

Recovery of data from perfectly twinned virus crystals revisited.

HELEN MARY GINN^a AND DAVID IAN STUART^{ab*}

^a*Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, Oxfordshire, OX3 7BN England, and* ^b*Diamond House, Harwell Science and Innovation Campus, Fermi Avenue, Didcot, Oxfordshire, OX11 0QX England. E-mail: dave@strubi.ox.ac.uk*

Abstract

Perfect merohedral twinning of crystals is not uncommon and complicates structural analysis. We have reimplemented an iterative method for the deconvolution of data from perfectly merohedrally twinned crystals in the presence of non-crystallographic symmetry. We show that the method recovers the data effectively using test data and now provide an independent metric of success, based on special classes of reflections that are unaffected by the twin operator. We apply the method to a real problem with five-fold NCS symmetry and rather poor quality diffraction data, and find that even in these circumstances the method appears to recover most of the information. We make the software available in a form that can be applied to other crystal systems.

1. Introduction

Biological crystals can occasionally, but not uncommonly, be subject to perfect or imperfect merohedral twinning (Yeates, 1997; Yeates & Fam, 1999), where unit cells

or mosaic domains are randomly distributed into two or more orientations without affecting the crystal lattice. This is particularly common in virus capsid crystallography where spherical capsids can rotate without significantly altering the minimal crystal contacts (Lerch *et al.*, 2009). For some crystal systems, twinning can be minimised or avoided by altering the concentration of nuclei for crystallisation (Chayen & Saridakis, 2008) or deliberately choosing crystals that grow at a slower rate (Borshchevskiy *et al.*, 2009). When the merohedral twinning fraction is measurably below 0.5, data recovery is comparatively easier, and quite frequently allows structure solution by *de novo* methods. For molecular replacement solutions there are a large number of examples (Breyer *et al.*, 1999; Igarashi *et al.*, 1997; Carr *et al.*, 1996; Luecke *et al.*, 1998; Chandra *et al.*, 1999; Contreras-Martel *et al.*, 2001). For anomalous phasing, notable examples include interleukin-1 (Rudolph *et al.*, 2003) and a SeMet variant of the capsid-stabilizing protein of bacteriophage λ , gpD, (Yang *et al.*, 2000), which were both solved by multiwavelength anomalous dispersion (MAD). Twinned crystals of bilirubin oxidase with a twin fraction of 0.487 were solved by SAD (Mizutani *et al.*, 2010). However, perfect merohedral twinning is often more challenging to overcome, and most commonly requires molecular replacement to solve the structure (Chandra *et al.*, 1999; Redinbo & Yeates, 1993; Lea & Stuart, 1995). However, solving gpD has been achieved by SAD, where the data were averaged to emulate a twinning fraction of 0.5 (Dauter, 2003). Twinning presents itself as a higher symmetry space group and may be more difficult to immediately detect if analysis of the crystal packing density is not unambiguous. However, it causes an enrichment of mid-intensity reflections due to the superposition of the two crystal orientations, where combinations of two low intensity or two high intensity reflections are less common. In fact, it is common for proteins to be submitted to the PDB with their partially twinned nature going unnoticed (Lebedev *et al.*, 2006). Programs such as Truncate, part of

the CCP4 suite, now test for this distorted intensity distribution as standard (Winn *et al.*, 2011).

Foot-and-mouth disease virus (FMDV) crystals of the O1M variant form perfectly merohedrally twinned crystals similar to that of the G67 variant caused by a 90° difference in orientation of 50% of virions in the crystal. The previously solved structures O1BFS (PDB code 1BBT) and O1K lack point mutations in residues 72-74 that were proposed to give rise to twinning and therefore form untwinned crystals in space group I23 (Acharya *et al.*, 1989; Lea *et al.*, 1995). In I23, ignoring anomalous differences, such perfect twinning makes reflections (h, k, l) and (k, h, l) equivalent, creating pseudo-four-fold symmetry that emulates the symmetry of the I432 space group. In this case, this can be distinguished from a true I432 space group as icosahedral viruses do not possess four-fold symmetry and the unit cell dimensions only permit a single virion in the unit cell. Reflections where $h = k$ are unaffected by twinning (hereby referred to as ‘singlet’ reflections). Note that depending on the definition of the asymmetric unit, this can also include $h = l$ and $k = l$. Twinning in the G67 variant has been shown to occur at the level of mosaic blocks as the paired structure factors correlate most strongly with the mean intensity of untwinned O1BFS structure factor twin pairs rather than the vector mean (Lea & Stuart, 1995). Importantly, the icosahedral virus capsid pentamers cannot be part of the crystallographic symmetry, and are therefore present in the non-crystallographic symmetry operations, which is key to this study.

We aim to recover a set of untwinned structure factors from these perfectly twinned data, using a method that has previously described to deconvolute similar data sets (Lea & Stuart, 1995). This is an unusual procedure, as it is said conventionally that untwinned intensities cannot be recovered from perfectly twinned data sets, unlike those that have a twinning fraction of less than 50%. The procedure is designed to obtain a set of untwinned structure factors that are consistent with the F_{obs} measure-

ments, while producing an electron density map that obeys the known five-fold NCS. In other words, after recovery of the untwinned intensities, averaging the intensities of each twin pair of reflections would be equal to the original twinned intensity. In order to generate the untwinned intensities, the intensities must be biased towards their true values. If a data set had no non-crystallographic symmetry, it would not be possible to bias the intensities enough to recover the untwinned structure factors. However, with five-fold non-crystallographic symmetry, which breaks the symmetry produced by the 90° rotation twinning operation, it is possible to bias the original intensities towards their untwinned values, and recover the untwinned intensities over several iterative cycles of refinement. Five-fold averaging across one axis causes constructive interference of signal for one orientation of the virion, whereas the 90° -related virions do not possess this symmetry and average out to noise. After this, one must ensure that paired reflection intensities respect the twinning law: this is done by rescaling individual pairs of reflections such that the average of the corresponding intensities matches that of the original twinned intensities. This is followed by additional cycles of NCS averaging and application of the twinning law until the procedure converges.

We have made the source code available for others to use, and a summary of the method (iterative cycles of NCS averaging, application of the twinning law and rescaling of the structure factors) is provided in Fig. 1. As a control, a set of structure factors were generated from FMDV O1BFS coordinates. These intensities came from naturally untwinned crystals that were artificially 'retwinned' by averaging the (h, k, l) and (k, h, l) intensities. This study reimplements the method and seeks to validate the procedure using these 'retwinned' O1BFS structure factors as a control and assess the quality of recovery from twinned O1M data in a more rigorous fashion than previously attempted. The experimental details of crystal preparation and the derived structure are reported in another paper (Kotecha *et al.*, 2015).

2. Materials and Methods

2.1. Artificial twinning of O1BFS reflections

Untwinned O1BFS structure factors were obtained from the PDB (entry 1BBT). To ‘retwin’ the data, intensities were averaged between the twin reflection pairs. To reduce the quality of the O1BFS phases to be similar to the quality of the O1M phases (derived as described below), rigid body refinement, positional minimisation and B factor refinement was performed using retwinned O1BFS amplitudes and the atomic coordinates of O1BFS in CNS version 1.3 (Brunger, 2007).

2.2. Generation of preliminary phases

The intensities for the O1M dataset were scaled and merged in space group I432 and expanded to space group I23. Preliminary phases for O1M were generated in CNS by rigid body refinement using atomic coordinates of O1BFS and the twinned amplitudes from the O1M data. The model was further refined by minimisation and B factor refinement.

2.3. NCS averaging

A solvent flattening envelope was generated for electron density maps by setting the interior and exterior of the protein capsid to a density of 0 using the General Averaging Program (GAP) (Grimes *et al.*, 1998). Electron density maps were averaged using the envelope and symmetry operators representing the five-fold non-crystallographic symmetry present in these data. The calculated data were transformed back to reciprocal space for scaling.

2.4. Resolution shell scaling

Reflections were categorised into twenty resolution shells, each containing a similar number of data. All calculated amplitudes were scaled to observed amplitudes using a

scale factor F_{obs} / F_{calc} generated using singlet reflections only within each resolution shell, as these are not affected by twinning. The number of such reflections was between 89 and 360, so the scale factors are likely to be statistically reliable.

2.5. Twinning law scaling

A scale factor k was generated and applied to each related pair of reflections in order to generate calculated amplitudes that are consistent with the observed amplitudes in the twinned data set according to Equation 1, while keeping the ratio between the pair of amplitudes the same. Except for the final cycle iteration, singlet data were adjusted to $(2F_{obs} - F_{calc})$ before scaling rather than setting them equal to their known values. On the last round of refinement, singlet reflections were set to the original amplitudes from the twinned data set. Structure factors were transformed to real space if sequential rounds of NCS averaging and scaling were required.

$$k = \frac{\sqrt{2}F_{hkl}^{obs}}{\sqrt{F_{hkl}^{23}{}^2 + F_{khl}^{23}{}^2}} \quad (1)$$

3. Results

Reflections for O1M and artificially twinned O1BFS were transformed into real space. These electron density maps were averaged using five-fold NCS and scaled according to resolution shell using only singlet reflections for a total of 20 cycles. R factors and correlation coefficients were measured between observed twinned data and partially detwinned data, for both the whole set of reflections (R_{all} , CC_{all}) and the singlet subset ($R_{singlets}$, $CC_{singlets}$), on each stage of cycle (R factors shown in Figure 2, including the result from incorrect NCS operators). The singlet reflections are treated specially, rather than setting them equal to the amplitudes in the twinned data set: they are only scaled globally. This allows them to be used as a measure of success by

tracking their agreement with the original amplitudes over several rounds of five-fold NCS averaging, as they are unaffected by twinning.

The O1BFS data set is of high quality with a standard error (σ_{obs}/F_{obs}) of 4.2%, reflecting the excellent diffraction from these crystals. R_{all} for the O1BFS control shows sequential divergence between the twinned and deconvoluted data sets, reaching a maximum of 28.3% and a correlation coefficient (CC_{all}) of 0.591. $R_{singlets}$ improves from 15.9% to 5.8% showing excellent prediction of singlet values by the deconvoluted dataset. This is corroborated by the maximum $CC_{singlets}$ value of 0.978. The R factor comparing all of the original untwinned O1BFS amplitudes and the deconvoluted amplitudes shows strong agreement, 9.3%. Algorithms used to re-assign negative reflection intensities during data processing of the diffraction patterns (French & Wilson, 1978) tend to skew the weakest original amplitudes towards slightly higher calculated values, which is corrected post-deconvolution. This suggests the original amplitudes can be largely recovered to the limitations of the standard error of the untwinned amplitudes.

The phases generated for the twinned O1M dataset were of poor quality, and resulted in a poor preliminary R factor of 38.6%, as shown in Table 1. R_{all} for O1M closely follows that of the O1BFS data, reaching a maximum of 28.8% with a CC_{all} of 0.715. The $R_{singlets}$ shows calculated singlet reflections more closely match the observed data at the final converged value of 21.2% and a $CC_{singlets}$ value of 0.941 before the final cycle. The major source of error for the higher $R_{singlets}$ and lower $CC_{singlets}$ values compared to the O1BFS data is likely to be the poorer crystal quality and diffraction; the high standard error (σ_{obs}/F_{obs}) for the O1M data set is 15.4% for all reflections. Other sources of error include the reassignment of negative intensities and the using the O1BFS coordinates to generate phases, which will be of poorer quality. However, the drop in $R_{singlets}$ to a final value that is within 6% of a one stan-

dard deviation discrepancy suggests that the near-maximal recovery of the detwinned amplitudes has been achieved compared to the control, despite the poorer quality of the dataset.

The improvement in density is seen immediately after deconvolution without any need for extensive structure refinement. After deconvolution the structure can be refined and generates good quality electron density maps in Phenix (an illustrative example is given in Figure 3). These refined coordinates can be refined against the twinned data set as well and the density compared. It is apparent from the shape of the F_{obs} to F_{calc} distribution from Phenix (Adams *et al.*, 2010) that the twinned data have a distorted distribution of F_{obs} values, with an enrichment of mid-intensity reflections that match a wide range of F_{calc} values. This is reflected in the CC_{work} increasing from 76.8% (twinned) to 81.5% (detwinned). The real-space correlation coefficient between individual residues increase from 87.7% against the twinned data to 89.4% against the detwinned data across each of the five NCS copies of 660 residues and is clearly elevated throughout the sequence of the protein chains (Figure 4).

4. Conclusion

The data analysis suggests that deconvolution of twinned crystals with rotational non-crystallographic symmetry, which is distinct from the symmetry of the twinning operators, is successful. The control data set used here also suggests the error can be reduced to within 6% of the error already present during data collection. The success of the deconvolution process can be measured by separately processing and tracking the R factor for singlet reflections only, and is verified visually by comparing the electron density. Furthermore this method will be highly applicable to other virus crystal structures that possess high rotational non-crystallographic symmetry and a high propensity for twinning due to their pseudo-spherical nature, as well as other twinned

197 structures that exhibit similar non-crystallographic symmetry and twinning operator
198 relationships. This could be applied to the six point groups that support true mero-
199 hedral twinning (Yeates, 1997). Tables of space groups that can lead to this problem,
200 point groups and possible twin operators have been discussed (Chandra *et al.*, 1999).
201 The source code for solving hemihedral twinning, written primarily in C++, is avail-
202 able along with an example structure and script (github.com/helenginn/deconvolute).
203 It requires the CCP4 tools to be installed, but provides the other external Fortran
204 tools required to run the program. Compilation has been tested on the gcc compiler
205 version 4.4.7.

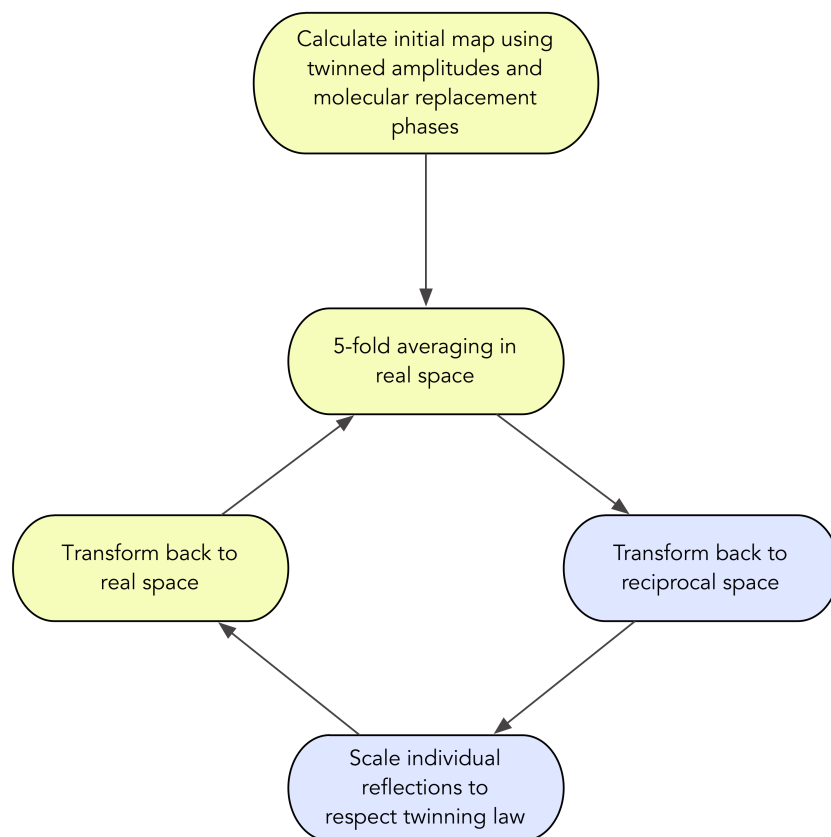
206 Note added in proof: Following submission of this paper, a study also dealing with
207 the use of non-crystallographic symmetry to aid in the handling of perfectly twinned
208 diffraction data was published by Sabin & Plevka (2016).

Acknowledgements

We thank Dr Claudine Porta who provided the O1M particles, and Drs Abhay Kotecha, Claudine, Ren Jingshan and Elizabeth Fry for providing the X-ray data for the O1M strain. We thank Wolfgang Kabsch for input into the code for the General Averaging Program. D.I.S. is supported by the Medical Research Council (grant G1000099) and H.G. is supported by a Wellcome Trust studentship (grant ALR00040). Admin support was received from the Wellcome Trust, grant 090532/Z/09/Z. This is a contribution from the Oxford Instruct Centre.

217

218



219

Fig. 1. Strategy for deconvolution of twinned data sets; yellow boxes within the cycle are carried out in real space and blue boxes are carried out in reciprocal space. Cycle is typically executed 20 times at which point convergence has been achieved.

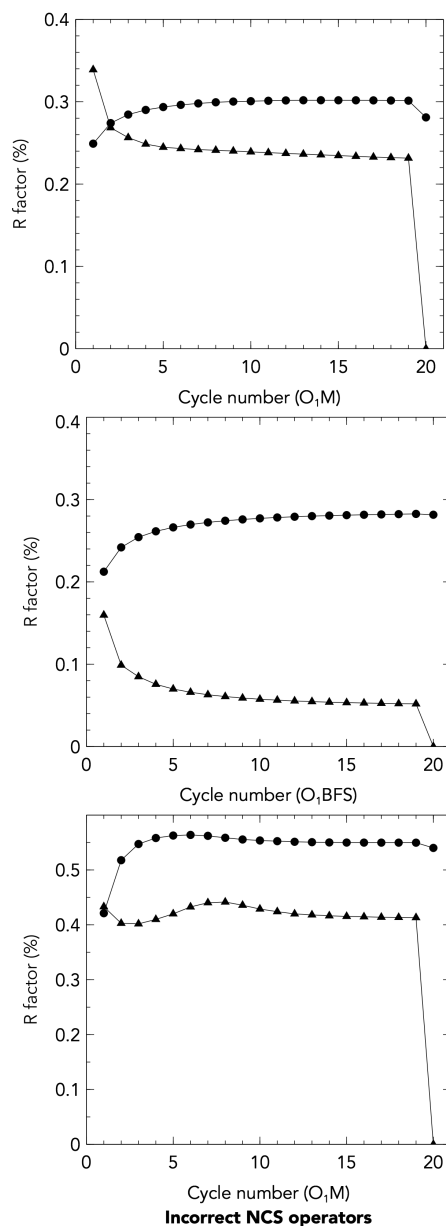


Fig. 2. $R_{singlets}$ (crosses) and R_{all} (noughts) values for deconvolution of the twinned O1M data set (top) and the artificially twinned O1BFS data set (middle). R_{all} values diverge while $R_{singlets}$ values converge; singlet reflections are not affected by twinning operators. If NCS operators are rotated by 90° in the x axis and deconvolution is attempted (bottom), the R factors do not show signs of success, as expected.

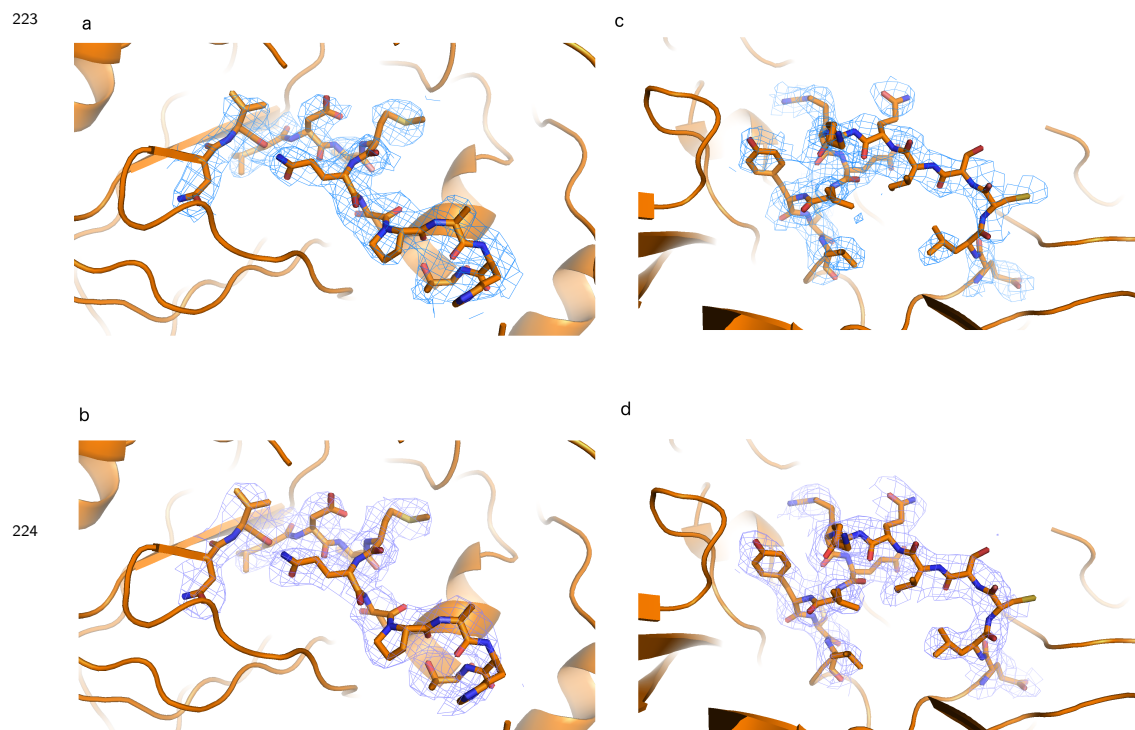


Fig. 3. Density around Met54 of VP3 in O1Manisa variant, showing breakage of main chain density in twinned structure (a), and recovery of main chain density in the detwinned structure (b). Similarly, twinned density (c) and detwinned density (d) is shown around residue Gln133 of VP1. Phases derived from refinement with Phenix in both maps. Density drawn at a sigma of 1.0.

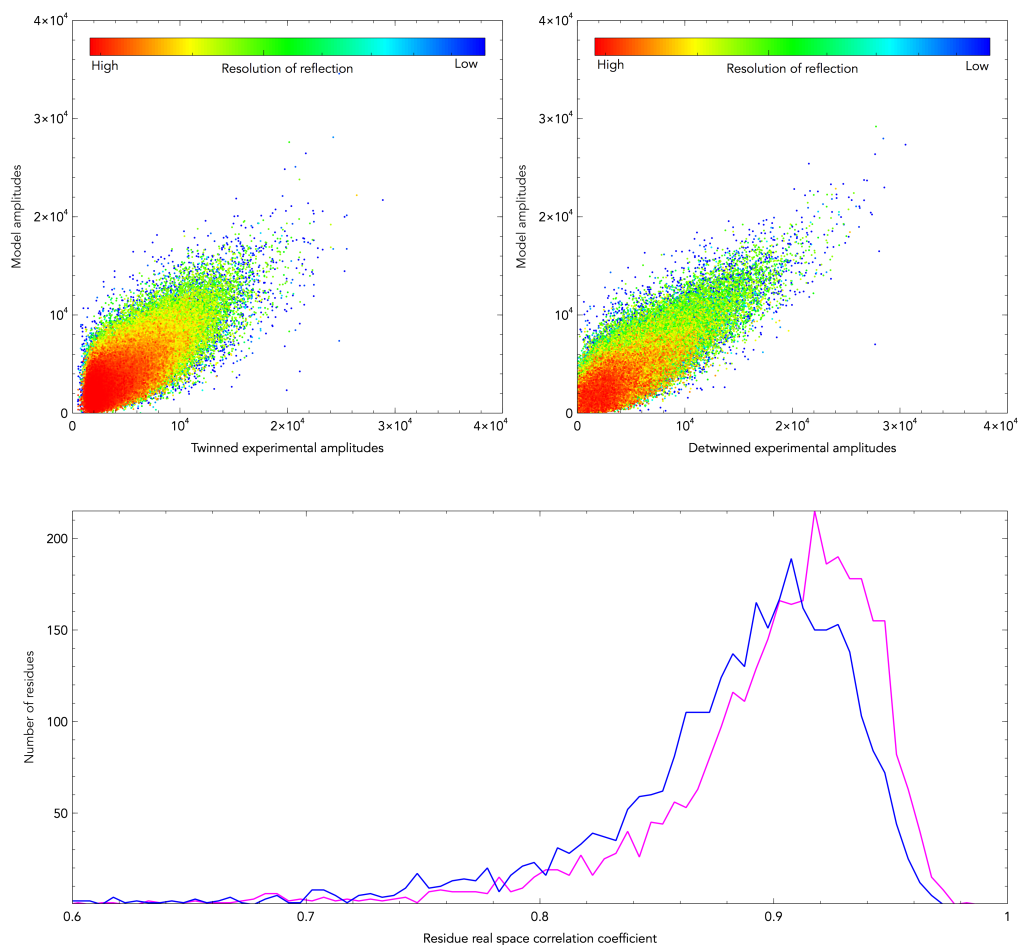


Fig. 4. Amplitude plots between F_{obs} and F_{calc} between the twinned data sets (top left) and detwinned data sets (top right). Frequency of real-space correlation coefficients per residue (below), where the blue line is derived from the map from refinement of the model against the twinned amplitude (mean correlation is 0.87), and the magenta line is derived similarly from the detwinned amplitudes (mean correlation is 0.89).

Table 1. *Preliminary statistics for O1M reflection data set prior to data recovery algorithm.*

Resolution range (°)	43.63 - 2.90
Space group	I23
Cell dimensions	
a = b = c	344.08
Number of unique reflections	69,889
Multiplicity	2.1
R_{merge} (%)	30.9
Completeness (outer shell) (%)	92.3 (77.4)
R_{work} pre-deconvolution (%)	38.6
R_{work} post-deconvolution (%)	37.5
R_{work} post model refinement (%)	33.9

229

References

230

- 231 Acharya, R., Fry, E., Stuart, D., Fox, G., Rowlands, D. & Brown, F. (1989). *Nature*, **337**,
 232 709–716.
- 233 Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J.,
 234 Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W. *et al.* (2010). *Acta Crystallographica*
 235 *Section D: Biological Crystallography*, **66**(2), 213–221.
- 236 Borshchevskiy, V., Efremov, R., Moiseeva, E., Büldt, G. & Gordeliy, V. (2009). *Acta Crystal-*
 237 *lographica Section D: Biological Crystallography*, **66**(1), 26–32.
- 238 Breyer, W. A., Kingston, R. L., Anderson, B. F. & Baker, E. N. (1999). *Acta Crystallographica*
 239 *Section D: Biological Crystallography*, **55**(1), 129–138.
- 240 Brunger, A. T. (2007). *Nature protocols*, **2**(11), 2728–2733.
- 241 Carr, P., Cheah, E., Suffolk, P., Vasudevan, S., Dixon, N. & Ollis, D. (1996). *Acta Crystallo-*
 242 *graphica Section D: Biological Crystallography*, **52**(1), 93–104.
- 243 Chandra, N., Acharya, K. R. & Moody, P. (1999). *Acta Crystallographica Section D: Biological*
 244 *Crystallography*, **55**(10), 1750–1758.
- 245 Chayen, N. E. & Saridakis, E. (2008). *Nature methods*, **5**(2), 147–153.
- 246 Contreras-Martel, C., Martinez-Oyanedel, J., Bunster, M., Legrand, P., Piras, C., Vernede, X.
 247 & Fontecilla-Camps, J.-C. (2001). *Acta Crystallographica Section D: Biological Crystal-*
 248 *lography*, **57**(1), 52–60.
- 249 Dauter, Z. (2003). *Acta Crystallographica Section D: Biological Crystallography*, **59**(11), 2004–
 250 2016.
- 251 French, S. & Wilson, K. (1978). *Acta Crystallographica Section A: Crystal Physics, Diffraction,*
 252 *Theoretical and General Crystallography*, **34**(4), 517–525.
- 253 Grimes, J. M., Burroughs, J. N., Gouet, P., Diprose, J. M., Malby, R., Zientara, S., Mertens,
 254 P. P. & Stuart, D. I. (1998). *Nature*, **395**(6701), 470–478.
- 255 Igarashi, N., Moriyama, H., Fujiwara, T., Fukumori, Y. & Tanaka, N. (1997). *Nature Structural*
 256 *& Molecular Biology*, **4**(4), 276–284.
- 257 Kotecha, A., Seago, J., Scott, K., Burman, A., Loureiro, S., Ren, J., Porta, C., Ginn, H. M.,
 258 Jackson, T., Perez-Martin, E. *et al.* (2015). *Nature structural & molecular biology*, **22**(10),
 259 788–794.
- 260 Lea, S., Abu-Ghazaleh, R., Blakemore, W., Curry, S., Fry, E., Jackson, T., King, A., Logan,
 261 D., Newman, J. & Stuart, D. (1995). *Structure*, **3**(6), 571–580.
- 262 Lea, S. & Stuart, D. (1995). *Acta Crystallographica Section D: Biological Crystallography*,
 263 **51**(2), 160–167.
- 264 Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Crystallographica Section D:*
 265 *Biological Crystallography*, **62**(1), 83–95.
- 266 Lerch, T. F., Xie, Q., Ongley, H. M., Hare, J. & Chapman, M. S. (2009). *Acta Crystallographica*
 267 *Section F: Structural Biology and Crystallization Communications*, **65**(2), 177–183.
- 268 Luecke, H., Richter, H.-T. & Lanyi, J. K. (1998). *Science*, **280**(5371), 1934–1937.
- 269 Mizutani, K., Toyoda, M., Sagara, K., Takahashi, N., Sato, A., Kamitaka, Y., Tsujimura, S.,
 270 Nakanishi, Y., Sugiura, T., Yamaguchi, S. *et al.* (2010). *Acta Crystallographica Section*
 271 *F: Structural Biology and Crystallization Communications*, **66**(7), 765–770.
- 272 Redinbo, M. R. & Yeates, T. O. (1993). *Acta Crystallographica Section D: Biological Crystal-*
 273 *lography*, **49**(4), 375–380.
- 274 Rudolph, M. G., Kelker, M. S., Schneider, T. R., Yeates, T. O., Oseroff, V., Heidary, D. K.,
 275 Jennings, P. A. & Wilson, I. A. (2003). *Acta Crystallographica Section D: Biological*
 276 *Crystallography*, **59**(2), 290–298.
- 277 Sabin, C. & Plevka, P. (2016). *Acta Crystallographica Section F: Structural Biology Commu-*
 278 *nications*, **72**(3).

- 279 Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan,
 280 R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N.,
 281 Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S.
 282 (2011). *Acta Crystallographica Section D: Biological Crystallography*, **67**(4), 235–242.
 283 **URL:** <http://dx.doi.org/10.1107/S0907444910045749>
- 284 Yang, F., Dauter, Z. & Wlodawer, A. (2000). *Acta Crystallographica Section D: Biological*
 285 *Crystallography*, **56**(8), 959–964.
- 286 Yeates, T. O. (1997). *Methods in enzymology*, **276**, 344–358.
- 287 Yeates, T. O. & Fam, B. C. (1999). *Structure*, **7**(2), R25–R29.

Synopsis

- 288 We reimplement and release an iterative method for map recovery for perfectly merohedrally
 twinned crystals in the presence of non-crystallographic symmetry and provide an independent
 metric of success.
-