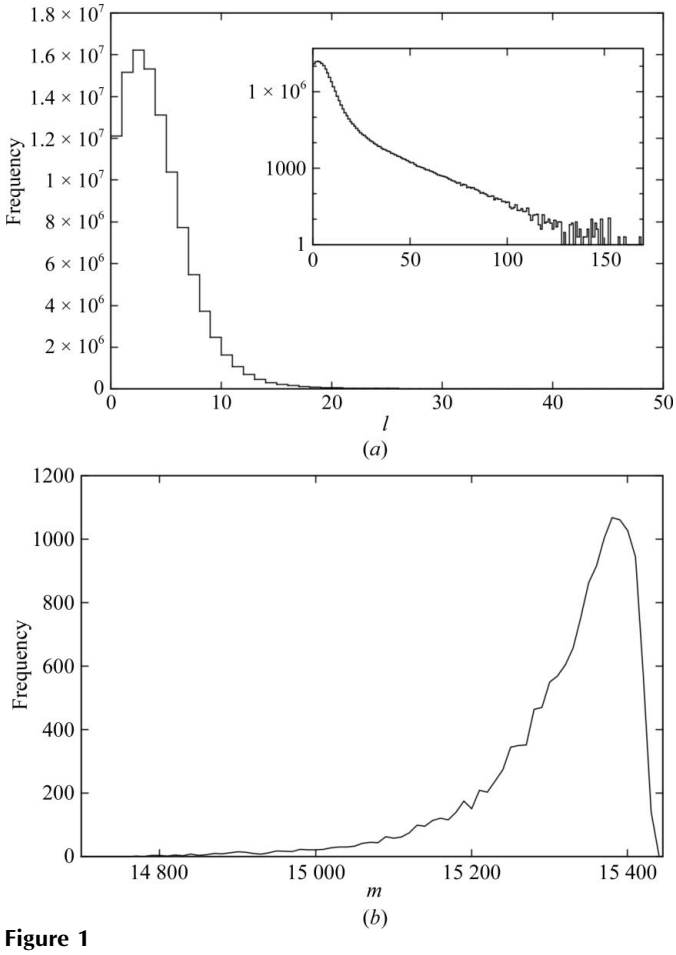


**Figure 1**
The number of snapshots pairs with given $l$ ($a$) and $m$ ($b$) are shown for the experimental and model data. The inset in ($a$) is a half-logarithmic plot.

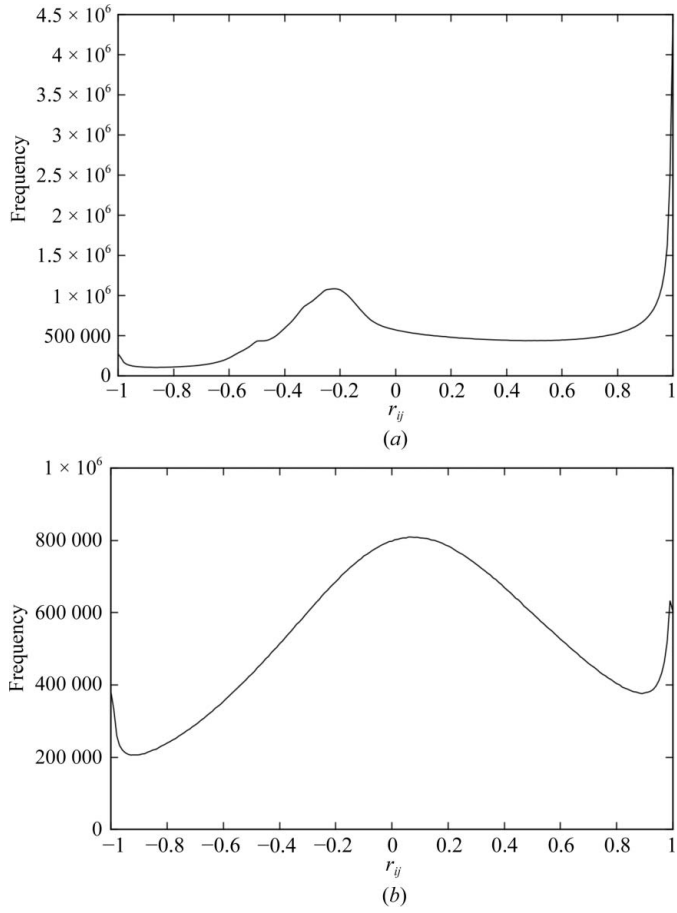


**Figure 2**
Histograms of $r_{ij}$ for the experimental data ($a$) and the noisy model data ($b$) in $P6_3$ are shown.

quarter of all shots and (ii) the coordinates of the shots assigned to each centre were averaged to give the new position of the centre.

## 3. Results

In the case of the data from photosystem I, we find $n = 15\,445$, $\langle m \rangle = 15\,319$ and $\langle l \rangle = 6.3$, which means that any snapshot has about six common reflections with any other snapshot. The triangular array of unique $r_{ij}$ values can be calculated in about 30 s of CPU time with a single processor (3 GHz). The L-BFGS iterations take roughly the same time, which demonstrates that even without parallelization the computational problem is easily tractable.

Fig. 1 shows the number of snapshots pairs with given $l$ (Fig. 1*a*) and $m$ (Fig. 1*b*). By construction, the histograms for the model data are identical.

Fig. 2 shows a histogram of $r_{ij}$ for the experimental data (Fig. 2*a*) and the noisy model data (Fig. 2*b*) in $P6_3$. These

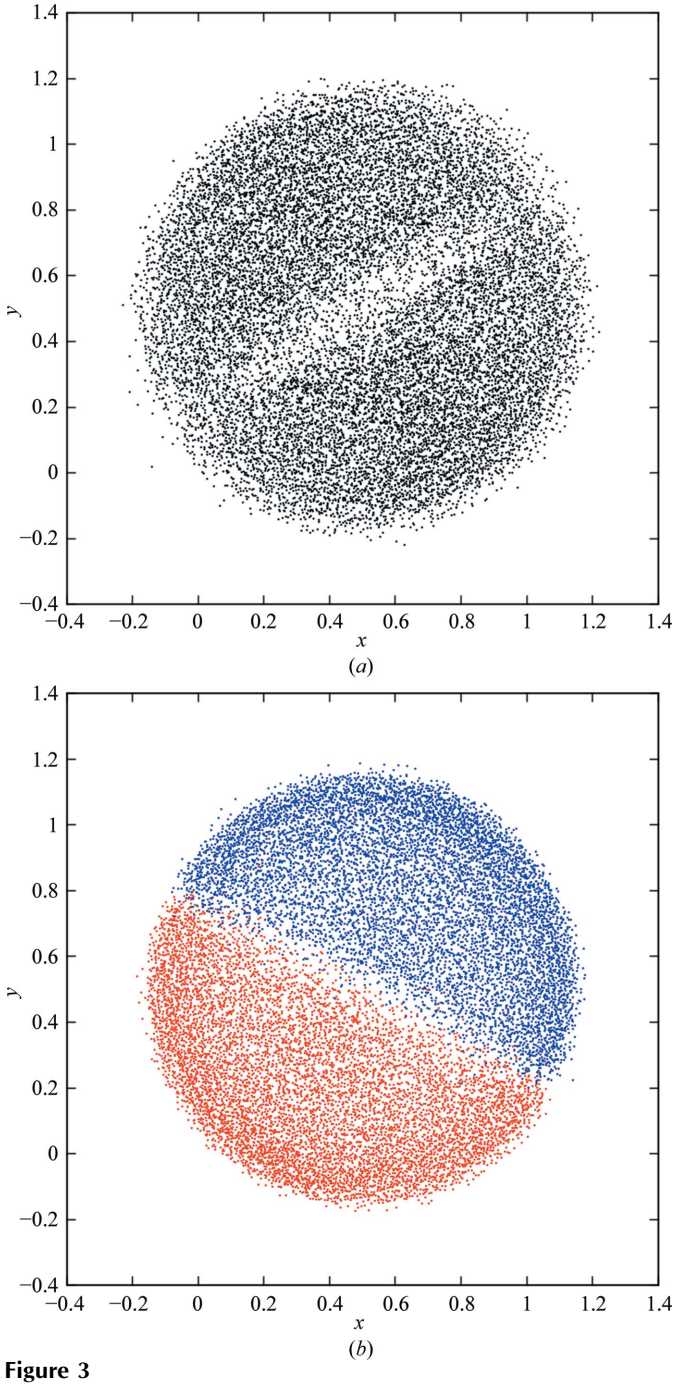


**Figure 3**
The application of algorithm 1 to the experimental data (*a*) and the noisy model data (*b*) is shown. The points in (*b*) are coloured according to the known indexing mode.

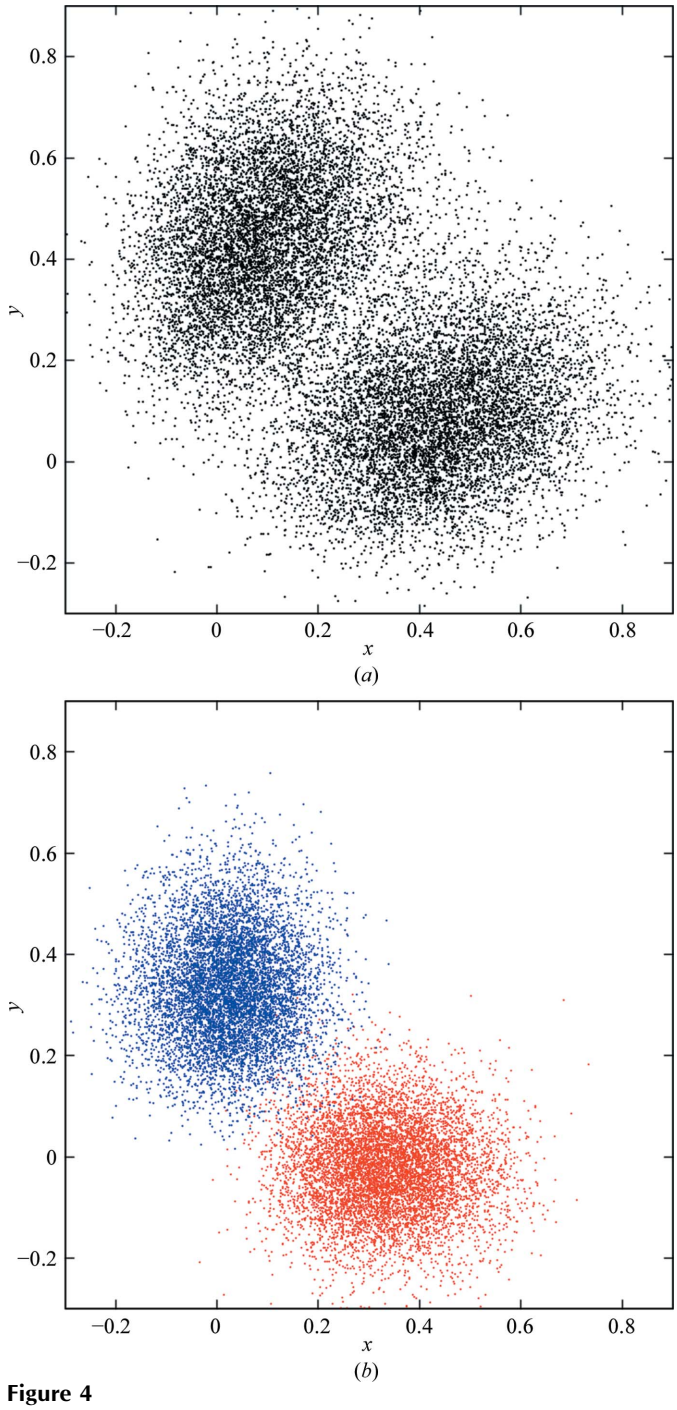


**Figure 4**
Application of algorithm 2 to the experimental data (*a*) and the noisy model data (*b*), respectively. The points in (*b*) are coloured according to the known indexing mode.

histograms differ substantially since there is a large number of high-correlation pairs of experimental snapshots, whereas for the model data high-correlation pairs are rare in comparison. We do not know the reason for this discrepancy, but it signifies that the pattern of error in the intensities of the noisy model data does not match that of the experimental data.

Fig. 3 shows the results from algorithm 1 applied to the experimental data (Fig. 3*a*) and the noisy model (Fig. 3*b*) data, respectively. The iterations were stopped after 200 cycles when no further improvement of the target function was observed; however, the gradient had not yet converged. The shapes of the clouds reminded us of cell-division processes, with a clear separation resulting from depletion of points at the line of division, but a roughly constant density of points within each half-sphere. The angle of the line of division differs between Fig. 3(*a*) and Fig. 3(*b*) since it depends on the random numbers used to assign starting coordinates for each snapshot.

Fig. 4 shows the application of algorithm 2 to the experimental data (Fig. 4*a*) and the noisy model data (Fig. 4*b*), respectively. 25 iterations were needed to obtain convergence to a minimum value of the target function where the gradient vanishes. In contrast to algorithm 1, algorithm 2 produces a higher density of points in the centre of the cloud.

Figs. 3 and 4, the most important results of our study, clearly show that not only the noisy model data but also the experimental data may be separated into the two possible indexing modes. For the model data and for both algorithms, less than 1% of the snapshots are placed in the wrong cloud, equivalent to an estimated twinning fraction $\alpha_{\rm est}$ of less than 1%. For the experimental data, the correctness of the indexing assignment can be checked by correlation coefficients against a reference data set consisting of intensities from PDB model 1jb0; this reveals a clear transition between the two indexing modes for the two clouds of points (Fig. 5). The assignment by algorithm 2 to indexing modes agrees for 81.4% of the experimental shots with the assignment obtained by using the model intensities of 1jb0 as a reference.

For both algorithms, we ran optimizations from different starting points and compared the results. In all cases we ended up with similar arrangements of points and closely similar values of the target function. This lends evidence, but not

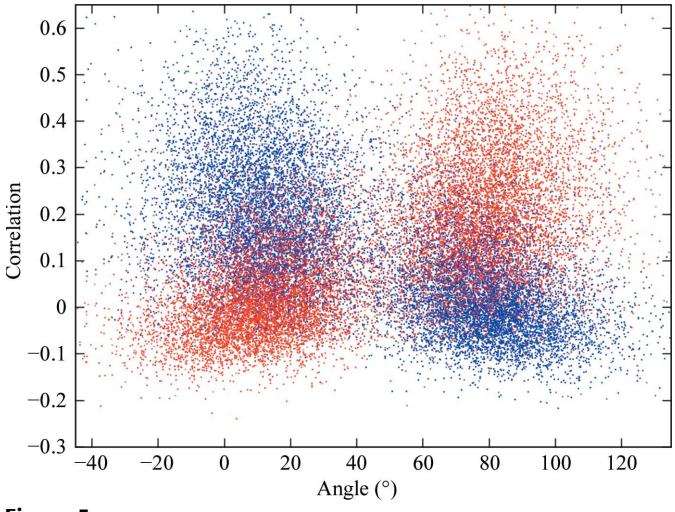


**Figure 5**
Checking the correctness of the indexing-mode assignment for the experimental data of Fig. 4(*a*) by comparison with the 1jb0 reference data set. The location of an experimental snapshot on the abscissa is given as the angle of its position in Fig. 4. The correlation coefficient of each snapshot against the 1jb0 reference data (red points) and against the re-indexed (using $\overline{h}-k,k,\overline{l}$) 1jb0 reference data (blue points) is plotted on the ordinate. There is a sharp transition between the indexing modes near 45°, the diagonal of the first quadrant.

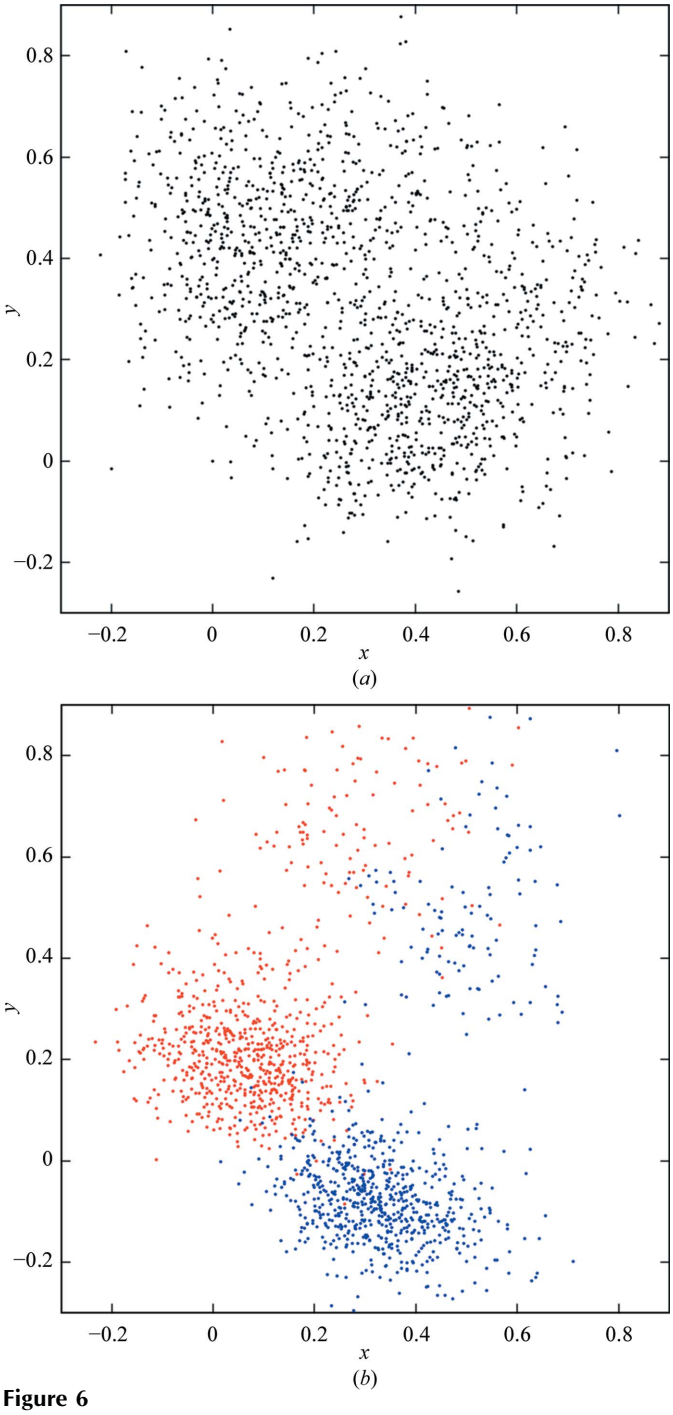


**Figure 6**
Application of algorithm 2 to a random selection (10%) of the experimental (*a*) and the noisy model (*b*) snapshots, respectively. The points in (*b*) are coloured according to the known indexing mode.

proof, to the hypothesis that after convergence the results are close to the global minima of the target functions. Since just an imperfect separation of points is needed to break the indexing ambiguity, this result is strong enough for our purposes.

Encouraged by these results, we reduced the number of snapshots to investigate whether a lower degree of over-determination can still produce a meaningful result. Fig. 6 shows the application of algorithm 2 to 1544 (10%) of the experimental (Fig. 6*a*) and the model (Fig. 6*b*) snapshots, respectively. The visual separation of the two indexing modes is not significantly worse than that seen in Fig. 4, but for the model data in Fig. 6(*b*) $\alpha_{\rm est}$ is significantly worse (10.7%), a value that may require twinning to be taken into account during refinement of a model.

Fig. 7 shows the application of algorithm 1 ($k$ = 3; Fig. 7*a*) and algorithm 2 ($k$ = 4; Fig. 7*b*) to the noisy model snapshots in *P*3. Again, the clouds are very distinct and their arrangement is indeed in the shape of a tetrahedron (Fig. 7*a*) or orthogonal (Fig. 7*b*), as was anticipated. We also tried algorithm 2 with $k$ = 3 and obtained a tetrahedron (not shown), but the target

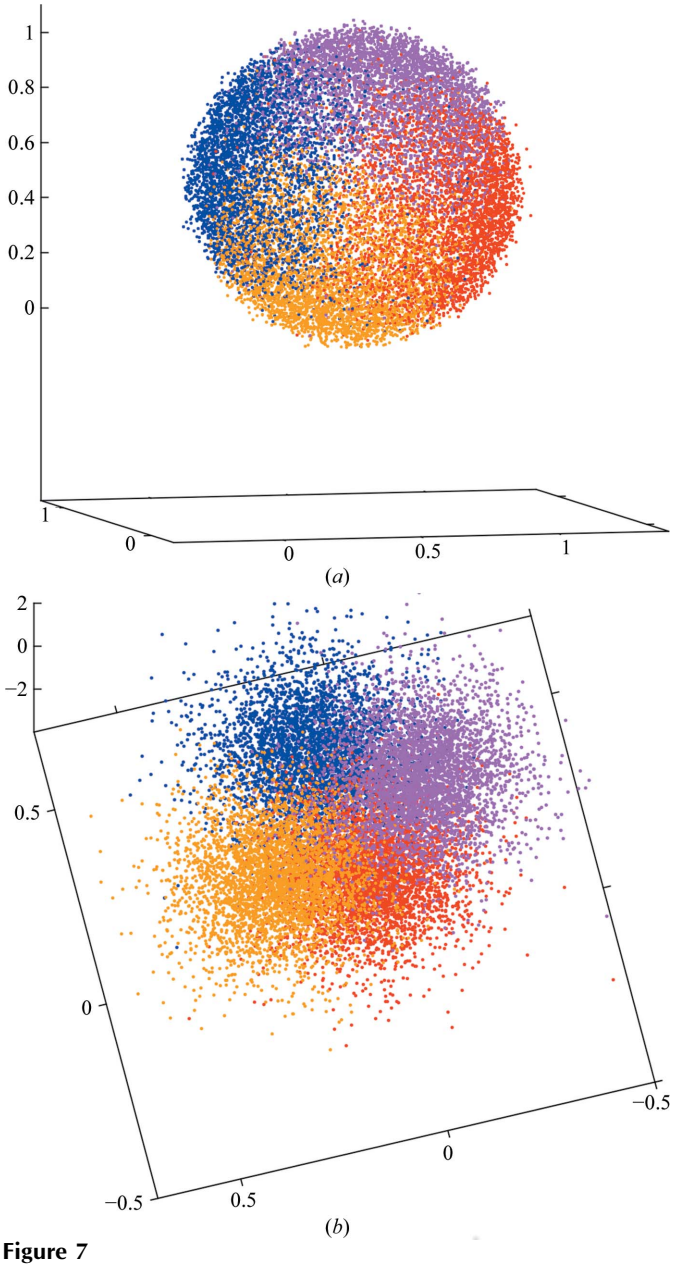


**Figure 7**
Application of algorithms 1 (*a*) and 2 (*b*) to the noisy model snapshots (space group *P*3) in three-dimensional space (*a*) and four-dimensional space (*b*), represented as two-dimensional projections. The points are coloured according to the known indexing mode.

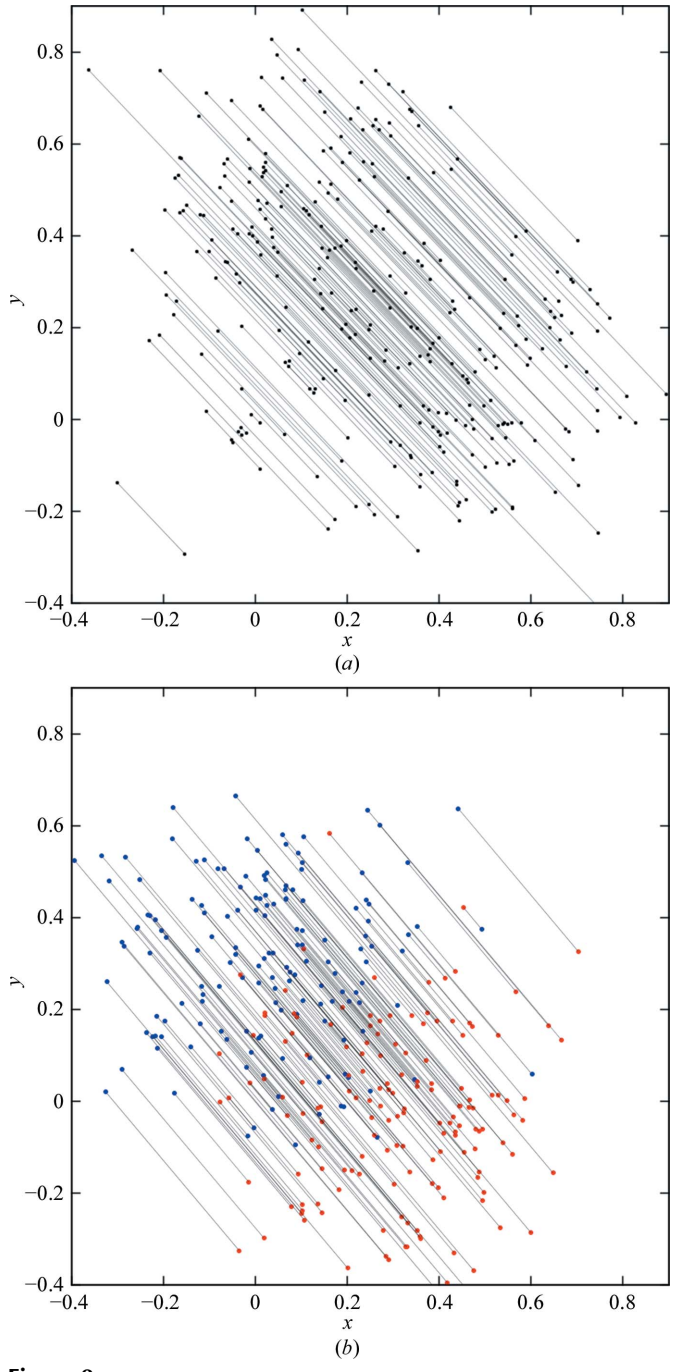


**Figure 8**
Application of algorithm 2 to a random selection (1%) of the experimental (*a*) and the noisy model (*b*) snapshots, respectively. This selection of 154 snapshots was augmented by the same snapshots after re-indexing (see text). Equivalent snapshots are connected by lines. The points in (*b*) are coloured according to the known indexing mode.

function value after convergence was higher than with $k = 4$, as we expected. Numerical evaluation of algorithm 2 (Fig. 7*b*) gave a value of 5.7% for $\alpha_{\rm est}$. This is probably low enough to ignore twinning in refinement, but is worse than the value obtained for $k = 2$. Nonetheless, it demonstrates that the fourfold indexing ambiguity in point group 3 can be broken.

We tried both algorithms with $k = 1$, but we did not observe a well separated bimodal distribution in the point density histogram (not shown).

## 4. Discussion

The successful separation of indexing modes achieved with both algorithms and in $k = 2$, 3 and 4 dimensions attests to the fact that our hypothesis, namely that a close *versus* distant relationship can be calculated from imprecise pairwise correlation coefficients between snapshots, is justified. On the other hand, the simplest approach with $k = 1$ did not work well. We believe that the target functions, despite their simplicity, may have local extrema and that the optimization of the target function works better in a higher than one-dimensional space, since local target-function maxima may be iteratively circumvented in higher dimensions, whereas they are insurmountable in one dimension.

The success for $k > 1$ is probably owing to the fact that only a single bit of information (or two bits in the case of point group 3) must be extracted for every snapshot, which is easily accomplished since this task is highly overdetermined with, on average, $\langle m \rangle = 15\,319$ pairwise correlation coefficients.

One obvious possibility to use even more information is to include more snapshots. In addition to this, we realised that any given snapshot can actually be used twice (or four times in $P3$, $P3_1$ and $P3_2$) by re-indexing it and also including it in the calculation of the $r_{ij}$. This not only increases the available number of snapshots, but some systematic errors may cancel if the difference vectors between the pairs (or quadruples) of points belonging to the same snapshot are directly compared.

We tried this approach with a randomly selected 154 snapshots (1% of the total) from both the experimental and noisy model data and connected the originally indexed and the re-indexed snapshots with each other (Fig. 8) after optimization with algorithm 2. Indeed, the pairs of snapshots are connected by parallel lines, and the endpoints of the lines can be used to define the centres of gravity corresponding to the two indexing modes. Furthermore, a separation of clouds is clearly visible for the experimental data, and the model data show a separation of the originally indexed and re-indexed snapshots, although some are found in the wrong cloud. Numerically, for the model data with 154 duplicated snapshots $\alpha_{\rm est}$ is somewhat worse (14.8%) than when using 1545 snapshots without duplication.

This demonstrates that even with a low number of snapshots it is possible to assign the indexing modes, as long as the individual snapshots have enough reflections in common. When the re-indexing duplication is performed for the data of Fig. 6, the separation of clouds becomes much better, and for the noisy model data all snapshots are correctly indexed (not shown). Thus, re-indexing duplication of snapshots definitely improves the robustness of the method and the clarity of its results.

We did not investigate the mathematical properties of our target functions, nor did we try to find analytical solutions instead of the iterative optimization we employed, because we wanted to keep our method as general as possible, allowing easy implementation of even better-suited target functions. One possibility for such an improvement is to include the standard error of $r_{ij}$ into the target function that is being optimized. This would give more weight to those $r_{ij}$ that are accurately determined and is straightforward to implement. Another possibility is to filter the list of $r_{ij}$ used for minimization of the target function by considering only those $r_{ij}$ that are based on a minimum $l_{ij}$. Along similar lines, the experimental $r_{ij}$ could be transformed by a linear ($r'_{ij} = ar_{ij} + b$) or higher-order function to make the histograms of experimental and model $r_{ij}$ more similar. Also, different relations involving $r_{ij}$, $\mathbf{x}_i$ and $\mathbf{x}_j$ could be investigated.

We have not tried these possible improvements because we obtained good results with algorithm 1 and, in particular, with algorithm 2. For future projects, we also anticipate significant improvements in the raw data arising from better integration software as well as from better detector hardware and from improved FEL beam quality. Any such improvement in the raw data will result in easier resolution of the indexing ambiguity.

Evidently, our method can also be applied to data from crystals in pseudo-merohedral settings, such as those of *Tobacco mosaic virus* (Bloomer *et al.*, 1978), which are orthorhombic but appear pseudo-tetragonal because the *a* and *b* axes have similar lengths. Lebedev *et al.* (2006) estimate that combinations of crystal and lattice symmetries allow merohedral and pseudo-merohedral twinning in more than 30% of PDB entries, almost doubling our estimate in Table 1.

A well known property of merohedral space groups (those that show an indexing ambiguity) is that twinning by merohedry can, and often does, occur. A search for 'twinned' in a subset of the PDB in which sequence homologues with >90% identical residues were removed identified 184 entries. Inspection reveals that many of these are close to perfectly twinned, as is the case for the first four entries in the list: *Escherichia coli* elongation factor Tu (Heffron *et al.*, 2006), phospholipase $A_2$ from the venom of *Ophiophagus hannah* (Xu *et al.*, 2003), a domain-opened mutant (R121D) of the human lactoferrin N-lobe (Jameson *et al.*, 2002) and calcium-depleted human C-reactive protein (Ramadan *et al.*, 2002).

Crystals of sizes larger than a few micrometres are composed of many mosaic blocks. If such a crystal is perfectly twinned its data cannot be detwinned. Nanocrystals, however, consist of only one mosaic block and cannot be twinned. Thus, nanocrystals of such proteins could be measured on an FEL or microbeam synchrotron beamline (Kirian *et al.*, 2011), allowing the solution and refinement of the structure from untwinned data after application of our method. Twinned microcrystals consisting of several mosaic blocks may be analyzed with our methods, and those with the lowest twinning

fraction may be selected for merging. Thus, the technique of serial crystallography could be used to experimentally detwin data from those crystal forms which otherwise only give close to perfectly twinned macroscopic crystals.

In summary, a problem in serial crystallography that has hampered its application to a significant fraction of projects can be easily solved with our method. As a byproduct, the method may allow the long-standing twinning problem in conventional crystallography to be overcome, a possibility that attests to the complementary properties of serial and conventional crystallography.

KD is grateful to J. Holton for information about the problem, encouragement to tackle it and fruitful discussion, and we thank H. Chapman, T. White, P. Fromme and J. Spence for making the PSI data available to us.

## References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. (1978). *Nature (London)*, **276**, 362–368.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.
Heffron, S. E., Moeller, R. & Jurnak, F. (2006). *Acta Cryst.* D**62**, 433–438.
Jameson, G. B., Anderson, B. F., Breyer, W. A., Day, C. L., Tweedie, J. W. & Baker, E. N. (2002). *Acta Cryst.* D**58**, 955–962.
Jordan, P., Fromme, P., Witt, H. T., Klukas, O., Saenger, W. & Krauss, N. (2001). *Nature (London)*, **411**, 909–917.
Kirian, R. A., White, T. A., Holton, J. M., Chapman, H. N., Fromme, P., Barty, A., Lomb, L., Aquila, A., Maia, F. R. N. C., Martin, A. V., Fromme, R., Wang, X., Hunter, M. S., Schmidt, K. E. & Spence, J. C. H. (2011). *Acta Cryst.* A**67**, 131–140.
Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Cryst.* D**62**, 83–95.
Liu, D. C. & Nocedal, J. (1989). *Math. Program. B*, **45**, 503–528.
Ramadan, M. A. M., Shrive, A. K., Holden, D., Myles, D. A. A., Volanakis, J. E., DeLucas, L. & Greenhough, T. J. (2002). *Acta Cryst.* D**58**, 992–1001.
White, T. A., Barty, A., Stellato, F., Holton, J. M., Kirian, R. A., Zatsepin, N. A. & Chapman, H. N. (2013). *Acta Cryst.* D**69**, 1231–1240.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
Xu, S., Gu, L., Wang, Q., Shu, Y., Song, S. & Lin, Z. (2003). *Acta Cryst.* D**59**, 1574–1581.
Yeates, T. O. & Rees, D. C. (1987). *Acta Cryst.* A**43**, 30–36.
Zhou, L., Liu, P. & Dong, Y.-H. (2013). *Chin. Phys. C.* **37**, 028101.