Acta Cryst D
STRUCTURAL BIOLOGY

ISSN 2059-7983

Received 17 January 2022
Accepted 25 April 2022

Edited by E. F. Garman, University of Oxford, United Kingdom

**Keywords:** *xia*2.*multiplex*; multi-crystal data sets; data processing; data analysis; partial data sets; SARS-CoV-2.

**PDB references**: SARS-CoV-2 main protease, complex with Z31792168, 7qt5; complex with Z1367324110, 7qt6; complex with Z4439011520, 7qt7; complex with ABT-957, 7qt8; complex with Z4439011584, 7qt9

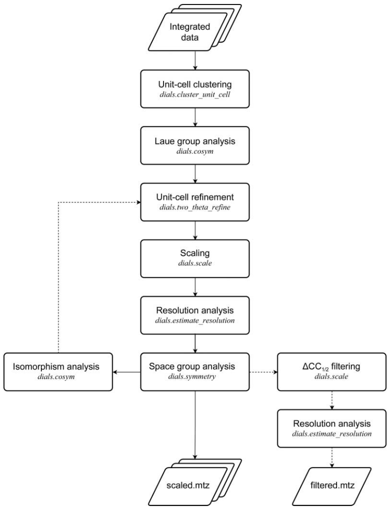
**Supporting information**: this article has supporting information at journals.iucr.org/d

# *xia*2.*multiplex*: a multi-crystal data-analysis pipeline

**Richard J. Gildea,[a]* James Beilsten-Edmands,[a] Danny Axford,[a] Sam Horrell,[a,b] Pierre Aller,[a] James Sandy,[a] Juan Sanchez-Weatherby,[a] C. David Owen,[a,b] Petra Lukacik,[a,b] Claire Strain-Damerell,[a,b] Robin L. Owen,[a] Martin A. Walsh[a,b] and Graeme Winter[a]**

[a]Diamond Light Source Ltd, Diamond House, Harwell Science and Innovation Campus, Didcot OX11 0DE, United Kingdom, and [b]Research Complex at Harwell, Harwell Science and Innovation Campus, Didcot OX11 0FA, United Kingdom. *Correspondence e-mail: richard.gildea@diamond.ac.uk

In macromolecular crystallography, radiation damage limits the amount of data that can be collected from a single crystal. It is often necessary to merge data sets from multiple crystals; for example, small-wedge data collections from micro-crystals, *in situ* room-temperature data collections and data collection from membrane proteins in lipidic mesophases. Whilst the indexing and integration of individual data sets may be relatively straightforward with existing software, merging multiple data sets from small wedges presents new challenges. The identification of a consensus symmetry can be problematic, particularly in the presence of a potential indexing ambiguity. Furthermore, the presence of non-isomorphous or poor-quality data sets may reduce the overall quality of the final merged data set. To facilitate and help to optimize the scaling and merging of multiple data sets, a new program, *xia*2.*multiplex*, has been developed which takes data sets individually integrated with *DIALS* and performs symmetry analysis, scaling and merging of multi-crystal data sets. *xia*2.*multiplex* also performs analysis of various pathologies that typically affect multi-crystal data sets, including non-isomorphism, radiation damage and preferential orientation. After the description of a number of use cases, the benefit of *xia*2.*multiplex* is demonstrated within a wider autoprocessing framework in facilitating a multi-crystal experiment collected as part of *in situ* room-temperature fragment-screening experiments on the SARS-CoV-2 main protease.



## 1. Introduction

Macromolecular structure determination routinely uses data sets obtained under cryogenic conditions from a single crystal. However, radiation damage limits the amount of data that can be collected from a single crystal. Cryocooling vastly increases the dose that can be tolerated by a single crystal, leading to the dominance of cryo-crystallography in macromolecular structure determination (Garman, 1999; Garman & Owen, 2007). However, it is often still necessary to merge multiple data sets from one or more crystals when dealing with radiation-sensitive samples and high-brilliance X-ray beams from third-generation light sources.

Multi-crystal data collection dates back to the early days of macromolecular crystallography (Kendrew *et al.*, 1960; Clemons *et al.*, 2001), but has seen a resurgence in recent years (Yamamoto *et al.*, 2017) as many scientifically important targets, such as membrane proteins and viruses, frequently yield small, weakly diffracting microcrystals. The development of crystallization in lipidic mesophases (Caffrey, 2003, 2015)

https://doi.org/10.1107/S2059798322004399

and the availability of microfocus beamlines (Evans *et al.*, 2011; Smith *et al.*, 2012) have facilitated data collection and structure solution of these difficult targets. Data-collection strategies for small weakly diffracting crystals rely on the collection of many small wedges of data, typically 5–10° per crystal, at cryogenic temperatures. For samples in the lipidic mesophase this is often preceded by X-ray raster scanning to identify the locations of crystals (Cherezov *et al.*, 2007, 2009; Rasmussen *et al.*, 2011; Rosenbaum *et al.*, 2011; Warren *et al.*, 2013). Such experiments are becoming increasingly automated thanks to developments such as *MeshAndCollect* (Zander *et al.*, 2015) and *ZOO* (Hirata *et al.*, 2019).

Multi-crystal data collections have also been applied to experimental phasing, where combining data from multiple crystals enhances weak anomalous signals, providing high-multiplicity data of sufficient quality to enable structure solution by single-wavelength anomalous dispersion (SAD; Liu *et al.*, 2011; Liu & Hendrickson, 2015) and sulfur SAD (S-SAD; Akey *et al.*, 2014; Liu *et al.*, 2014; Huang *et al.*, 2015, 2016; Olieric *et al.*, 2016).

Although cryogenic structures have provided the gold standard for structural analysis of macromolecules for decades, it has been shown that cryocooling can hide biologically significant structural features (Fraser *et al.*, 2009, 2011; Fischer *et al.*, 2015). Certain classes of macromolecular crystals, such as viruses, can also suffer when cryocooled. However, room-temperature data collection presents its own challenges, namely that radiation damage occurs at an absorbed dose one to two orders of magnitude lower than at cryogenic temperatures (Helliwell, 1988; Nave & Garman, 2005). In contrast to cryogenic data collections, an inverse dose-rate effect on crystal lifetime has been observed in room-temperature data (Southworth-Davies *et al.*, 2007). As a result, obtaining a complete room-temperature data set from a single crystal is difficult, so combining data from multiple crystals becomes necessary.

As the demand for room-temperature methods has increased, beamline developments have enabled routine room-temperature data collection directly from crystals in crystallization plates (*in situ*). This has the added benefit of eliminating the need for crystal harvesting (Axford *et al.*, 2012, 2015; Aller *et al.*, 2015), and a beamline, VMXi at Diamond Light Source, now exists that is dedicated to *in situ* data collection (Sanchez-Weatherby *et al.*, 2019). Advances in beamline and detector technology have enabled the collection of room-temperature data at a higher dose rate (Owen *et al.*, 2012, 2014; Schubert *et al.*, 2016), increasing the general applicability of room-temperature data collection (Aller *et al.*, 2015; Broecker *et al.*, 2018).

Merging multiple data sets from small wedges presents a number of challenges. For novel structures with unknown space group and unit-cell parameters, identifying a consensus symmetry can be problematic, particularly in the presence of indexing ambiguities (Brehm & Diederichs, 2014; Kabsch, 2014; Gildea & Winter, 2018). The presence of non-isomorphous or poor-quality data sets may also degrade the overall quality of the merged data set. Various methods have been developed to identify individual non-isomorphous data sets based on the comparison of unit-cell parameters (Foadi *et al.*, 2013; Zeldin *et al.*, 2015) or intensities (Giordano *et al.*, 2012; Santoni *et al.*, 2017; Diederichs, 2017) in order to combat this. Rogue data sets, or even individual bad images, can be identified by algorithms such as the $\Delta CC_{1/2}$ method described by Assmann *et al.* (2016) and implemented within *dials.scale* (Beilsten-Edmands *et al.*, 2020).

Microcrystal and room-temperature data-collection strategies are a compromise between maximizing the useful signal and minimizing the effects of radiation damage. By analysing manifestations of radiation damage, we can provide rapid feedback to guide an ongoing experiment and truncate the number of images used to produce the best final composite data set. The $R_{\rm cp}$ statistic introduced by Winter *et al.* (2019) can also be applied to multi-crystal data, under the assumption that the dose per image is approximately constant for all data sets. This may be appropriate for multi-crystal data collections where approximately uniformly sized crystals are bathed in the X-ray beam.

Preferential orientation of crystals can be a concern for some multi-crystal data collections, depending on the crystal symmetry and morphology, such as plate-like crystals *in situ* within a flat-bottomed crystallization well. Preferential orientation can lead to under-sampled regions of reciprocal space with systematically low-multiplicity or missing reflections, which may have adverse consequences on downstream phasing or refinement. Providing feedback on preferential orientation provides the opportunity for a user to make modifications to their experiment to minimize any resulting issues, for example by fully exploiting the available experimental geometry or changing the crystallization conditions or platform (Maeki *et al.*, 2016).

Structural biologists have become accustomed to the highly automated data analysis provided by synchrotron beamlines around the world (Holton & Alber, 2004; Winter, 2010; Vonrhein *et al.*, 2011; Winter & McAuley, 2011; Winter *et al.*, 2013; Monaco *et al.*, 2013; Yamashita *et al.*, 2018), typically obtaining automated data-processing results within minutes of the end of data collection for routine experiments. Multi-crystal experiments can generate large volumes of data in minutes, which brings new challenges in terms of bookkeeping and data analysis.

There are two primary aspects in which automated data analysis can support multi-crystal experiments. Firstly, rapid feedback from data analysis during beamtime can help to guide ongoing experiments, enabling a more efficient use of beamtime and allowing a user to more selectively screen sample conditions. Relevant feedback may include suitable metrics on merged data quality, *i.e.* completeness, multiplicity and resolution, and feedback on experimental pathologies, such as non-isomorphism, radiation damage and preferential orientation, that may hinder the experimental goals.

Secondly, after the completion of beamtime the user may be prepared to invest more time and effort in interactively optimizing the best overall data set for any given sample group. Automation is still highly relevant in this context, as the user may have collected data on many sample groups which they wish to process in a similar manner.

Standard autoprocessing pipelines such as *xia*2 (Winter, 2010) can handle multi-crystal data sets to some extent. However, they are optimized to process a small number of relatively complete data sets, rather than the many tens to hundreds of severely incomplete data sets that comprise a multi-crystal experiment. Recent software developments, for example *KAMO* (Yamashita *et al.*, 2018), have focused on automating the data processing of multi-crystal experiments.

Here, we present a new program, *xia*2*.multiplex*, which has been developed to facilitate the scaling and merging of multiple data sets. It takes data sets individually integrated with *DIALS* as input and performs symmetry analysis, scaling and merging, and analyses the various pathologies that typically affect multi-crystal data sets, including non-isomorphism, radiation damage and preferential orientation.

*xia*2*.multiplex* has been deployed as part of the autoprocessing pipeline at Diamond Light Source, including integration with downstream phasing pipelines such as *DIMPLE* (http://ccp4.github.io/dimple/) and *Big EP* (Sikharulidze *et al.*, 2016).

Using data sets collected as part of *in situ* room-temperature fragment-screening experiments on the SARS-CoV-2 main protease (M$^{\rm pro}$), we demonstrate the use of *xia*2*.multiplex* within a wider autoprocessing framework to give rapid feedback during a multi-crystal experiment, and how the program can be used to further improve the quality of the final merged data set.

## 2. Methods

Prior to using *xia*2*.multiplex*, each data set should be processed individually with *DIALS* (Winter *et al.*, 2018). Data may be processed either in the primitive *P*1 setting, or alternatively Bravais symmetry may be determined prior to integration using *dials.refine_bravais_settings*. It is not necessary to individually scale the data at this point.

Preliminary filtering of data sets is performed using hierarchical unit-cell clustering methods (Zeldin *et al.*, 2015). Histograms and scatterplots of the unit-cell distribution are generated for visual analysis, after which symmetry analysis and indexing-ambiguity resolution are performed with *dials.cosym*. Finally, the data are scaled with *dials.scale*, followed by radiation-damage and isomorphism analysis. The main sequence of steps taken by *xia*2*.multiplex* is outlined in Fig. 1.

### 2.1. Symmetry analysis

Initial analysis of the Patterson symmetry of the data is performed using *dials.cosym* (Gildea & Winter, 2018). This is an extension of the methods of Brehm & Diederichs (2014) for resolving indexing ambiguities in partial data sets and for completeness is reviewed here.

The maximum possible lattice symmetry compatible with the averaged unit cell is used to compile a list of all potential symmetry operations. The matrix of pairwise correlation coefficients is constructed, of size $(n \times m)^2$, where $n$ is the

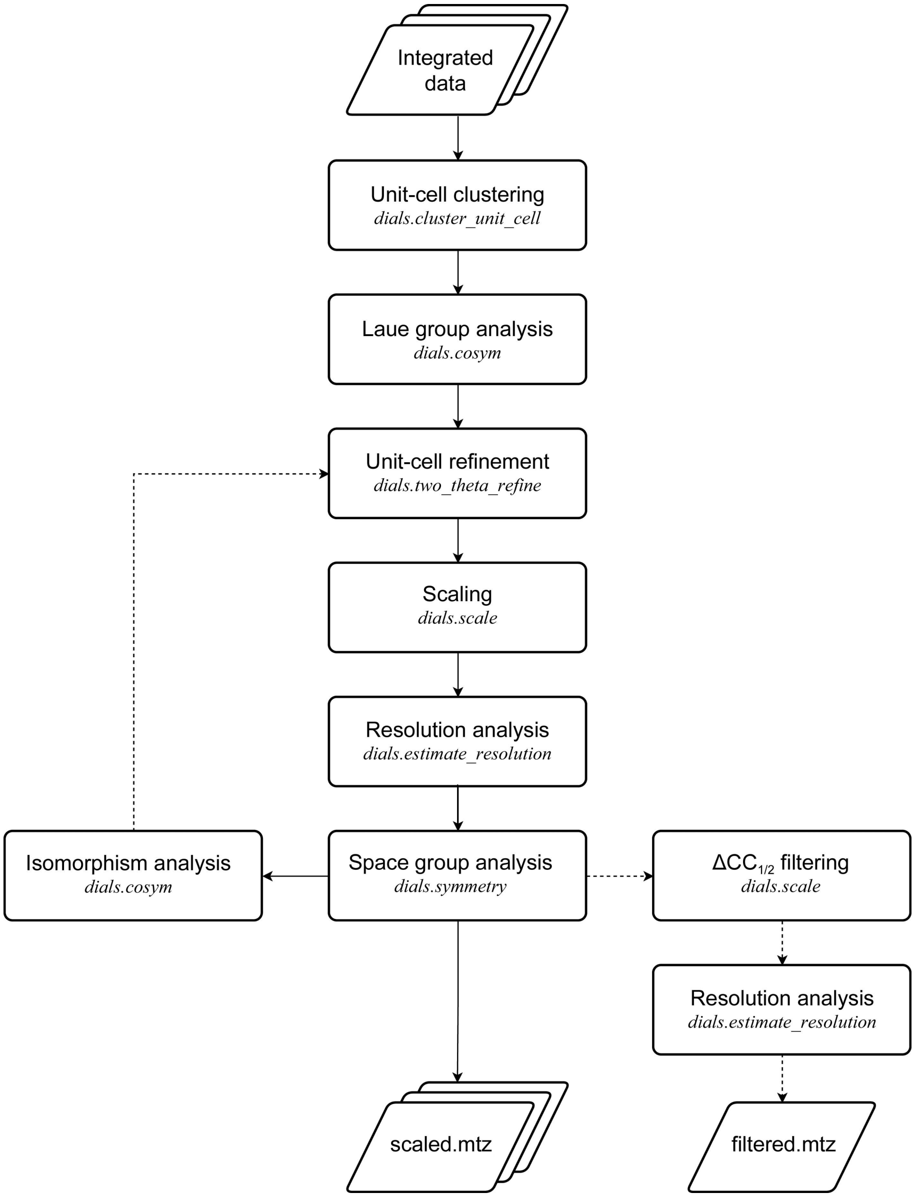


**Figure 1**
Flowchart outlining the main sequence of steps taken by *xia*2*.multiplex*. Optional steps are indicated by dashed arrows. The command-line programs used at each step are indicated.

number of data sets and $m$ is the number of symmetry operations in the lattice group. The Pearson correlation coefficient between data sets $i$ and $j$, after the application of the $k$th and $l$th symmetry operators respectively, is defined according to

$$r_{i_k,j_l} = \frac{\sum_h [I_{i_k}(h) - \overline{I}_{i_k}][I_{j_l}(h) - \overline{I}_{j_l}]}{\left\{\sum_h [I_{i_k}(h) - \overline{I}_{i_k}]^2 \sum_h [I_{j_l}(h) - \overline{I}_{j_l}]^2\right\}^{1/2}}, \qquad (1)$$

where $I_{i_k}(h)$ is the scaled intensity for data set $i$ of the reflection with Miller index $h$ after application of the $k$th symmetry operator.

Similarly to Brehm & Diederichs (2014), correlation coefficients are only calculated for pairs of data sets with three or more reflections in common. If a pair of data sets have two or fewer common reflections, then the correlation coefficient for that pair is assumed to be zero. The minimum number of common reflections required for the calculation of correlation coefficients is configurable in *dials.cosym* and *xia2.multiplex*.

Each data set is represented as $n \times m$ coordinates in an $m$-dimensional space. Use of an $m$-dimensional space allows the presence of up to $m$ orthogonal $\mathbf{x}_i$ clusters, where the orthogonality between clusters corresponds to a correlation coefficient $r_{i_k,j_l}$ close to zero. A modification of algorithm 2 of Brehm & Diederichs (2014), accounting for the additional symmetry-related copies of each data set, is used to iteratively minimize the function

$$\Phi = \sum_{i=1}^{n\times m} \sum_{j=1}^{n\times m} \left(r_{i_k,j_l} - \mathbf{x}_i \cdot \mathbf{x}_j\right)^2 \qquad (2)$$

using the L-BFGS minimization algorithm (Liu & Nocedal, 1989), with randomly assigned starting coordinates $\mathbf{x}$ in the range 0–1.

**2.1.1. Determination of the number of dimensions**. It is necessary to use a sufficient number of dimensions to represent any systematic variation that is present between data sets. Using $m$-dimensional space, where $m$ is equal to the number of symmetry operations in the maximum possible lattice symmetry, should be sufficient to represent any systematic variation present due to pseudosymmetry. However, choosing the optimal number of dimensions is a balance between underfitting and overfitting. Using more dimensions than is strictly necessary may reduce the stability of the minimization, particularly in the case of sparse data, where there is minimal overlap between data sets. As a result, we devised the following procedure to automatically determine the necessary number of dimensions.

(i) For each dimension in the range 2–$m$ minimize equation (2) and record the final value of the function.

(ii) Plot the resulting values as a function of the number of dimensions.

(iii) Determine the 'elbow' point of the plot, in a similar manner to that used by Zhang *et al.* (2006), to give the optimal number of dimensions.

Alternatively, the user may specify the number of dimensions to be used for the analysis.

**2.1.2. Identification of symmetry**. Patterson group symmetry is determined using an algorithm heavily influenced by the program *POINTLESS* (Evans, 2006, 2011).

Evans (2011) estimates the likelihood of a symmetry element $S_k$ being present, given the correlation coefficient $\mathrm{CC}_k$, as

$$p(S_k;\, \mathrm{CC}_k) = \frac{p(\mathrm{CC}_k;\, S_k)}{p(\mathrm{CC}_k;\, S_k) + p(\mathrm{CC}_k;\, !S_k)}. \qquad (3)$$

The probability of observing the correlation coefficient $\mathrm{CC}_k$ if the symmetry is present, $p(\mathrm{CC}_k;\, S_k)$, is modelled as a truncated Lorentzian centred on the expected value of CC if the symmetry is present, $E(\mathrm{CC};\, S)$, with a width parameter $\gamma = \sigma(\mathrm{CC}_k)$.

The distribution of $\mathrm{CC}_k$ if the symmetry is not present is modelled as

$$p(\mathrm{CC}_k;\, !S_k) = \frac{\int_0^1 p(\mathrm{CC};\, u)p(u)\,\mathrm{d}u}{\int_0^1 p(u)\,\mathrm{d}u}, \qquad (4)$$

$$p(u) = (1 - u^2)^{1/2}. \qquad (5)$$

Diederichs (2017) makes clear that the relationship between the results of the clustering procedure outlined above and the correlation coefficient $r_{ij}$ between two data sets $i$ and $j$ is

$$r_{ij} = \mathrm{CC}_i^* \times \mathrm{CC}_j^* \times \cos[\angle(\mathbf{x}_i, \mathbf{x}_j)]. \qquad (6)$$

The lengths of the vectors $|\mathbf{x}_i|$ are inversely related to the amount of random error, *i.e.* they provide an estimate of CC*. The maximum possible correlation coefficient between two data sets is given by the product of their CC* values. The angles between two vectors represent genuine systematic differences. For points related by genuine symmetry operations we expect $\cos[\angle(\mathbf{x}_i, \mathbf{x}_j)] \simeq 1$, whereas for points related by symmetry operations that are not present we expect $\cos[\angle(\mathbf{x}_i, \mathbf{x}_j)] = 0$.

We can therefore use $\cos[\angle(\mathbf{x}_i, \mathbf{x}_j)]$ in place of $\mathrm{CC}_k$, with $E(\mathrm{CC};\, S) = 1$. The estimated error $\sigma(\mathrm{CC}_k)$ used by Evans (2011) has a lower bound of 0.1, which is intended to avoid very small values of $\sigma(\mathrm{CC}_k)$ when large numbers of reflections contribute to the calculation of $\mathrm{CC}_k$. Since many reflections are contributing indirectly to the angles between any one pair of vectors, we can assume a value for the truncated Lorentzian width parameter of $\gamma = \sigma(\mathrm{CC}_k) = 0.1$. The average of all observations of $\cos[\angle(\mathbf{x}_i, \mathbf{x}_j)]$ corresponding to a given symmetry operator $S_k$ is used as an estimate of $\mathrm{CC}_k$.

Once a score has been assigned to each potential symmetry operator, all possible point groups compatible with the lattice group are scored as in Appendix *A*2 of Evans (2011),

(i) Find the highest lattice symmetry compatible with the unit-cell dimensions.

(ii) Score each potential rotation operation using all reflections related by that operation.

(iii) Score possible subgroups (Patterson groups) according to combinations of symmetry elements.

Once the most likely Patterson group has been identified by the above procedure, it is then relatively straightforward to assign a suitable re-indexing operation to each data set to ensure that all data sets are consistently indexed. Firstly, a high-density point is chosen as a seed for the cluster. Then, for each data set, the nearest symmetry copy of that data set to the seed is identified. The symmetry operation corresponding to this symmetry copy is then the re-indexing operation for this data set.

### 2.2. Unit-cell refinement

After symmetry determination, an overall best estimate of the unit cell is obtained by refinement of the unit-cell parameters against the observed $2\theta$ angles using *dials.two_theta_refine* (Winter *et al.*, 2022). This program minimizes the unit-cell constants against the difference between observed and calculated $2\theta$ values, which are determined from background-subtracted integrated centroids. This provides an overall best estimate of the unit cell that is a suitable representative average for use in subsequent downstream phasing and refinement.

### 2.3. Scaling

Data are then scaled using the *physical* scaling model in *dials.scale* (Beilsten-Edmands *et al.*, 2020). *xia2.multiplex* uses the automatic scaling-model selection within *dials.scale* to enable a suitable model parameterization for both the cases of small-wedge data sets and large-wedge data sets. For small-wedge data sets, each data set is corrected by an overall scale factor and relative $B$ factor that vary smoothly as a function of rotation angle, whereas the absorption correction of the *physical* scaling model is not used as this correction requires the sampling of a diverse set of scattering paths through the sample. For large-wedge data sets, the absorption correction of the *physical* scaling model is used in addition to the smoothly varying scale and $B$-factor corrections. The strength of the absorption correction can optionally be set to low (the default), medium or high. This option adjusts the absorption model parameterization and restraints to enable a correction that more closely matches the expected relative absorption, which can be high at long wavelengths or for crystals containing heavy atoms.

Several rounds of outlier rejection are performed during scaling to remove individual reflections that have poor agreement with their symmetry equivalents. The uncertainties of the intensities are also adjusted during scaling by optimizing a single error model across all data sets in order to account for the effects of systematic errors, which tend to increase the variability of intensities within each symmetry-equivalent group. Optionally, for anomalous data, Friedel pairs can be treated separately in scaling, which can increase the strength of the detected anomalous signal.

### 2.4. Estimation of resolution cutoff

After the data have successfully been scaled, *dials.estimate_resolution* is used to estimate a suitable resolution cutoff for the data. By default, this is determined from a fit of a hyperbolic tangent to $CC_{1/2}$ calculated in resolution bins, similar to that used by *AIMLESS* (Evans & Murshudov, 2013). The resolution cutoff is chosen as the resolution where the fit curve reaches $CC_{1/2} = 0.3$ (this cutoff value can be controlled by the user). A second round of scaling with *dials.scale* is then performed after application of the resolution cutoff. The default cutoff value of $CC_{1/2} = 0.3$ is chosen as one that works well in the context of autoprocessing in order to provide a consistent set of merging statistics for judging data quality during an ongoing experiment. Suitable cutoff values may depend on the downstream data-processing requirements, but the current gold standard for publication is to use 'paired refinement' to determine the resolution at which including higher resolution data in refinement no longer improves the model (Karplus & Diederichs, 2012).

### 2.5. Space-group identification

After the data have been scaled in the Patterson group identified by *dials.cosym* (Section 2.1), analysis of potential systematic absences is performed by *dials.symmetry* in order to assign a final space group. In this analysis, the existence of each potential screw axis allowed by the Patterson group is tested by calculating the $z$-score based on the deviation from zero of the merged $\langle I/\sigma(I)\rangle$ for the expected absent reflections. From the individual $z$-scores, a likelihood of the presence of each screw axis is determined; these are combined to score and select the most likely non-enantiogenic space group.

### 2.6. Analysis of radiation-damage indicators

*xia2.multiplex* performs a number of analyses that can be useful in assessing the extent of any radiation damage which may be present. Plots of scale factor and $R_{\rm merge}$ versus image number are generated to look for any trends associated with radiation damage. The $R_{\rm cp}$ statistic introduced by Winter *et al.* (2019) can also be applied to multi-crystal data. This statistic accumulates the pairwise measured intensity differences as a function of dose (or image number). In the absence of accurate dose information for each data set, it is necessary to make the assumption that the dose per image is approximately constant for all data sets. In order to assess how many images per crystal are necessary to achieve a complete data set, a plot of completeness versus dose is also generated.

### 2.7. Isomorphism analysis

Unit-cell clustering, as implemented in *BLEND* (Foadi *et al.*, 2013) and elsewhere (Zeldin *et al.*, 2015), is used by *xia2.multiplex* as a preliminary filtering step to reject any highly non-isomorphous data sets.

*xia2.multiplex* implements two alternative intensity-based clustering methods that are suitable for the identification and analysis of non-isomorphism, once symmetry determination, resolution of indexing ambiguities and scaling have been carried out as described above. Clustering on correlation coefficients (Giordano *et al.*, 2012; Santoni *et al.*, 2017; Yamashita *et al.*, 2018) begins by calculating a matrix of pairwise correlation coefficients:

$$r_{i,j} = \frac{\sum_h [I_i(h) - \overline{I}_i][I_j(h) - \overline{I}_j]}{\left\{\sum_h [I_i(h) - \overline{I}_i]^2 \sum_h [I_j(h) - \overline{I}_j]^2\right\}^{1/2}}. \qquad (7)$$

A distance matrix defined as $d_{i,j} = 1 - r_{i,j}$ is provided as input to the *SciPy* (Virtanen *et al.*, 2020) hierarchical clustering routine using the *average* linkage method. Clusters are sorted by distance, and the completeness and multiplicity of each cluster are reported. Optionally, *xia*2.*multiplex* can scale and merge the data sets defined by each cluster that meet user-defined criteria for minimum completeness or multiplicity.

A second intensity-based clustering method follows that described by Diederichs (2017), who demonstrated that the methods of Brehm & Diederichs (2014) could be generalized to search for any systematic differences between data sets, not just those caused by an indexing ambiguity. In addition to its use for identifying the Patterson symmetry (Section 2.1), *dials.cosym* can also be used for analysis of non-isomorphism. In this mode, rather than searching for the presence of potential additional symmetry operators, the matrix of pairwise correlation coefficients of size $n^2$ reduces to equation (7). The function defined by equation (2) is minimized as before to obtain a representation of the similarity between data sets in a reduced dimensional space.

As made clear by Diederichs (2017), the length of a vector $\mathbf{x}_i$ is inversely proportional to the random error in data set $\mathbf{X}_i$. The angle between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ corresponds to the level of systematic error between data sets $\mathbf{X}_i$ and $\mathbf{X}_j$, and thus can be used to estimate the degree of non-isomorphism between these data sets. Analysis of the angular separation of vectors $\mathbf{x}$ can be used to identify groups of systematically different data sets. Hierarchical clustering on the cosines of the angles between vectors is performed to identify possible groupings of data sets for further investigation. Optionally, *xia*2.*multiplex* can rescale multiple subsets of data, which can be controlled by specifying a maximum number of clusters to merge and/or the minimum required completeness or multiplicity for a cluster.

The final approach to isomorphism analysis implemented within *xia*2.*multiplex* is the $\Delta CC_{1/2}$ method described by Assmann *et al.* (2016) and implemented within *dials.scale* (Beilsten-Edmands *et al.*, 2020). If $\Delta CC_{1/2}$ filtering is selected then *xia*2.*multiplex* will perform additional scaling with *dials.scale*, rejecting any data sets that are identified as significant outliers according to $\Delta CC_{1/2}$ analysis. Whilst this approach may not be suitable if there are two or more significant non-isomorphous populations, it may give useful results if there are a small number of data sets that are systematically different from the majority.

### 2.8. Preferential orientation

The report generated by *xia*2.*multiplex* includes stereographic projections of the crystal orientation relative to the laboratory frame generated with *dials.stereographic_projection*. A random distribution of points (each point corresponds to a crystal or its symmetry equivalent) in a stereographic projection suggests a random distribution of crystal orientations, whereas a systematic nonrandom distribution may be indicative of preferential crystal orientation.

*xia*2.*multiplex* also generates a number of plots that can aid in the analysis of the distribution of multiplicities.

A new command, *dials.missing_reflections*, has been developed to identify connected regions of missing reflections in reciprocal space. Prior to performing the analysis, it is necessary to map centred unit cells to the primitive setting in order to avoid systematically absent reflections complicating the analysis. The complete set of possible Miller indices is generated and expanded to cover the full sphere of reciprocal space by the application of symmetry operators belonging to the known space group. This allows the identification of connected regions that cross the boundary of the asymmetric unit. Nearest-neighbour analysis is used to construct a graph of connected regions, which is then used to perform connected components analysis to identify each connected region of missing reflections. Miller indices for missing reflections are then mapped back to the asymmetric unit in order to identify the set of unique Miller indices belonging to each region. A sorted list of connected regions is reported to the user, detailing the resolution range spanned by each region and the number and proportion of total reflections comprising each region.

## 3. Deployment of *xia*2.*multiplex* at Diamond Light Source

*xia*2.*multiplex*, as described above, has been deployed as part of the autoprocessing pipeline at Diamond Light Source. A series of partial data sets are collected from a set of related crystals, for example from multiple crystals within one or more drops in a crystallization plate (Sanchez-Weatherby *et al.*, 2019), sample loop or sample mesh. After the end of each data collection, the partial data set is processed individually with *DIALS* via *xia*2. On the successful completion of *xia*2, a *xia*2.*multiplex* processing job is triggered using all successful *xia*2 results from this and prior data collections as input. The *xia*2.*multiplex* results, including merging statistics, are recorded in ISPyB (Delagenière *et al.*, 2011) for presentation to the user via SynchWeb (Fisher *et al.*, 2015), where results are typically available within minutes of the end of data collection. Prior to data collection, users may define groups of related samples for combining with *xia*2.*multiplex* either via SynchWeb or via a configuration file in a pre-defined location. In the absence of this information, *xia*2.*multiplex* will only combine data collected from the same *sample*, *i.e.* loop, mesh or well within a crystallization plate.

If a PDB file has been associated with the data collection, then automated structure refinement is performed

Table 1
Data-collection, merging and refinement statistics for lysozyme room-temperature *in situ* heavy-atom soaks using all data sets.

Values in parentheses are for the highest resolution shell.

| Heavy atom | Au | KBr | KI | NaBr | NaI | Sm |
|---|---|---|---|---|---|---|
| Data collection | | | | | | |
| Exposure time (s) | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| Ω width (°) | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Wavelength (Å) | 0.9028 | 0.9193 | 1.8233 | 0.9193 | 1.8233 | 1.6947 |
| No. of images | 200 | 200 | 200 | 200 | 200 | 200 |
| No. of data sets | 26 | 60 | 73 | 77 | 49 | 24 |
| Crystal parameters | | | | | | |
| Space group | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ |
| $a$, $b$, $c$ (Å) | 78.58, 78.58, 38.27 | 79.09, 79.09, 37.98 | 79.16, 79.16, 38.01 | 79.10, 79.10, 38.03 | 79.16, 79.16, 38.01 | 79.11, 79.11, 37.91 |
| Data statistics | | | | | | |
| Resolution range (Å) | 78.73–1.28 (1.33–1.28) | 79.23–1.37 (1.42–1.37) | 79.23–1.98 (2.06–1.98) | 79.23–1.38 (1.43–1.38) | 79.22–1.98 (2.06–1.98) | 79.19–1.82 (1.89–1.82) |
| No. of unique reflections | 59067 (5906) | 48155 (4545) | 15387 (967) | 47418 (4671) | 15152 (763) | 18825 (684) |
| Multiplicity | 16.9 (7.0) | 30.9 (4.0) | 38.0 (2.1) | 39.9 (5.8) | 25.7 (1.8) | 13.2 (1.4) |
| $R_{\rm merge}$ | 0.398 (−8.055) | 0.118 (1.322) | 0.161 (0.370) | 0.232 (15.279) | 0.160 (0.421) | 0.473 (−5.938) |
| $R_{\rm meas}$ | 0.409 (−8.728) | 0.119 (1.513) | 0.163 (0.456) | 0.234 (17.102) | 0.162 (0.540) | 0.490 (−7.623) |
| $R_{\rm p.i.m.}$ | 0.094 (−3.287) | 0.018 (0.710) | 0.023 (0.261) | 0.033 (7.304) | 0.028 (0.331) | 0.127 (−4.735) |
| Completeness (%) | 100.0 (100.0) | 99.5 (94.4) | 96.3 (60.9) | 100.0 (98.8) | 94.7 (47.8) | 91.7 (33.3) |
| $\langle I/\sigma(I)\rangle$ | 6.8 (0.2) | 19.2 (0.7) | 19.7 (1.0) | 21.8 (0.9) | 16.5 (0.9) | 26.2 (3.4) |
| $CC_{1/2}$ | 0.991 (0.016) | 0.999 (0.280) | 0.998 (0.756) | 0.992 (0.005) | 0.997 (0.568) | 0.954 (0.040) |
| $CC_{\rm anom}$ | −0.031 (−0.011) | 0.415 (0.055) | 0.421 (0.203) | −0.017 (0.022) | 0.107 (0.008) | −0.178 (−0.174) |
| Phasing | | | | | | |
| Substructure solution | Yes | Yes | Yes | Yes | Yes | Yes |
| Residues autotraced | 127 | 120 | 79 | 120 | 77 | 91 |
| $R_{\rm work}$ | 0.2682 | 0.2120 | 0.2353 | 0.2126 | 0.2396 | 0.2476 |
| $R_{\rm free}$ | 0.2834 | 0.2346 | 0.2705 | 0.2322 | 0.2771 | 0.2737 |
| Anomalous peak height ($\sigma$) | 45.45 | 45.40 | 23.22 | 46.35 | 21.47 | 44.04 |

with *DIMPLE* using the merged reflections output by *xia*2*.multiplex*.

## 4. Examples

### 4.1. Room-temperature *in situ* experimental phasing

Using data from Lawrence *et al.* (2020), we showcase the application of *xia*2*.multiplex* to multi-crystal room-temperature *in situ* data sets from heavy-atom soaks of lysozyme crystals, demonstrating successful experimental phasing using the resulting *xia*2*.multiplex* output. Data from lysozyme crystals soaked with six different heavy-atom solutions were processed individually with *DIALS* via *xia*2 followed by symmetry determination (Figs. 3*a* and 3*b*), scaling and merging with *xia*2*.multiplex*. Partial data sets identified as outliers according to $\Delta CC_{1/2}$ were rejected in an automated iterative process with *xia*2*.multiplex*. Data-processing statistics for each heavy-atom soak, with and without $\Delta CC_{1/2}$ filtering of outlier data sets, are shown in Tables 1 and 2. Phasing was performed with *fast_ep* using *SHELXC*/*D*/*E* (Sheldrick, 2010). Structure refinement was performed by *REFMAC*5 (Murshudov *et al.*, 2011) via *DIMPLE* using PDB entry 6qqf (Gotthard *et al.*, 2019) as the reference structure. Anomalous difference maps were calculated by *ANODE* (Thorn & Sheldrick, 2011) via the `--anode` option in *DIMPLE*.

Significant anomalous signal was observed, as indicated in the *SHELXC* plot of $\langle d''/\sigma(I)\rangle$ versus resolution (Fig. 2*a*). Substructure searches with *SHELXD* were successful (Fig. 2*b*), and traceable electron-density maps were obtained by *SHELXE*. Anomalous difference maps calculated by *ANODE* via *DIMPLE* indicated the presence of significant anomalous difference peaks (Figs. 2*c* and 2*d*).

To assess the impact of $\Delta CC_{1/2}$ filtering on the resulting anomalous signal, we performed experimental phasing and structure refinement (via *DIMPLE*) and calculated anomalous difference maps using data both with and without $\Delta CC_{1/2}$ filtering of outliers. Substructure solution and autotracing were successful in both cases. $\Delta CC_{1/2}$ filtering also resulted in improved merging statistics, typically in $CC_{1/2}$, $CC_{\rm anom}$, $\langle d''/\sigma(I)\rangle$, $\langle I/\sigma(I)\rangle$ and $R_{\rm p.i.m.}$ versus resolution (Tables 1 and 2). For the NaBr and Sm soaks there are particularly significant improvements in $R_{\rm work}$ and $R_{\rm free}$ after $\Delta CC_{1/2}$ filtering. These two soaks also correspond to the data sets that showed the largest improvement in anomalous difference peak height after the removal of outlier data sets (Fig. 2*d*).

We note that merging statistics such as correlation coefficients and *R* factors, which are calculated only on the unmerged intensity values without taking into account their errors, can be affected by regions of lower data quality that are suitably down-weighted with larger errors during scaling. The presence of these regions, however, does not adversely affect the resulting merged intensities, which are appropriately weighted. This disparity is most likely to be evident for high-multiplicity data with regions of significant radiation damage, in which case merged data-quality indicators are most representative of the data quality.

As outlined in Section 2.5, several different methods are available in *xia*2*.multiplex* for identifying outlier data sets. Above, we used $\Delta CC_{1/2}$ filtering to identify and exclude outlier partial data sets. Visualization of the distribution and

**Table 2**
Data-collection, merging and refinement statistics for lysozyme room-temperature *in situ* heavy-atom soaks after the removal of data sets identified by $\Delta CC_{1/2}$ analysis.

Values in parentheses are for the highest resolution shell.

| Heavy atom | Au | KBr | KI | NaBr | NaI | Sm |
|---|---|---|---|---|---|---|
| Data collection | | | | | | |
| Exposure time (s) | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| Ω width (°) | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Wavelength (Å) | 0.9028 | 0.9193 | 1.8233 | 0.9193 | 1.8233 | 1.6947 |
| No. of images | 200 | 200 | 200 | 200 | 200 | 200 |
| No. of data sets | 22 | 59 | 72 | 75 | 48 | 22 |
| Crystal parameters | | | | | | |
| Space group | $P4_12_12$ | $P4_12_1\ 2$ | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ | $P4_12_12$ |
| $a$, $b$, $c$ (Å) | 78.58, 78.58, 38.27 | 79.09, 79.09, 37.98 | 79.17, 79.17, 38.01 | 79.10, 79.10, 38.03 | 79.16, 79.16, 38.01 | 79.11, 79.11, 37.91 |
| Data statistics | | | | | | |
| Resolution range (Å) | 39.31–1.27 (1.32–1.27) | 79.23–1.35 (1.40–1.35) | 79.23–1.98 (2.06–1.98) | 79.24–1.33 (1.38–1.33) | 79.22–1.98 (2.06–1.98) | 79.19–1.82 (1.89–1.82) |
| No. of unique reflections | 60440 (6006) | 49744 (4213) | 15363 (945) | 51260 (3621) | 15096 (726) | 18509 (589) |
| Multiplicity | 13.1 (5.1) | 29.1 (2.7) | 36.1 (2.0) | 34.4 (2.2) | 24.2 (1.7) | 10.7 (1.3) |
| $R_{\rm merge}$ | 0.137 (2.345) | 0.115 (1.161) | 0.163 (0.336) | 0.111 (1.106) | 0.156 (0.346) | 0.074 (0.162) |
| $R_{\rm meas}$ | 0.142 (2.620) | 0.116 (1.390) | 0.165 (0.417) | 0.112 (1.361) | 0.159 (0.440) | 0.077 (0.216) |
| $R_{\rm p.i.m.}$ | 0.036 (1.131) | 0.018 (0.741) | 0.023 (0.241) | 0.015 (0.772) | 0.028 (0.266) | 0.020 (0.141) |
| Completeness (%) | 100.0 (99.6) | 98.4 (83.7) | 96.1 (59.5) | 96.8 (68.5) | 94.4 (45.5) | 90.2 (28.6) |
| $\langle I/\sigma(I)\rangle$ | 7.2 (0.2) | 20.3 (0.8) | 20.1 (1.3) | 21.9 (0.6) | 17.3 (1.4) | 25.3 (3.7) |
| $CC_{1/2}$ | 0.997 (0.187) | 0.999 (0.313) | 0.997 (0.802) | 0.999 (0.315) | 0.994 (0.736) | 0.996 (0.894) |
| $CC_{\rm anom}$ | 0.313 (0.011) | 0.455 (−0.020) | 0.565 (−0.100) | 0.423 (0.089) | 0.485 (0.055) | 0.656 (0.024) |
| Phasing | | | | | | |
| Substructure solution | Yes | Yes | Yes | Yes | Yes | Yes |
| Residues autotraced | 116 | 103 | 85 | 114 | 54 | 119 |
| $R_{\rm work}$ | 0.2668 | 0.2116 | 0.2355 | 0.2140 | 0.2420 | 0.2078 |
| $R_{\rm free}$ | 0.2820 | 0.2333 | 0.2704 | 0.2332 | 0.2753 | 0.2424 |
| Anomalous peak height ($\sigma$) | 46.47 | 45.43 | 23.22 | 47.63 | 21.85 | 47.00 |

hierarchical clustering on unit-cell parameters for the Sm soak (Figs. 3*e* and 3*f*) identifies data set 11 as an outlier, which was also the first data set to be excluded by $\Delta CC_{1/2}$ filtering. Similarly, hierarchical clustering on pairwise correlation coefficients (Fig. 4*a*) and on the cosines of the angles between vectors $\mathbf{x}$ (Figs. 3*c*, 3*d* and 4*b*) both identify data set 11 as an outlier. Whilst in this case all available methods for isomorphism analysis identified data set 11 as the least compatible data set, it is beneficial to have an array of different methods available, as the best method for a particular system may depend on the nature of any non-isomorphism involved.

### 4.2. TehA

Previously published *in situ* data for *Haemophilus influenzae* TehA (Axford *et al.*, 2015) were used to further demonstrate the applicability of *xia*2.*multiplex* and the tools contained therein. 73 partial data sets were processed individually with *DIALS* via *xia*2, providing no prior space group or unit-cell information. 71 successfully integrated data sets were provided as input to *xia*2.*multiplex*, where data were combined and scaled using *dials.cosym* and *dials.scale*. Two data sets were identified as having inconsistent unit cells by preliminary filtering and were removed, leaving 69 data sets for subsequent symmetry analysis and scaling. Structure refinement was performed by *REFMAC*5 via *DIMPLE*. Data-processing and refinement statistics using all data and only those remaining after filtering by $\Delta CC_{1/2}$ are shown in Table 3.

The maximum possible lattice symmetry was determined to be $R{-}3m{:}H$, with a maximum of six symmetry operations. Analysis of the value given by equation (2) as a function of the number of dimensions identified that two dimensions were sufficient to explain the variation between data sets. Further symmetry analysis with *dials.cosym* correctly identified the Patterson group as $R{-}3{:}H$, resolving the indexing ambiguity present in this space group (Figs. 5*a* and 5*b*).

The best overall unit cell was determined by *dials.two_theta_refine* as $a = b = 98.76$, $c = 136.77$ Å, and data were scaled together with *dials.scale*. Resolution analysis with *dials.estimate_resolution* identified 2.14 Å as the resolution where the fit of a hyperbolic tangent to $CC_{1/2} \simeq 0.3$.

Six cycles of scaling and filtering were performed by *dials.scale*, where exclusion was performed on whole data sets. A single outlier data set (using a cutoff of $3\sigma$) was removed in each of the first five cycles, removing a total of 6.2% of the reflections. No significant outliers were identified in the sixth and final cycle.

Structure refinement was performed by *REFMAC*5 via *DIMPLE* with the model from PDB entry 4ycr (Axford *et al.*, 2015), using all scaled data and after filtering of outliers using the $\Delta CC_{1/2}$ method. Filtering of outlier data sets leads to a slight improvement in the merging statistics, particularly in $\langle I/\sigma(I)\rangle$ and $R_{\rm p.i.m.}$. There is also a slight reduction in the $R_{\rm work}$ and $R_{\rm free}$ reported by *REFMAC*5.

Stereographic projections of crystal orientations with *dials.stereographic_projection* shows that preferential crystal orientatation may be an issue for this experiment (Figs. 5*c* and

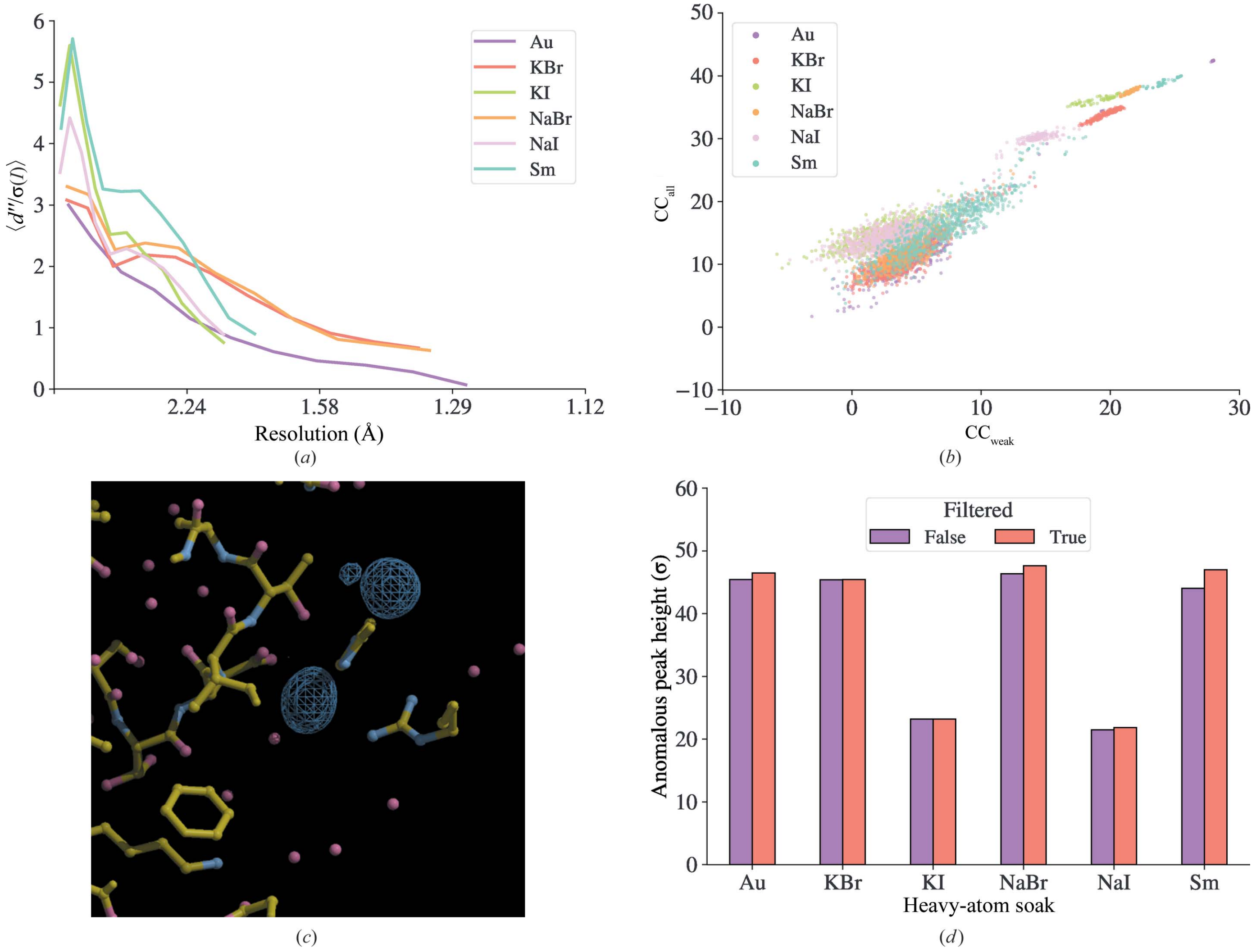


**Figure 2**
Experimental phasing and anomalous signal from multi-crystal room-temperature *in situ* experiments using lysozyme crystals soaked with various heavy-atom solutions. (*a*) *SHELXC* plot of $\langle d''/\sigma(I)\rangle$. (*b*) $CC_{\rm all}$ versus $CC_{\rm weak}$ after substructure solution with *HKL2MAP*/*SHELXD*. (*c*) Anomalous difference map peaks identified by *ANODE* via *DIMPLE* for lysozyme Au soaks. Contours are drawn at $4\sigma$. (*d*) Anomalous difference map peak heights identified by *ANODE* via *DIMPLE* with and without filtering of outlier regions of data sets.

5*d*). Fig. 5(*e*) and 5(*f*) show the consequences that this has on the distribution of multiplicities in the resulting data set. Analysis with *dials.missing_reflections* identifies a single region of missing reflections, comprising 1390 reflections (5.2%) covering the range 53.41–2.14 Å.

## 5. Applications

### 5.1. *In situ* ligand-screening studies of SARS-CoV-2 M$^{\rm pro}$

With the emergence of the novel coronavirus SARS-CoV-2 and the associated coronavirus disease 2019 (COVID-19), SARS-CoV-2 M$^{\rm pro}$ quickly emerged as one of the primary targets for antiviral drug development (Jin *et al.*, 2020, 2021; Walsh *et al.*, 2021). Fragment-screening experiments using the XChem platform at Diamond Light Source (Cox *et al.*, 2016; Collins *et al.*, 2017; Krojer *et al.*, 2017) screened over 1250 unique chemical fragments, yielding 74 fragment hits (Douangamath *et al.*, 2020).

Fragment-screening experiments such as these are typically carried out using conventional cryogenic conditions to minimize the effects of radiation damage, with each structure being obtained from a single crystal. Room-temperature data, however, can usefully identify or rule out structural artefacts induced by pushing the temperature far from the biologically relevant level (Durdagi *et al.*, 2021; Guven *et al.*, 2021).

Over the course of several beamline visits, room-temperature *in situ* data were collected for 30 ligand soaks that had previously shown ligand binding under cryogenic conditions. Here, we highlight room-temperature data collections for five ligand soaks that showed evidence of ligand binding at room temperature: Z1367324110 (PDB entry 5r81) and Z31792168 (PDB entry 5r84) (Douangamath *et al.*, 2020), Z4439011520 (PDB entry 5rh5), Z4439011584 (PDB entry 5rh7) and ABT-957 (PDB entry 7aeh) (Redhead *et al.*, 2021).

Data were collected on beamline I24 at Diamond Light Source with a Dectris PILATUS 3 6M detector using a 30 ×

30 µm beam with a flux of approximately 2 × $10^{11}$ photons $s^{-1}$. 20° of data were collected per crystal with an oscillation range of 0.1° and an exposure time of 0.02 s per image. The starting angle was varied to maximize the total angular range within

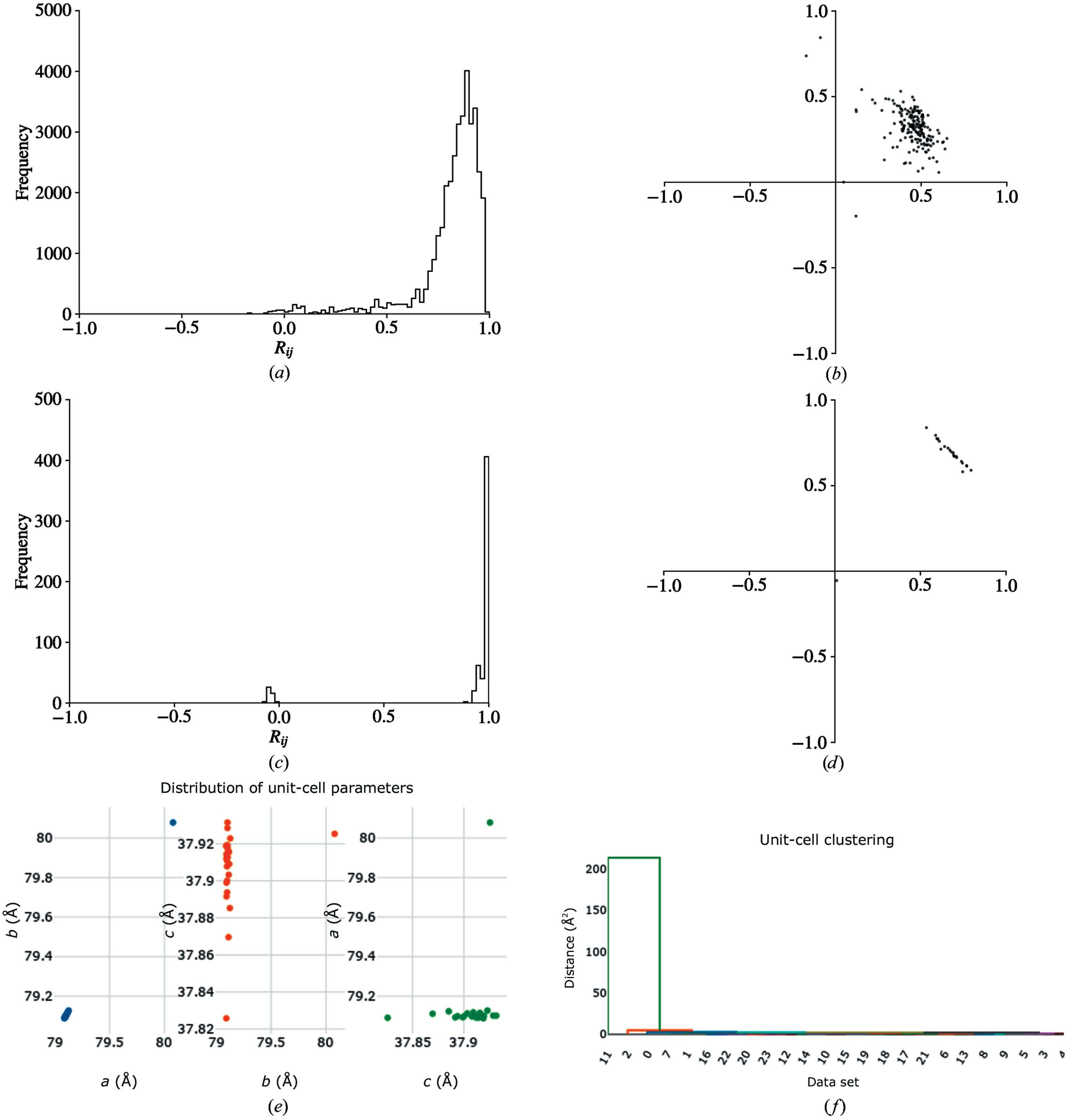


**Figure 3**

*dials.cosym* plots for data from lysozyme Sm soaks as described in Section 4.1. (*a*) Histogram of $(n \times m)^2$ pairwise $R_{ij}$ correlation coefficients and (*b*) the $(n \times m)$ vectors **x** determined by the minimization of equation (2) during symmetry determination with *dials.cosym*. The $R_{ij}$ correlation coefficients are clustered towards 1 and the majority of the vectors **x** form a single cluster, suggesting the absence of an indexing ambiguity, *i.e.* the Patterson group of the data set corresponds to the maximum lattice symmetry. (*c*, *d*) As above but after symmetry determination and scaling. The distribution of the $n^2$ $R_{ij}$ correlation coefficients is sharpened towards 1 as scaling improves the internal consistency of the data. There is also an effect from multiplicity when comparing with (*a*), as here the $n^2$ $R_{ij}$ values are calculated in the highest symmetry group for the lattice. All but one of the *n* vectors **x** form a tight cluster, with the vector lengths close to 1. Visualization of (*e*) the distribution of unit-cell parameters and (*f*) clustering on unit-cell parameters suggests the presence of an outlier data set.

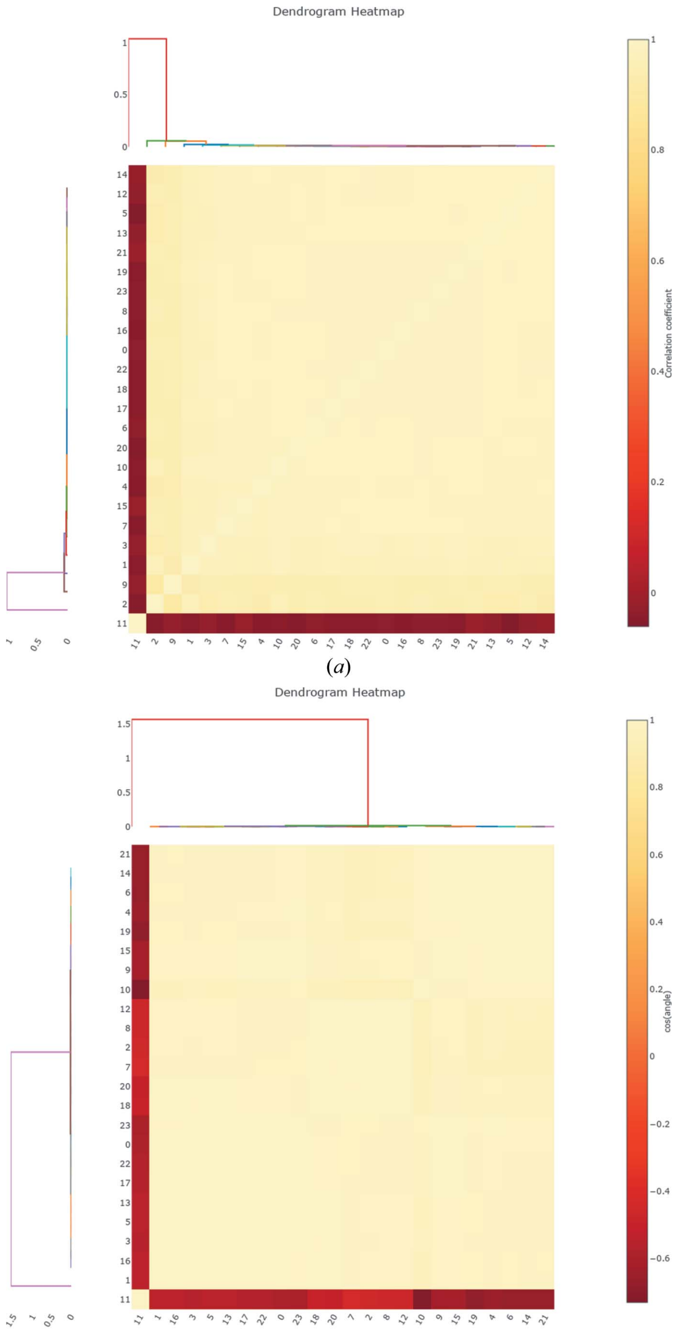
the constraints imposed by the experimental setup. Based on typical crystal dimensions of 50 × 50 × 5 µm, the X-ray dose per data collection was estimated to be in the range 50–67 kGy using *RADDOSE-3D* (Zeldin *et al.*, 2013; Bury *et al.*, 2018). *RADDOSE-3D* input and output files are included in the supporting information.

As described in Section 3, data sets were automatically processed individually with *DIALS* via *xia2*, followed by combined scaling and merging after each data collection with *xia2.multiplex*. Automatic structure refinement and difference map calculations were performed using *DIMPLE*.

**Table 3**
Data-collection, merging and refinement statistics for TehA.

Values in parentheses are for the highest resolution shell.

| | All data | $\Delta CC_{1/2}$-filtered data |
|---|---|---|
| Data collection | | |
| Exposure time (s) | 0.04 | 0.04 |
| Ω width (°) | 0.2 | 0.2 |
| Transmission (%) | 12.34 | 12.34 |
| No. of images | 20–50 | 20–50 |
| No. of data sets | 69 | 64 |
| Crystal parameters | | |
| Space group | *R*3:*H* | *R*3:*H* |
| *a*, *b*, *c* (Å) | 98.76, 98.76, 136.77 | 98.76, 98.76, 136.77 |
| Data statistics | | |
| Resolution range (Å) | 72.56–2.13 (2.21–2.13) | 72.56–2.14 (2.22–2.14) |
| No. of unique reflections | 26203 (2415) | 25851 (2396) |
| Multiplicity | 13.7 (6.8) | 13.0 (6.7) |
| $R_{\rm merge}$ | 0.315 (−1033.253) | 0.162 (2.508) |
| $R_{\rm meas}$ | 0.326 (−1113.925) | 0.167 (2.703) |
| $R_{\rm p.i.m.}$ | 0.078 (−406.346) | 0.040 (0.981) |
| Completeness (%) | 94.1 (86.3) | 94.2 (86.8) |
| $\langle I/\sigma(I)\rangle$ | 13.1 (1.3) | 13.9 (1.5) |
| $CC_{1/2}$ | 0.988 (0.285) | 0.996 (0.360) |
| $CC_{\rm anom}$ | −0.002 (0.004) | 0.073 (0.045) |
| $R_{\rm work}$ | 0.1515 | 0.1499 |
| $R_{\rm free}$ | 0.1726 | 0.1711 |



(*a*)

(*b*)

**Figure 4**
Hierarchical clustering (*a*) on pairwise correlation coefficients and (*b*) on the cosines of the angles between vectors in Fig. 3(*d*) identify the presence of an outlier data set.

410 data sets were collected in a single visit at a maximum throughput of 46 data sets per hour. The median time from the end of data collection to the completion of the associated processing job was 222.5 and 352 s for *xia2.multiplex* and *DIMPLE*, respectively. 98% of *DIMPLE* results were reported within 10 min of data collection finishing (see also Supplementary Fig. S1).

Figs. 6(*a*)–6(*c*) show the improvement in the merging statistics for the autoprocessed data on the addition of each new data set. There is a visible improvement in the quality of the *DIMPLE* electron-density map with the number of crystals (Figs. 6*d*–6*g*).

Analysis of the distribution of unit-cell parameters and clustering on unit-cell parameters indicated the presence of potential outlier data sets (Figs. 7*a* and 7*b*). Reprocessing with a lower unit-cell clustering threshold resulted in improved merging statistics for some data sets (Figs. 7*e* and 7*f*). Alternatively, $\Delta CC_{1/2}$ analysis may be useful in identifying outlier data sets. For ligand soak Z4439011520, $\Delta CC_{1/2}$ analysis by *dials.scale* identified two outlier data sets over two rounds of scaling and filtering (Figs. 7*c* and 7*d*). $\Delta CC_{1/2}$ filtering removed data sets 0 and 18, which were also the two least compatible data sets identified by unit-cell clustering, although only the latter was identified as an outlier according to the chosen unit-cell clustering threshold.

Using the data improved by the rejection of outlier data sets as above, initial structure solution was performed using *MOLREP* (Vagin & Teplyakov, 2010) with PDB entry 7aeh as the search model. Structures were refined for 200 cycles in *REFMAC*5 using rigid-body refinement, followed by iterative rounds of restrained refinement with automatic TLS and assisted model building in *Coot* (Emsley *et al.*, 2010). Final

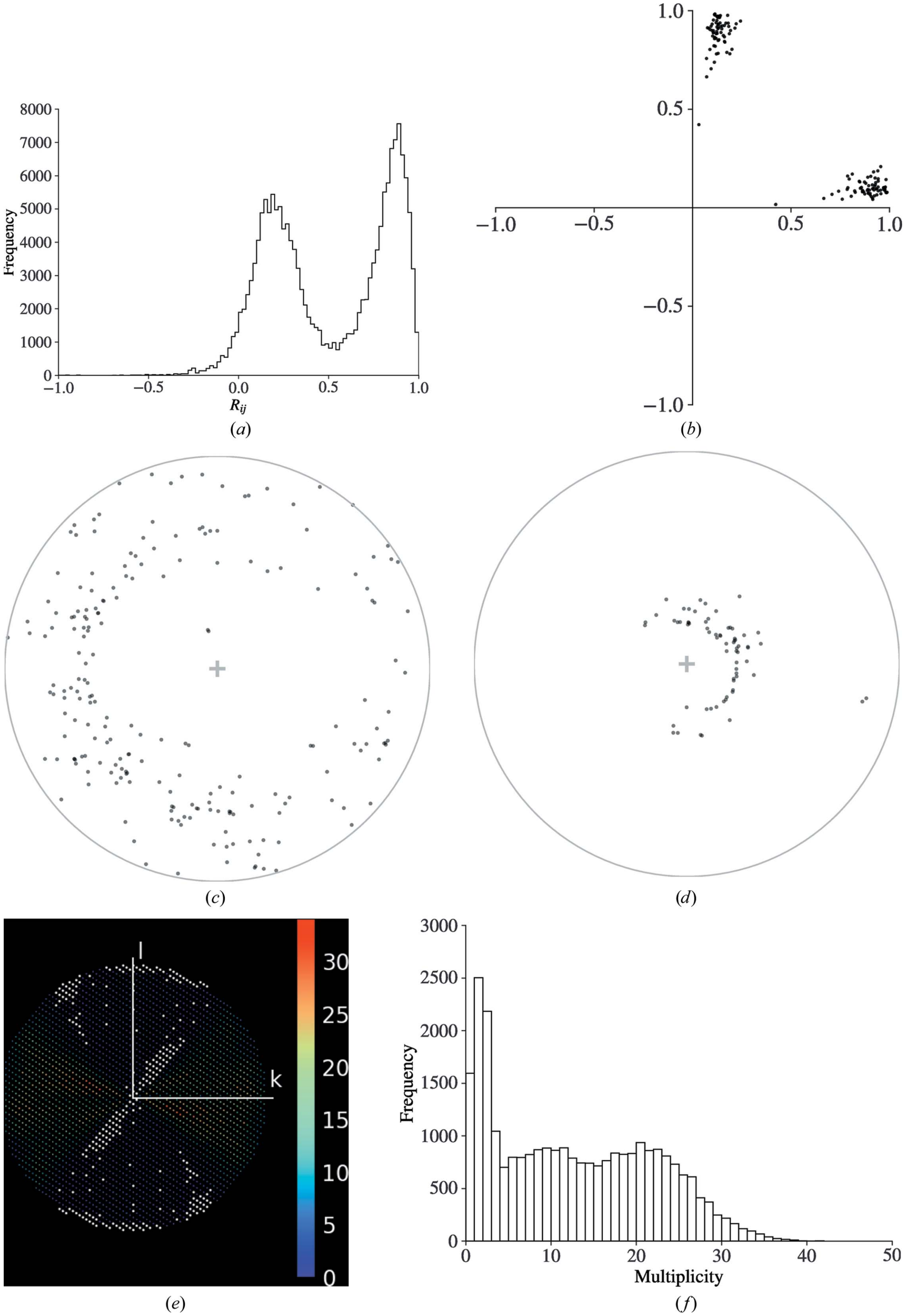


**Figure 5**
(*a*) A clear bimodal distribution of the histogram of pairwise $R_{ij}$ values is a strong indicator of the presence of an indexing ambiguity. (*b*) The vectors $\mathbf{x}$ determined by the minimization of equation (2) in *dials.cosym*. The separation of the vectors into two clusters indicates the presence of an indexing ambiguity. (*c*, *d*) Stereographic projections of crystal orientations for TehA crystals, representing the direction of *hkl* = 100 and *hkl* = 001 for each crystal, respectively, relative to the beam direction ($z$), which is shown as the central '+' into the page. A point close to the centre of the circle indicates that the crystal axis is close to parallel to the beam, whereas a point close to the edge of the unit circle indicates that the crystal axis is close to perpendicular to the beam. Preferential orientation can lead to regions with systematically low multiplicity or missing reflections. (*e*) shows the reflection multiplicities in the *0 kl* plane, where white corresponds to missing reflections. (*f*) The bivariate distribution of multiplicities is also indicative of an uneven distribution of multiplicities.

data-processing and refinement statistics for five ligand soaks, Z1367324110, Z31792168, Z4439011520, Z4439011584 and ABT-957, are reported in Table 4. Final coordinates and structure factors have been deposited in the Protein Data Bank (PDB entries 7qt6, 7qt5, 7qt7, 7qt9 and 7qt8, respectively) and raw data were uploaded to Zenodo (https://doi.org/10.5281/zenodo.5837942, https://doi.org/10.5281/zenodo.5837946, https://doi.org/10.5281/zenodo.5837903, https://doi.org/10.5281/zenodo.5836055 and https://doi.org/10.5281/zenodo.5837958).

Ligand soak ABT-957 is of particular interest as this unexpectedly crystallized in space group $P2_1$, in contrast to the space group $C2$ typical of this protein and indeed observed

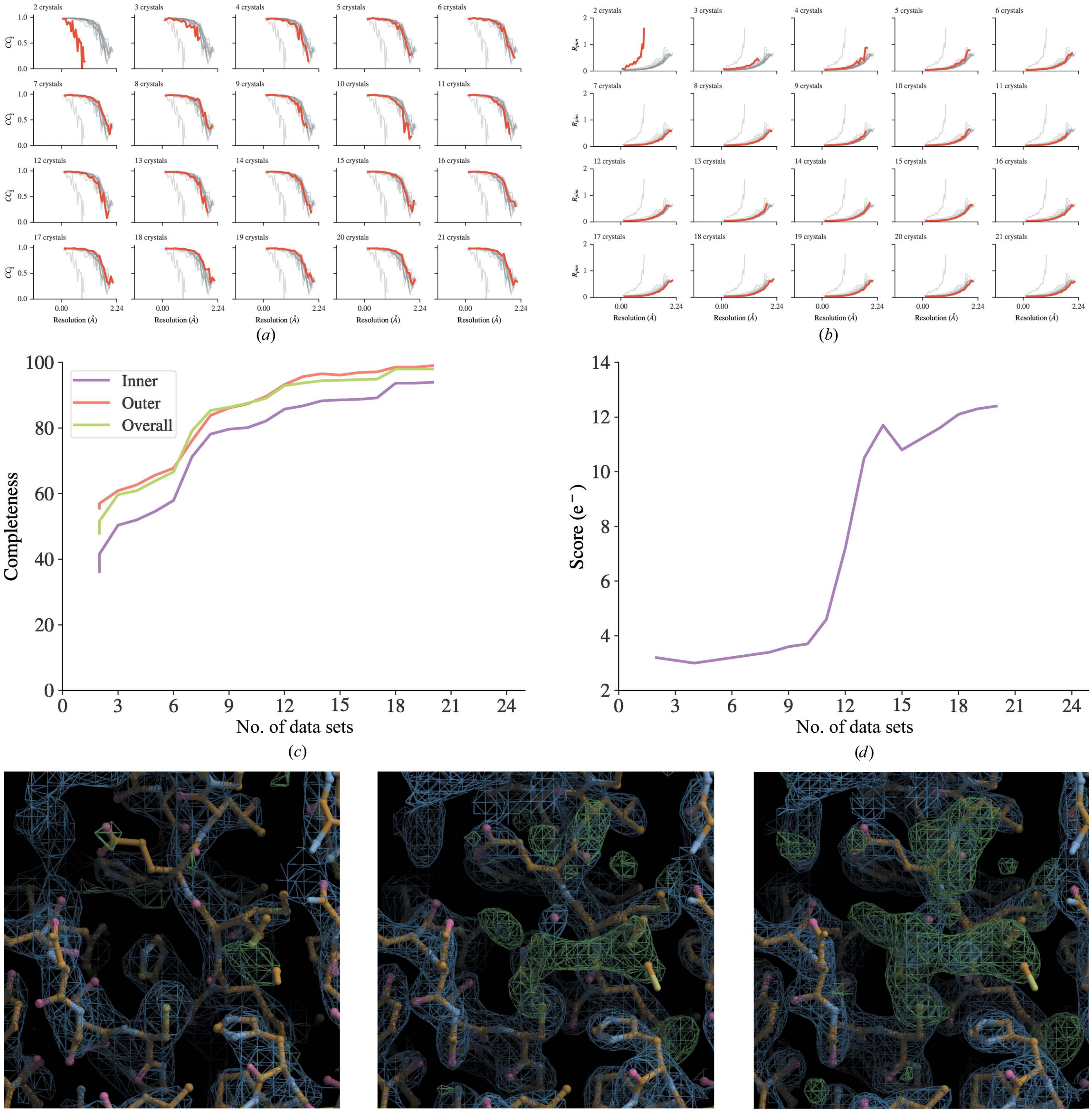


**Figure 6**
Incremental processing with *xia2.multiplex* and *DIMPLE* of *in situ* data collections of SARS-CoV-2 M$^{\rm pro}$ ligand soak Z4439011520. (*a*, *b*) $CC_{1/2}$ and $R_{\rm p.i.m.}$ data-processing statistics for ligand Z4439011520 with the inclusion of progressively more data sets in data-collection order from top left to bottom right. (*c*, *d*) Overall data completeness and *gemmi* (https://gemmi.readthedocs.io) blob search scores. (*e*, *f*, *g*) The ligand density in the autoprocessed *DIMPLE* maps for two, nine and 20 crystals, respectively. All contours are drawn at 3$\sigma$.

**Table 4**
Data-collection, merging and refinement statistics for SARS-CoV-2 M$^{pro}$ *in situ* data sets after filtering of outliers according to $\Delta CC_{1/2}$.

Values in parentheses are for the highest resolution shell.

| | Z1367324110 | Z31792168 | Z4439011520 | Z4439011584 | ABT-957 |
|---|---|---|---|---|---|
| Data collection | | | | | |
| Exposure time (s) | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| Ω width (°) | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Wavelength (Å) | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| Transmission (%) | 2.9 | 2.9 | 2.9 | 2.9 | 2.9 |
| No. of images | 200 | 200 | 200 | 200 | 200 |
| No. of data sets | 27 | 19 | 19 | 16 | 33 |
| Crystal parameters | | | | | |
| Space group | *C*2 | *C*2 | *C*2 | *C*2 | $P2_1$ |
| *a*, *b*, *c* (Å) | 115.21, 54.78, 45.34 | 114.77, 54.59, 45.31 | 115.69, 54.47, 45.25 | 115.86, 54.48, 45.20 | 45.23, 54.68, 116.54 |
| $\alpha$, $\beta$, $\gamma$ (°) | 90, 101.24, 90 | 90, 101.48, 90 | 90, 101.70, 90 | 90, 101.42, 90 | 90, 100.35, 90 |
| Data statistics | | | | | |
| Resolution range (Å) | 49.31–2.11 (2.19–2.11) | 44.42–2.26 (2.34–2.26) | 44.32–2.25 (2.33–2.25) | 56.80–2.43 (2.52–2.43) | 49.37–2.01 (2.08–2.01) |
| No. of unique reflections | 16050 (1586) | 12834 (1277) | 12607 (1272) | 10345 (1038) | 37112 (3748) |
| Multiplicity | 9.8 (9.9) | 7.0 (7.0) | 7.2 (7.3) | 5.9 (5.9) | 12.1 (12.2) |
| $R_{\rm merge}$ | 0.170 (2.429) | 0.170 (1.956) | 0.162 (1.538) | 0.166 (1.241) | 0.291 (2.409) |
| $R_{\rm meas}$ | 0.179 (2.551) | 0.184 (2.110) | 0.174 (1.652) | 0.183 (1.380) | 0.304 (2.511) |
| $R_{\rm p.i.m.}$ | 0.053 (0.755) | 0.067 (0.767) | 0.061 (0.577) | 0.073 (0.578) | 0.084 (0.691) |
| Completeness (%) | 99.7 (99.9) | 98.6 (99.6) | 95.2 (97.2) | 97.9 (97.7) | 98.8 (99.8) |
| $\langle I/\sigma(I)\rangle$ | 7.9 (0.6) | 9.9 (1.2) | 8.9 (1.1) | 10.8 (2.0) | 3.6 (0.4) |
| $CC_{1/2}$ | 0.996 (0.331) | 0.994 (0.425) | 0.987 (0.311) | 0.987 (0.305) | 0.992 (0.360) |
| Refinement | | | | | |
| $R_{\rm work}$ | 0.177 | 0.168 | 0.163 | 0.150 | 0.204 |
| $R_{\rm free}$ | 0.222 | 0.231 | 0.216 | 0.200 | 0.237 |
| R.m.s.d., bond lengths (Å) | 0.0133 | 0.0107 | 0.1070 | 0.0132 | 0.0123 |
| R.m.s.d., bond angles (°) | 1.843 | 1.691 | 1.810 | 1.903 | 1.752 |
| Average *B* factor (Å$^2$) | | | | | |
| Protein | 55.54 | 52.98 | 50.45 | 48.72 | 36.75 |
| Water | 47.13 | 46.09 | 46.44 | 41.28 | 29.50 |
| Ligand | 90.25 | 58.13 | 69.91 | 63.09 | 47.13 |
| Ramachandran statistics (%) | | | | | |
| Favoured | 96.69 | 96.04 | 97.35 | 96.36 | 97.03 |
| Allowed | 2.32 | 2.97 | 1.66 | 2.65 | 2.31 |
| PDB code | 7qt6 | 7qt5 | 7qt7 | 7qt9 | 7qt8 |

for the cryo-structure with this ligand (Redhead *et al.*, 2021). Autoprocessing (including both *xia*2 and *xia*2*.multiplex*) was performed both using the user-specified target space group, *C*2, and with automatic space-group determination. Out of 42 data sets collected, 18 data sets were successfully autoprocessed with *DIALS* via *xia*2 in the target space group *C*2 and combined with *xia*2*.multiplex*. In contrast, all 42 data sets individually processed successfully with automatic space-group determination in a mixture of space groups *P*1, *P*2, $P2_1$ and *C*2. 33 data sets remained after filtering for inconsistent unit cells. Analysis of symmetry with *dials.cosym* identified the Patterson group *P*2/*m*, which features an indexing ambiguity due to the approximate pseudo-symmetry of the supergroup *C*2 (Tables 5 and 6).

Of the ligand-soaked structures obtained, all showed a near-identical binding conformation in the cryogenic and room-temperature structures. A minor difference was observed in the conformation of ABT-957, with the C9—N—C1(*R*) amide bond in the room-temperature structure being flipped compared with the cryogenic structure (Fig. 8). This amide flip had a knock-on effect on the rotomer of the $\gamma$-lactam ring and the benzylic side chain which stems from N1 of the $\gamma$-lactam.

Inspection of a plot of $R_{\rm cp}$ versus image number (Supplementary Fig. S2) showed slight signs of radiation damage for some ligand soaks. Whilst limiting the number of images used from each data set may lead to improvements in some merging statistics (Supplementary Fig. S3), at the cost of completeness and multiplicity, this did not lead to any appreciable difference in the ligand density in the final structures (Supplementary Fig. S4).

**Table 5**
*dials.cosym* scores for individual symmetry elements for SARS-CoV-2 M$^{pro}$ ligand soak ABT-957.

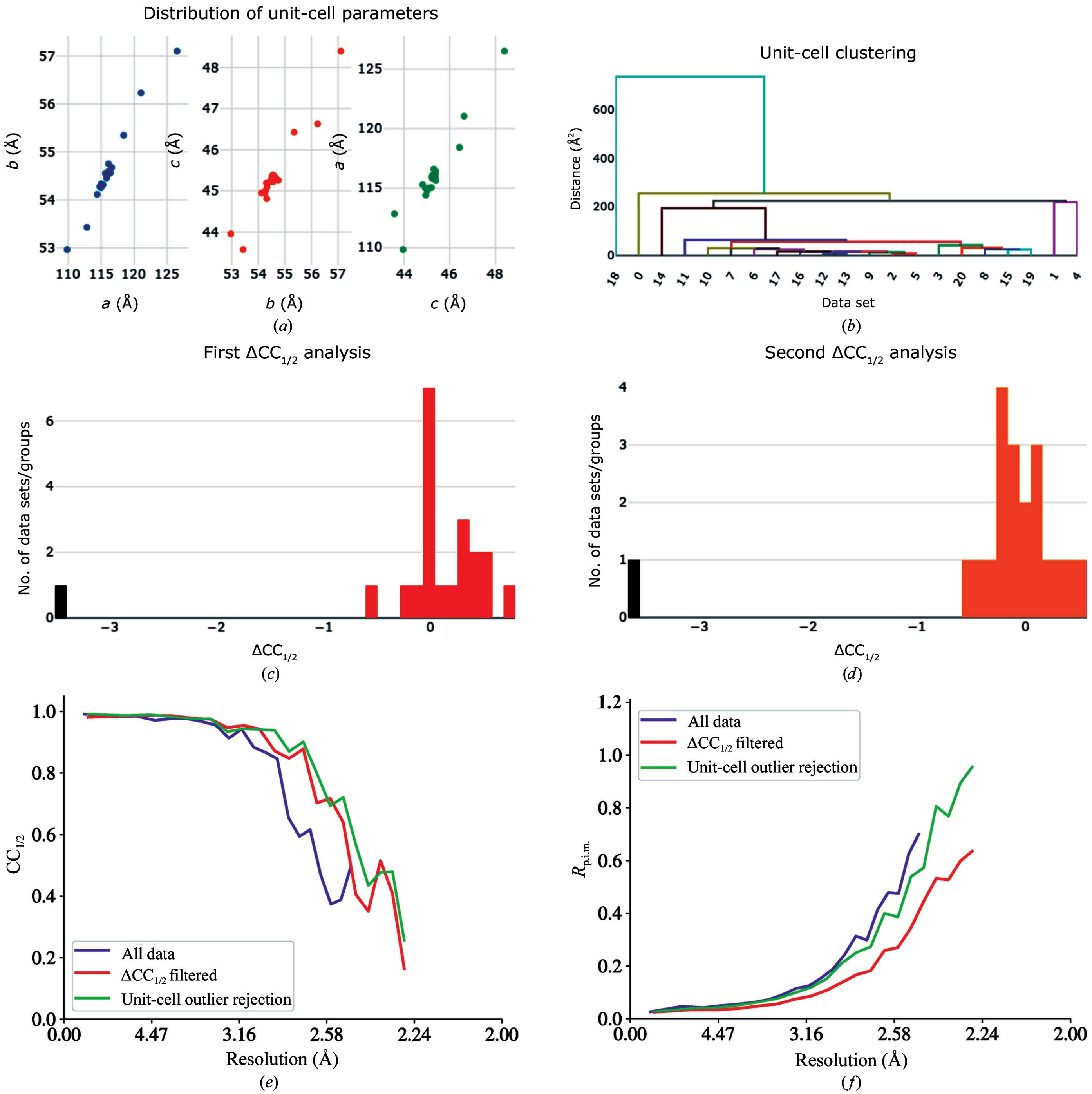
| Likelihood | Z-CC | CC | Symmetry element |
|---|---|---|---|
| 0.085 | 1.833 | 0.183 | 2\|(1, 0, 2) |
| 0.085 | 1.833 | 0.183 | 2\|(1, 0, 0) |
| 0.949 | 10.000 | 1.000 | 2\|(0, 1, 0) |

## 6. Conclusions

*xia*2*.multiplex* has been developed to perform symmetry analysis, scaling and merging of multiple data sets. It is distributed with *DIALS* and hence *CCP*4, and is available as part of the autoprocessing pipelines across the MX beamlines at Diamond Light Source, including integration with downstream phasing pipelines such as *DIMPLE* and *Big EP*. It is capable of providing near real-time feedback on data quality and completeness during ongoing multi-crystal data collec-

**Table 6**
*dials.cosym* subgroup scores for SARS-CoV-2 M$^{\rm pro}$ ligand soak ABT-957.

| Patterson group | Likelihood | NetZcc | Zcc+ | Zcc− | delta | Re-index operator |
|---|---|---|---|---|---|---|
| *P*2/*m* | 0.933 | 8.17 | 10.00 | 1.83 | 0.0 | *h*, *k*, *l* |
| $P\bar{1}$ | 0.050 | −5.96 | 0.00 | 5.96 | 0.0 | *h*, *k*, *l* |
| *Cmmm* | 0.008 | 5.96 | 5.96 | 0.00 | 0.9 | −*h*, *h* + 2*l*, *k* |
| *C*2/*m* | 0.005 | −5.36 | 1.83 | 7.19 | 0.9 | *h* + 2*l*, *h*, *k* |
| *C*2/*m* | 0.005 | −5.36 | 1.83 | 7.19 | 0.9 | −*h*, *h* + 2*l*, *k* |



(*a*) (*b*) (*c*) (*d*) (*e*) (*f*)

**Figure 7**
Outlier identification and removal for SARS-CoV-2 M$^{\rm pro}$ ligand soak Z4439011520. Visualization of (*a*) the distribution of unit-cell parameters and (*b*) clustering on unit-cell parameters may suggest possible outlier data sets. (*c*, *d*) $\Delta CC_{1/2}$ filtering with *dials.scale* can also remove data sets that strongly disagree with the majority of data sets. (*e*, *f*) Removing outlier data sets can improve the overall merging statistics.

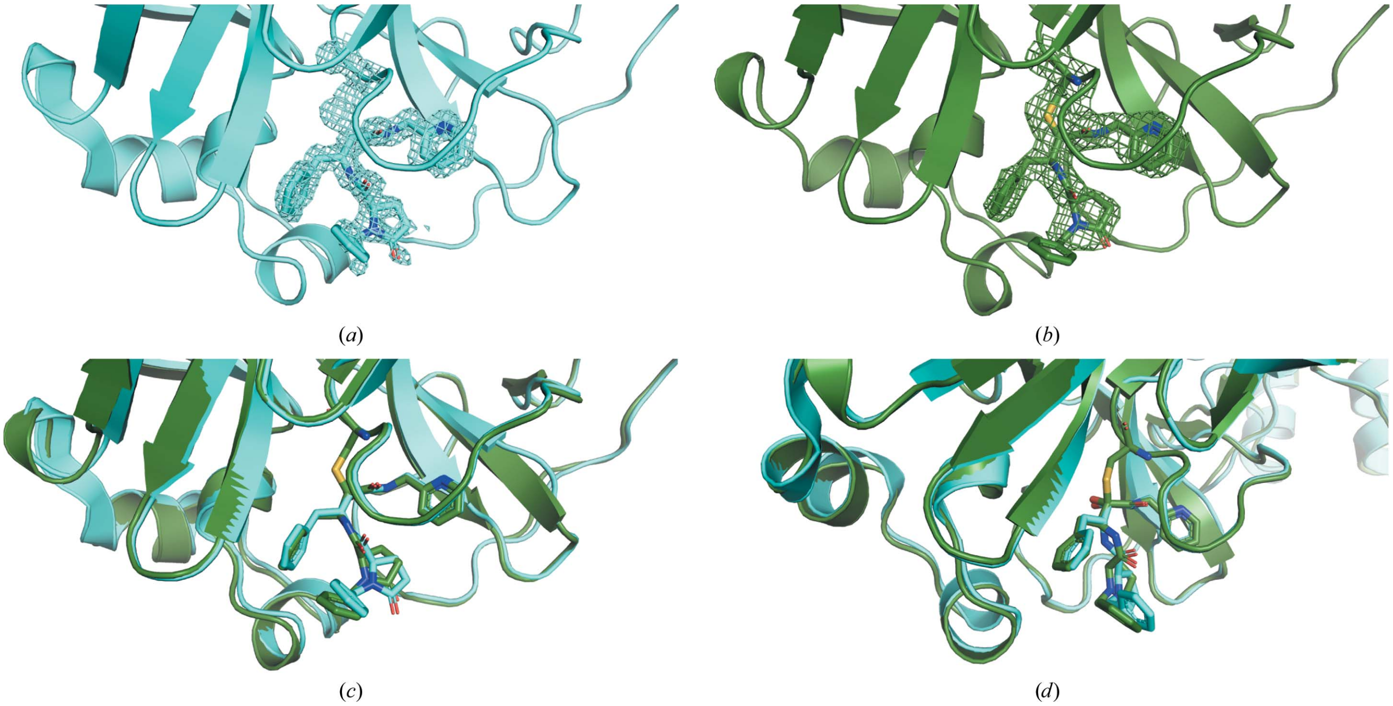


**Figure 8**
Views of the active site of SARS-CoV-2 M$^{pro}$ in complex with ABT-957 (*a*) under cryogenic conditions (Redhead *et al.*, 2021) and (*b*) at room temperature. Contours for the ligand density are drawn at 3$\sigma$. (*c*, *d*) Two slightly displaced views of the active site of SARS-CoV-2 M$^{pro}$ in complex with ABT-957 to show the conformational differences observed, particularly for the oxopyrrolidine and benzyl moieties of ABT-957 when bound to M$^{pro}$, at cryo temperature (cyan) and room temperature (green). The structures were superimposed using *PyMOL* (Schrödinger)

tions, and can be used as part of an iterative workflow to obtain the best possible final data set after an experiment.

We have demonstrated its applicability using two previously published room-temperature *in situ* multi-crystal data sets, including an example of experimental phasing. Using data sets collected as part of *in situ* room-temperature fragment-screening experiments on SARS-CoV-2 M$^{pro}$, we have shown the ability of *xia*2.*multiplex* to provide rapid feedback during multi-crystal experiments, including the identification of an unexpected change in space group on ligand addition.

Remaining challenges include the automatic identification of the best subset(s) of data to use for downstream analyses, and providing a user interface via applications such as SynchWeb or *CCP*4 to view results and facilitate an interactive workflow using *xia*2.*multiplex*. Support for MTZ files as input is planned in order to enable running *xia*2.*multiplex* on the output of other data-processing software such as *XDS* (Kabsch, 2010) and *MOSFLM* (Battye *et al.*, 2011).

## Acknowledgements

The authors would like to thank the *DIALS* development team for the various components that provide the foundations of *xia*2.*multiplex* and those within the wider Diamond Light Source software team who have assisted in the deployment of *xia*2.*multiplex*. We would also like to thank the Diamond XChem team for their assistance with the SARS-CoV-2 M$^{pro}$ ligand-screening experiments and all MX beamline staff at Diamond Light Source who have provided feedback on *xia*2.*multiplex* throughout its development. The authors acknowledge Diamond Light Source for the award of beam-time through the COVID-19 dedicated call (proposal IDs LB26986 and MX27088).

## Funding information

Funding for this research was provided by Diamond Light Source. Development of *DIALS* has been or is supported by Diamond Light Source, STFC via CCP4, Biostruct-X project No. 283570 of the EU FP7, the Wellcome Trust (grant No. 202933/Z/16/Z and 218270/Z/19/Z) and US National Institutes of Health grants GM095887 and GM117126.

## References

Akey, D. L., Brown, W. C., Konwerski, J. R., Ogata, C. M. & Smith, J. L. (2014). *Acta Cryst.* D**70**, 2719–2729.

Aller, P., Sanchez-Weatherby, J., Foadi, J., Winter, G., Lobley, C. M. C., Axford, D., Ashton, A. W., Bellini, D., Brandao-Neto, J., Culurgioni, S., Douangamath, A., Duman, R., Evans, G., Fisher, S., Flaig, R., Hall, D. R., Lukacik, P., Mazzorana, M., McAuley, K. E., Mykhaylyk, V., Owen, R. L., Paterson, N. G., Romano, P., Sandy, J., Sorensen, T., von Delft, F., Wagner, A., Warren, A., Williams, M., Stuart, D. I. & Walsh, M. A. (2015). *Methods Mol. Biol.* **1261**, 233–253.

Assmann, G., Brehm, W. & Diederichs, K. (2016). *J. Appl. Cryst.* **49**, 1021–1028.

Axford, D., Foadi, J., Hu, N.-J., Choudhury, H. G., Iwata, S., Beis, K., Evans, G. & Alguel, Y. (2015). *Acta Cryst.* D**71**, 1228–1237.

Axford, D., Owen, R. L., Aishima, J., Foadi, J., Morgan, A. W., Robinson, J. I., Nettleship, J. E., Owens, R. J., Moraes, I., Fry, E. E., Grimes, J. M., Harlos, K., Kotecha, A., Ren, J., Sutton, G., Walter, T. S., Stuart, D. I. & Evans, G. (2012). *Acta Cryst.* D**68**, 592–600.

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. (2011). *Acta Cryst.* D**67**, 271–281.

Beilsten-Edmands, J., Winter, G., Gildea, R., Parkhurst, J., Waterman, D. & Evans, G. (2020). *Acta Cryst.* D**76**, 385–399.

Brehm, W. & Diederichs, K. (2014). *Acta Cryst.* D**70**, 101–109.

Broecker, J., Morizumi, T., Ou, W.-L., Klingel, V., Kuo, A., Kissick, D. J., Ishchenko, A., Lee, M.-Y., Xu, S., Makarov, O., Cherezov, V., Ogata, C. M. & Ernst, O. P. (2018). *Nat. Protoc.* **13**, 260–292.

Bury, C. S., Brooks-Bartlett, J. C., Walsh, S. P. & Garman, E. F. (2018). *Protein Sci.* **27**, 217–228.

Caffrey, M. (2003). *J. Struct. Biol.* **142**, 108–132.

Caffrey, M. (2015). *Acta Cryst.* F**71**, 3–18.

Cherezov, V., Hanson, M. A., Griffith, M. T., Hilgart, M. C., Sanishvili, R., Nagarajan, V., Stepanov, S., Fischetti, R. F., Kuhn, P. & Stevens, R. C. (2009). *J. R. Soc. Interface.* **6**, s587.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., Choi, H.-J., Kuhn, P., Weis, W. I., Kobilka, B. K. & Stevens, R. C. (2007). *Science*, **318**, 1258–1265.

Clemons, W. M. Jr, Brodersen, D. E., McCutcheon, J. P., May, J. L., Carter, A. P., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. (2001). *J. Mol. Biol.* **310**, 827–843.

Collins, P. M., Ng, J. T., Talon, R., Nekrosiute, K., Krojer, T., Douangamath, A., Brandao-Neto, J., Wright, N., Pearce, N. M. & von Delft, F. (2017). *Acta Cryst.* D**73**, 246–255.

Cox, O. B., Krojer, T., Collins, P., Monteiro, O., Talon, R., Bradley, A., Fedorov, O., Amin, J., Marsden, B. D., Spencer, J., von Delft, F. & Brennan, P. E. (2016). *Chem. Sci.* **7**, 2322–2330.

Delagenière, S., Brenchereau, P., Launer, L., Ashton, A. W., Leal, R., Veyrier, S., Gabadinho, J., Gordon, E. J., Jones, S. D., Levik, K. E., McSweeney, S. M., Monaco, S., Nanao, M., Spruce, D., Svensson, O., Walsh, M. A. & Leonard, G. A. (2011). *Bioinformatics*, **27**, 3186–3192.

Diederichs, K. (2017). *Acta Cryst.* D**73**, 286–293.

Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O'Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F. & Walsh, M. A. (2020). *Nat. Commun.* **11**, 5047.

Durdagi, S., Dağ, Ç., Dogan, B., Yigin, M., Avsar, T., Buyukdag, C., Erol, I., Ertem, F. B., Calis, S., Yildirim, G., Orhan, M. D., Guven, O., Aksoydan, B., Destan, E., Sahin, K., Besler, S. O., Oktay, L., Shafiei, A., Tolu, I., Ayan, E., Yuksel, B., Peksen, A. B., Gocenler, O., Yucel, A. D., Can, O., Ozabrahamyan, S., Olkan, A., Erdemoglu, E., Aksit, F., Tanisali, G., Yefanov, O. M., Barty, A., Tolstikova, A., Ketawala, G. K., Botha, S., Dao, E. H., Hayes, B., Liang, M., Seaberg, M. H., Hunter, M. S., Batyuk, A., Mariani, V., Su, Z., Poitevin, F., Yoon, C. H., Kupitz, C., Sierra, R. G., Snell, E. H. & DeMirci, H. (2021). *Structure*, **29**, 1382–1396.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Evans, G., Axford, D., Waterman, D. & Owen, R. L. (2011). *Crystallogr. Rev.* **17**, 105–142.

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Evans, P. R. (2011). *Acta Cryst.* D**67**, 282–292.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.

Fischer, M., Shoichet, B. K. & Fraser, J. S. (2015). *ChemBioChem*, **16**, 1560–1564.

Fisher, S. J., Levik, K. E., Williams, M. A., Ashton, A. W. & McAuley, K. E. (2015). *J. Appl. Cryst.* **48**, 927–932.

Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Armour, W., Waterman, D. G., Iwata, S. & Evans, G. (2013). *Acta Cryst.* D**69**, 1617–1632.

Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. (2009). *Nature*, **462**, 669–673.

Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N. & Alber, T. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 16247–16252.

Garman, E. (1999). *Acta Cryst.* D**55**, 1641–1653.

Garman, E. & Owen, R. L. (2007). *Methods Mol. Biol.* **364**, 1–18.

Gildea, R. J. & Winter, G. (2018). *Acta Cryst.* D**74**, 405–410.

Giordano, R., Leal, R. M. F., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Cryst.* D**68**, 649–658.

Gotthard, G., Aumonier, S., De Sanctis, D., Leonard, G., von Stetten, D. & Royant, A. (2019). *IUCrJ*, **6**, 665–680.

Guven, O., Gul, M., Ayan, E., Johnson, J. A., Cakilkaya, B., Usta, G., Ertem, F. B., Tokay, N., Yuksel, B., Gocenler, O., Buyukdag, C., Botha, S., Ketawala, G., Su, Z., Hayes, B., Poitevin, F., Batyuk, A., Yoon, C. H., Kupitz, C., Durdagi, S., Sierra, R. G. & DeMirci, H. (2021). *Crystals*, **11**, 1579.

Helliwell, J. (1988). *J. Cryst. Growth*, **90**, 259–272.

Hirata, K., Yamashita, K., Ueno, G., Kawano, Y., Hasegawa, K., Kumasaka, T. & Yamamoto, M. (2019). *Acta Cryst.* D**75**, 138–150.

Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.

Huang, C.-Y., Olieric, V., Ma, P., Howe, N., Vogeley, L., Liu, X., Warshamanage, R., Weinert, T., Panepucci, E., Kobilka, B., Diederichs, K., Wang, M. & Caffrey, M. (2016). *Acta Cryst.* D**72**, 93–112.

Huang, C.-Y., Olieric, V., Ma, P., Panepucci, E., Diederichs, K., Wang, M. & Caffrey, M. (2015). *Acta Cryst.* D**71**, 1238–1256.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X., Yang, X., Bai, F., Liu, H., Liu, X., Guddat, L. W., Xu, W., Xiao, G., Qin, C., Shi, Z., Jiang, H., Rao, Z. & Yang, H. (2020). *Nature*, **582**, 289–293.

Jin, Z., Wang, H., Duan, Y. & Yang, H. (2021). *Biochem. Biophys. Res. Commun.* **538**, 63–71.

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

Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.

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

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). *Nature*, **185**, 422–427.

Krojer, T., Talon, R., Pearce, N., Collins, P., Douangamath, A., Brandao-Neto, J., Dias, A., Marsden, B. & von Delft, F. (2017). *Acta Cryst.* D**73**, 267–278.

Lawrence, J. M., Orlans, J., Evans, G., Orville, A. M., Foadi, J. & Aller, P. (2020). *Acta Cryst.* D**76**, 790–801.

Liu, D. C. & Nocedal, J. (1989). *Math. Program.* **45**, 503–528.

Liu, Q., Guo, Y., Chang, Y., Cai, Z., Assur, Z., Mancia, F., Greene, M. I. & Hendrickson, W. A. (2014). *Acta Cryst.* D**70**, 2544–2557.

Liu, Q. & Hendrickson, W. A. (2015). *Curr. Opin. Struct. Biol.* **34**, 99–107.

Liu, Q., Zhang, Z. & Hendrickson, W. A. (2011). *Acta Cryst.* D**67**, 45–59.

Maeki, M., Yamazaki, S., Pawate, A. S., Ishida, A., Tani, H., Yamashita, K., Sugishima, M., Watanabe, K., Tokeshi, M., Kenis, P. J. & Miyazaki, M. (2016). *CrystEngComm*, **18**, 7722–7727.

Monaco, S., Gordon, E., Bowler, M. W., Delagenière, S., Guijarro, M., Spruce, D., Svensson, O., McSweeney, S. M., McCarthy, A. A., Leonard, G. & Nanao, M. H. (2013). *J. Appl. Cryst.* **46**, 804–810.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

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

Olieric, V., Weinert, T., Finke, A. D., Anders, C., Li, D., Olieric, N., Borca, C. N., Steinmetz, M. O., Caffrey, M., Jinek, M. & Wang, M. (2016). *Acta Cryst.* D**72**, 421–429.

Owen, R. L., Axford, D., Nettleship, J. E., Owens, R. J., Robinson, J. I., Morgan, A. W., Doré, A. S., Lebon, G., Tate, C. G., Fry, E. E., Ren, J., Stuart, D. I. & Evans, G. (2012). *Acta Cryst.* D**68**, 810–818.

Owen, R. L., Paterson, N., Axford, D., Aishima, J., Schulze-Briese, C., Ren, J., Fry, E. E., Stuart, D. I. & Evans, G. (2014). *Acta Cryst.* D**70**, 1248–1256.

Rasmussen, S. G. F., Choi, H.-J., Fung, J. J., Pardon, E., Casarosa, P., Chae, P. S., DeVree, B. T., Rosenbaum, D. M., Thian, F. S., Kobilka, T. S., Schnapp, A., Konetzki, I., Sunahara, R. K., Gellman, S. H., Pautsch, A., Steyaert, J., Weis, W. I. & Kobilka, B. K. (2011). *Nature*, **469**, 175–180.

Redhead, M. A., Owen, C. D., Brewitz, L., Collette, A. H., Lukacik, P., Strain-Damerell, C., Robinson, S. W., Collins, P. M., Schäfer, P., Swindells, M., Radoux, C. J., Hopkins, I. N., Fearon, D., Douangamath, A., von Delft, F., Malla, T. R., Vangeel, L., Vercruysse, T., Thibaut, J., Leyssen, P., Nguyen, T., Hull, M., Tumber, A., Hallett, D. J., Schofield, C. J., Stuart, D. I., Hopkins, A. L. & Walsh, M. A. (2021). *Sci. Rep.* **11**, 13208.

Rosenbaum, D. M., Zhang, C., Lyons, J. A., Holl, R., Aragao, D., Arlow, D. H., Rasmussen, S. G. F., Choi, H.-J., DeVree, B. T., Sunahara, R. K., Chae, P. S., Gellman, S. H., Dror, R. O., Shaw, D. E., Weis, W. I., Caffrey, M., Gmeiner, P. & Kobilka, B. K. (2011). *Nature*, **469**, 236–240.

Sanchez-Weatherby, J., Sandy, J., Mikolajek, H., Lobley, C. M. C., Mazzorana, M., Kelly, J., Preece, G., Littlewood, R. & Sørensen, T. L.-M. (2019). *J. Synchrotron Rad.* **26**, 291–301.

Santoni, G., Zander, U., Mueller-Dieckmann, C., Leonard, G. & Popov, A. (2017). *J. Appl. Cryst.* **50**, 1844–1851.

Schubert, R., Kapis, S., Gicquel, Y., Bourenkov, G., Schneider, T. R., Heymann, M., Betzel, C. & Perbandt, M. (2016). *IUCrJ*, **3**, 393–401.

Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.

Sikharulidze, I., Winter, G. & Hall, D. R. (2016). *Acta Cryst.* A**72**, s193.

Smith, J. L., Fischetti, R. F. & Yamamoto, M. (2012). *Curr. Opin. Struct. Biol.* **22**, 602–612.

Southworth-Davies, R. J., Medina, M. A., Carmichael, I. & Garman, E. F. (2007). *Structure*, **15**, 1531–1541.

Thorn, A. & Sheldrick, G. M. (2011). *J. Appl. Cryst.* **44**, 1285–1287.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O. & Vázquez-Baeza, Y. (2020). *Nat. Methods*, **17**, 261–272.

Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.

Walsh, M. A., Grimes, J. M. & Stuart, D. I. (2021). *Biochem. Biophys. Res. Commun.* **538**, 40–46.

Warren, A. J., Armour, W., Axford, D., Basham, M., Connolley, T., Hall, D. R., Horrell, S., McAuley, K. E., Mykhaylyk, V., Wagner, A. & Evans, G. (2013). *Acta Cryst.* D**69**, 1252–1259.

Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.

Winter, G., Beilsten-Edmands, J., Devenish, N., Gerstel, M., Gildea, R. J., McDonagh, D., Pascal, E., Waterman, D. G., Williams, B. H. & Evans, G. (2022). *Protein Sci.* **31**, 232–250.

Winter, G., Gildea, R. J., Paterson, N., Beale, J., Gerstel, M., Axford, D., Vollmar, M., McAuley, K. E., Owen, R. L., Flaig, R., Ashton, A. W. & Hall, D. R. (2019). *Acta Cryst.* D**75**, 242–261.

Winter, G., Lobley, C. M. C. & Prince, S. M. (2013). *Acta Cryst.* D**69**, 1260–1273.

Winter, G. & McAuley, K. E. (2011). *Methods*, **55**, 81–93.

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.

Yamamoto, M., Hirata, K., Yamashita, K., Hasegawa, K., Ueno, G., Ago, H. & Kumasaka, T. (2017). *IUCrJ*, **4**, 529–539.

Yamashita, K., Hirata, K. & Yamamoto, M. (2018). *Acta Cryst.* D**74**, 441–449.

Zander, U., Bourenkov, G., Popov, A. N., de Sanctis, D., Svensson, O., McCarthy, A. A., Round, E., Gordeliy, V., Mueller-Dieckmann, C. & Leonard, G. A. (2015). *Acta Cryst.* D**71**, 2328–2343.

Zeldin, O. B., Brewster, A. S., Hattne, J., Uervirojnangkoorn, M., Lyubimov, A. Y., Zhou, Q., Zhao, M., Weis, W. I., Sauter, N. K. & Brunger, A. T. (2015). *Acta Cryst.* D**71**, 352–356.

Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013). *J. Appl. Cryst.* **46**, 1225–1230.

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