A comparative anatomy of protein crystals: lessons from the automatic processing of 56 000 samples

Olof Svensson, Maciej Gilski, Didier Nurizzo, Matthew W. Bowler

IUCrJ
ISSN 2052-2525
NEUTRON | SYNCHROTRON

A comparative anatomy of protein crystals: lessons from the automatic processing of 56 000 samples

Olof Svensson,[a] Maciej Gilski,[b] Didier Nurizzo[a] and Matthew W. Bowler[b]*

[a]European Synchrotron Radiation Facility, 71 Avenue des Martyrs, CS 40220, F-38043 Grenoble, France, and [b]European Molecular Biology Laboratory, Grenoble Outstation, 71 Avenue des Martyrs, CS 90181, F-38042 Grenoble, France. *Correspondence e-mail: mbowler@embl.fr

Received 6 March 2019
Accepted 4 June 2019

Edited by V. T. Forsyth, Institut Laue-Langevin, France, and Keele University, UK

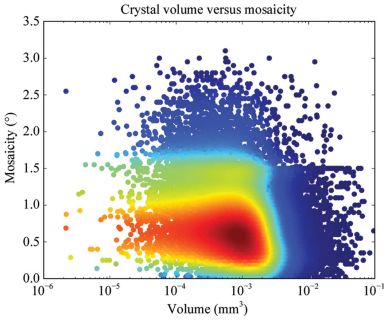
Keywords: MASSIF-1; data collection; biological macromolecules; crystal morphology.

Supporting information: this article has supporting information at www.iucrj.org

The fully automatic processing of crystals of macromolecules has presented a unique opportunity to gather information on the samples that is not usually recorded. This has proved invaluable in improving sample-location, characterization and data-collection algorithms. After operating for four years, MASSIF-1 has now processed over 56 000 samples, gathering information at each stage, from the volume of the crystal to the unit-cell dimensions, the space group, the quality of the data collected and the reasoning behind the decisions made in data collection. This provides an unprecedented opportunity to analyse these data together, providing a detailed landscape of macromolecular crystals, intimate details of their contents and, importantly, how the two are related. The data show that mosaic spread is unrelated to the size or shape of crystals and demonstrate experimentally that diffraction intensities scale in proportion to crystal volume and molecular weight. It is also shown that crystal volume scales inversely with molecular weight. The results set the scene for the development of X-ray crystallography in a changing environment for structural biology.

1. Introduction

Macromolecular crystallography (MX) has been the primary method for the determination of biological structures over the last 70 years. As such, much effort has been devoted to the development of methods to improve the ability to grow crystals, optimize their quality and collect the best possible data from them once they have been placed in an X-ray beam. The end results from these often long and tortuous experiments, structure factors and atomic coordinates, are deposited in one of the earliest examples of a searchable scientific open database: the Protein Data Bank (PDB; Berman *et al.*, 2000). Many studies have used this resource to draw conclusions on the properties of crystals, often with interesting conclusions (Abad-Zapatero, 2012; Berman *et al.*, 2013, 2015; Robert *et al.*, 2017). However, while the database is incredibly useful as a general repository of atomic structures for biologists, it has two fundamental limitations when attempting to draw conclusions on the properties of the crystals themselves. Firstly, the deposited data represent probably the best data that were obtained for a sample and were, as such, the result of extensive screening, thereby hiding the potentially thousands of crystals that stand behind the final structure. Secondly, experimental details, such as the size, shape and quality variation of the crystals, the data-collection strategy *etc.*, are often lost, even if recorded in the primary citation. These data are therefore difficult to collate.



Detailed studies have been made on individual systems, studying the morphology of the crystals and the packing of the protein, but a general survey across different proteins has never been made, presumably owing to the difficulty in gathering such information. More general studies have been made on the PDB itself, producing valuable results on the trends seen in protein crystals (Berman *et al.*, 2013, 2015) and also on their physical properties (Robert *et al.*, 2017; Bagaria *et al.*, 2013), most famously producing the Matthews coefficient (Matthews, 1968; Weichenberger & Rupp, 2014). However, both of these approaches lack a more general overview of how protein crystal morphology is distributed in general and how this is related to the macromolecule being studied. This is important as it has a direct effect on the requirements of the instrument used to study the crystal (Holton & Frankel, 2010).

The fully autonomous beamline MASSIF-1 at the ESRF (Bowler *et al.*, 2015) not only automates the process of sample handling (Nurizzo *et al.*, 2016), but also runs complex crystal-location, characterization and decision-making routines for every sample processed (Svensson *et al.*, 2015, 2018). This level of automation allows a wide range of projects to use the beamline, from those that require extensive screening to find the best diffracting crystal (Li *et al.*, 2018; Na *et al.*, 2017; Naschberger *et al.*, 2017; Sorigué *et al.*, 2017; Xu *et al.*, 2018) to small-molecule fragment screening (Cheeseman *et al.*, 2017; Hiruma *et al.*, 2017) and experimental phasing at high and low resolutions (Kharde *et al.*, 2015; Schulze *et al.*, 2018). The routines optimize data collection by centring crystals using X-ray diffraction quality to determine the location of the best volumes to centre (Svensson *et al.*, 2015) and measuring crystal volumes to dynamically adapt the beam diameter to match the crystal (Svensson *et al.*, 2018) and also to determine the dose that the sample can receive before sustaining significant radiation damage (Bowler *et al.*, 2016; Svensson *et al.*, 2015; Zeldin *et al.*, 2013; Bourenkov & Popov, 2010). Samples are then characterized, and optimized data sets are collected (Bourenkov & Popov, 2010; Incardona *et al.*, 2009), with subsequent autoprocessing of the data (Monaco *et al.*, 2013; Kabsch, 2010; Vonrhein *et al.*, 2011). All of the results from each step of these processes are stored with a unique identification for each sample (Brockhauser *et al.*, 2012; Delagenière *et al.*, 2011). Crucially, as the samples have been run without any human involvement, the reasons for the decisions that have been taken are known and the strategies have not been altered before the final data set is collected.

These data have not only allowed individual data-collection strategies to be improved, but have also improved general strategies by allowing, for example, the most commonly observed crystal dimension to determine the default beam diameter (Svensson *et al.*, 2015), improved low-resolution data-collection strategies or simply the correlation between the predicted and the obtained resolution to be assessed (Svensson *et al.*, 2018). While these data have proved invaluable in the improvement of the beamline, they also have inherent value in that they allow the first global survey of crystals of biological macromolecules. Here, we analyse the properties of the 56 459 samples sent to MASSIF-1 between September 2014 and December 2018. The results provide the first general overview of the morphology of crystals of biological macromolecules and how these properties relate to the macromolecule itself, and this is the first study of its kind in the history of macromolecular crystallography. Together, the results allow many long-held assumptions to be tested experimentally and provide a framework to direct the development of future beamline facilities.

Table 1
The fate of crystals sent to MASSIF-1.

Stage achieved	No. of samples	Percentage of total
Samples received at MASSIF-1	56459	100
Samples successfully centred	33905	60
Samples indexed and a strategy calculated	14495	26
Samples with a processed data set	16034	28

2. Methods

2.1. The crystal cohort

MASSIF-1 started taking user samples in September 2014 and has been gathering data on all aspects of these crystals since then. To date, the beamline has processed 56 459 samples from 1306 declared projects from laboratories across Europe and the world (Bowler *et al.*, 2016). We believe that this number and distribution of samples represents a reasonable snapshot of crystals for modern structural biology projects. This cohort can therefore form the basis of an analysis that we hope will be generally applicable. The fate of these samples is shown in Table 1. Of the samples received, only 60% were successfully centred; these 33 905 crystals form the basis of the study presented here. The remaining 40% either diffracted too weakly, were salt crystals or the sample mount was empty, preventing further analysis. The number itself demonstrates the need for extensive screening in MX, with only 28% of samples yielding a data set. This average covers a wide variety of projects, from fragment-screening projects with data-set yields close to 100% to more extreme cases, for example membrane proteins, with yields of only 1–2% (Bowler *et al.*, 2016; Svensson *et al.*, 2018), but it provides a good idea of the general attrition rate in MX. All data used in this study (for the 33 905 centred crystals) have been anonymized and are available to download from the ESRF data portal (https://doi.esrf.fr/10.15151/ESRF-DC-186715792; Svensson *et al.*, 2019).

2.2. Databases and analysis

All of the information gathered and used in the automatic location, characterization and data collection from crystals processed on MASSIF-1 is stored in two databases: one is related to the sample location, positioning and characterization processes (Brockhauser *et al.*, 2012; Svensson *et al.*, 2015; BES-DB, Support Square; https://supportsquare.io/products/), while the second, ISPyB (Delagenière *et al.*, 2011), records the results of characterization and data processing. The

information for samples contained in each database can be correlated using a data-collection ID that is unique to each sample. This allowed us to reconcile the data for samples between databases. A simple Python GUI was developed to access data from both databases and store relevant parameters for each sample in JSON format. Analysis of these data was performed using *SciPy* and *matplotlib* (Hunter, 2007; Oliphant, 2007).

Crystal dimensions are measured from the X-ray centring routine. The dimensions x, y and z are the measured crystal width parallel to the spindle axis, the height orthogonal to the spindle axis and the depth orthogonal to the spindle axis 90° away in ω, respectively. The full widths at half maximum (FWHMs) of diffraction signal over images are used to determine crystal dimensions. Oversampling means that the minimum distance that can be measured using the 50 µm diameter beam is 25 µm; dimensions smaller than this are determined using smaller beam apertures. As the automesh algorithm (Svensson *et al.*, 2015) will place the sample mount at either the smallest or widest orientation of the mount in ω, depending on whether single or multiple data collections are requested (Svensson *et al.*, 2018), we are confident that in most cases the dimensions measured will be consistent with the orientation of the crystals, as they tend to lie parallel to the mount. This would reduce the overestimation of sample height and depth, for instance if a plate-shaped crystal was presented at an angle. All samples are assumed to be cuboid.

Unit-cell volumes were calculated from the dimensions obtained during indexing using the equation

$$V = abc(1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma)^{1/2}. \quad (1)$$

The molecular weight (kDa) of the entity in the asymmetric unit was estimated using the equation

$$\mathrm{MW} = \frac{(V_{\mathrm{cell}} \times 0.47)/n_{\mathrm{symops}}}{V_{\mathrm{p}}}, \quad (2)$$

where a solvent content of 47% was assumed (the average in the PDB), n_{symops} is the number of symmetry operators determined from the point group and V_{p} is derived from the partial specific volume for globular proteins of 0.73 $\mathrm{cm^3\ g^{-1}}$ (Harpaz *et al.*, 1994), here expressed as 1210 $\mathrm{Å^3\ kDa^{-1}}$. These assumptions are reasonable, but will lead to some overestimates of molecular weight for smaller proteins and an underestimation for larger proteins, where the solvent contents may be significantly different from 47%. The molecular weight will also be inaccurate if the crystal is incorrectly indexed, particularly if the triclinic point group is incorrectly selected. The process was verified by comparing seven proteins with known molecular weights to the values calculated using these methods (see Supplementary Table S1) and the average was the same as that for the PDB (Berman *et al.*, 2013). A histogram of the molecular-weight distribution is shown in Supplementary Fig. S1.

Table 2
Volume guide.

Volume in $\mathrm{mm^3}$	Volume in $\mu\mathrm{m^3}$	Volume in pl or nl	Equivalent cube dimension (µm)
10^{-6}	1000	1 pl	Cube with 10 µm edge
10^{-5}	10000	10 pl	Cube with 21.5 µm edge
10^{-4}	100000	100 pl	Cube with 46.4 µm edge
10^{-3}	1000000	1 nl	Cube with 100 µm edge
10^{-2}	10000000	10 nl	Cube with 215 µm edge
10^{-1}	100000000	100 nl	Cube with 464 µm edge

3. Results and discussion

3.1. The size and shape of protein crystals

Accurately determining the dimensions, and therefore the volume, of a protein crystal is primarily important in determining the dose that the crystal can absorb before significant radiation damage (Bourenkov & Popov, 2010; Bowler *et al.*, 2016; Svensson *et al.*, 2015; Zeldin *et al.*, 2013) and in determining the diameter of the beam that should be used to maximize the signal-to-noise ratio (Evans *et al.*, 2011; Svensson *et al.*, 2018). However, the gathering of volumetric data has many other potential uses, not least in being able to correlate crystal size with quality for individual projects. The data collected for all samples processed on MASSIF-1 provide an opportunity, for the first time, to define the broad distribution of protein crystal dimensions. Here, we show volumes in cubic millimetres. These can be difficult to convert into real-world quantities, so a comparison between units is shown in Table 2.

To our knowledge, there has never been a general study of protein crystal volumes. While there are some careful studies of individual proteins (Frey *et al.*, 1991; Joachim & Markus, 2015; Liu *et al.*, 2013; Mayans & Wilmanns, 1999), we have no data for the general distribution. Programs that account for crystal volume when computing absorbed doses, such as *BEST*

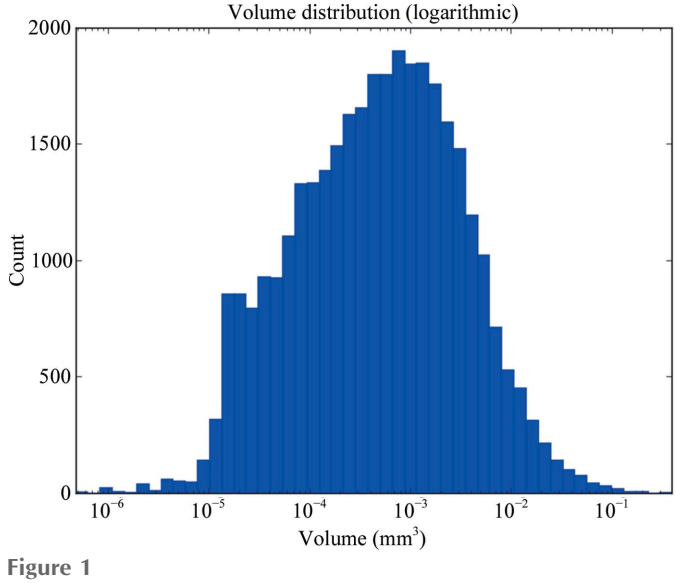


Figure 1
Histogram of crystal volumes. The histogram shows the distribution of crystal volumes measured on MASSIF-1. Note the logarithmic scale. N = 33 905.

(Bourenkov & Popov, 2010; Popov & Bourenkov, 2003) and *RADDOSE3D* (Zeldin *et al.*, 2013), are extremely useful when users input the correct crystal dimensions. However, a default volume, a cube with sides of 100 µm, is defined in the programs and is used in the absence of measured dimensions. This value originated in the early days of MX (Helliwell, 1984), but does it relate to the reality of protein crystals today?

The distribution of measured crystal volumes is shown in Fig. 1. The mean volume of 0.002303221 mm^3 (2 303 221 µm^3) is the same volume as a cube with edges of 132 µm. While the majority of crystals are smaller than this average, the distribution is lognormal and the mode volume is 0.000020209 mm^3 (20 209 µm^3; a cube with edges of 27 µm), it does seem to validate the choice of the default average crystal volume used. However, volumetric data alone hide an important factor: morphology. The best way to demonstrate the relationship between a shape and its volume is the surface area-to-volume ratio. Here, we have plotted the surface area against the volume (Fig. 2). The plot shows that most crystals have a surface area greater than that expected for a cube, with crystals with very large volumes being more cuboid. What is important is that many crystals that have a large volume (*e.g.* 0.01 mm^3) have shapes that are best matched by a cylinder of 40 µm in diameter or a plate of thickness 50 µm (Fig. 2, magnified panel). This reflects the distribution of measured dimensions, which have a modal value of around 50 µm (Bowler *et al.*, 2016; Svensson *et al.*, 2015). This implies that using X-ray beams of larger than 100 µm will have limited returns and most crystals will require diameters of 10–50 µm with, of course, the possibility of collecting data from multiple volumes in plate-shaped or needle-shaped crystals.

3.2. The internal properties

The recording of volumetric data along with the results from data collection now allows us to test certain maxims within the MX community. It is generally accepted that larger crystals will be more difficult to cryocool, for example. The mosaic spread of a crystal has been demonstrated to be closely correlated with the effectiveness of the cooling protocol employed (Mitchell & Garman, 1994; Kriminski *et al.*, 2002), and we have used this measure to relate to volumetric data. The mosaic spread value used here is the *MOSFLM*-estimated value from four characterization images (Leslie, 2006) and is used as it is the only value that is calculated in the same manner for all samples. This value can be higher than that calculated by, for example, *XDS* (Powell *et al.*, 2017; Kabsch, 2010), but will be consistent and will allow trends to be discerned. Plotting crystal volume against mosaic spread [Fig. 3(*a*)] does not show a correlation; in fact, there is a weak negative correlation (Spearman $R = -0.26$). As the cooling rate is an important factor in cryocooling (Garman, 1999; Teng & Moffat, 1998) perhaps the surface-area-to-volume ratio (S/V) is more important? Crystals with a higher S/V should cool faster. Plotting S/V against mosaic spread [Fig. 3(*b*)] again does not show that crystals that could potentially cool more rapidly have lower mosaic spread values, and again there is a weak opposite correlation (Spearman $R = 0.23$). If the mosaic spread is independent of the crystal shape and size, is it more closely related to the entity crystallized? Plotting the molecular weight against mosaic spread [Fig. 3(*c*)] does seem to point to a trend to higher mosaic spread values for larger macromolecules, but again the correlation is weak (Spearman $R = 0.2$). Is then the order of the crystal

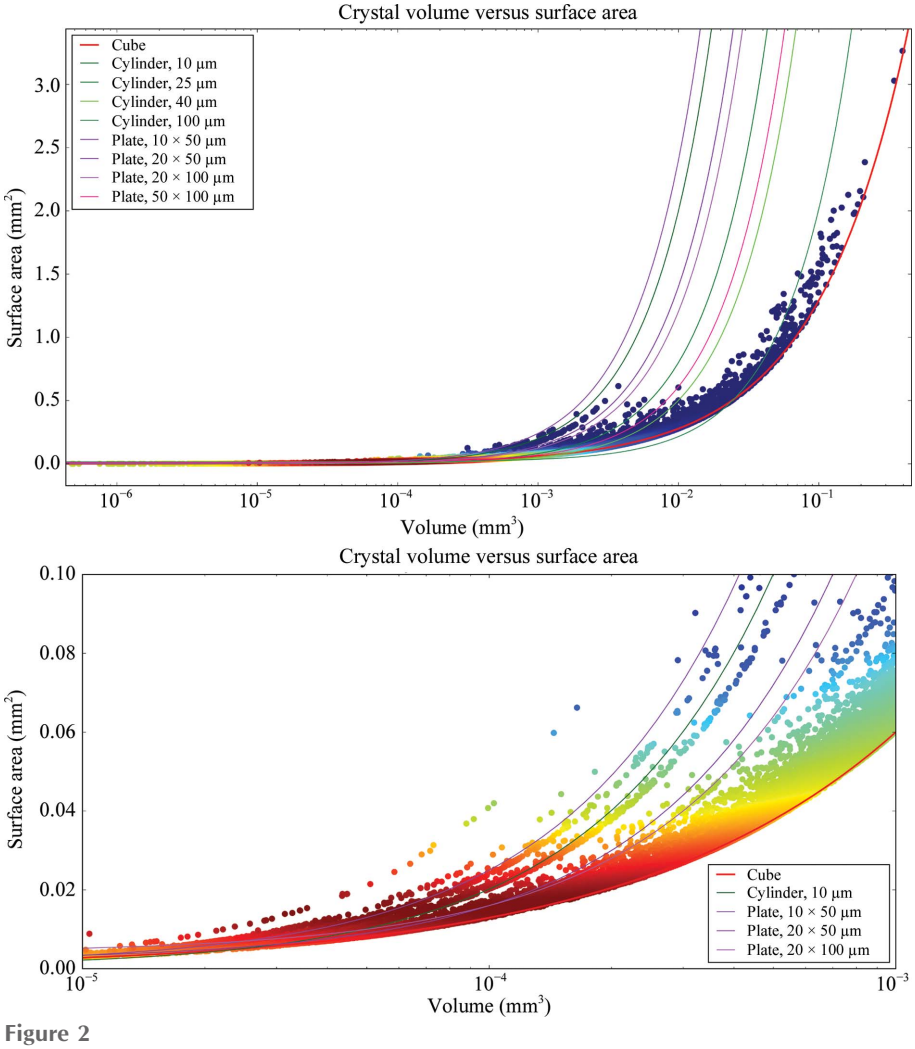


Figure 2
Crystal surface-area-to-volume ratio. The relationship between surface area and volume is shown with lines representing the curves described by different shapes and points coloured by the kernel density estimate (KDE). $N = 33\,905$. The lower panel shows a magnified area with the highest counts. Most crystals are small and have one dimension that is in the 10–30 µm range.

more dependent on the entity crystallized? While the correlation shown here is weak, small-molecule crystallographers have observed that smaller crystals have higher mosaic spread values (Andrews *et al.*, 1987, 1988; Papiz *et al.*, 1990), proposing that crystal growth could be limited by the disorder in the crystal. While the variation in mosaic spread values that we have measured here is most likely to be attributable to the mounting of crystals for data collection, it should not be ruled out that mosaic spread could be an inherent property of the crystal as grown. This seems to be supported by the lower mosaic spread values for more cuboid crystals [Fig. 3(*b*)], implying that well ordered growth in all lattice directions is a better predictor of lower mosaic spread. Several studies have shown that mosaic spread can be high at room temperature and can be reduced via controlled dehydration (Bowler *et al.*, 2006; Sanchez-Weatherby *et al.*, 2009; Russi *et al.*, 2011; Amunts *et al.*, 2007; Kiefersauer *et al.*, 2000), indicating that mosaic spread can already be high before cryocooling. This seems to counter the received wisdom that protein crystals tend to have lower mosaic spread values at room temperature (Garman, 1999), but only a few systematic studies have been made (Fischer *et al.*, 2015; Low *et al.*, 1966; Juers & Matthews, 2001) and further studies will be required.

What practical implications does this have? It seems to be clear that crystals should not be selected based on their size and shape. It is therefore probably more important to focus on careful crystal handling to minimize mosaic spread through the choice of cryoprotectant, soaking protocol and speed of cooling (Garman, 1999; Warkentin *et al.*, 2006). It is worth spending time obtaining lower mosaic spreads; when plotted against resolution there is a good correlation [Spearman R = 0.44; Fig. 3(*d*)].

Another important parameter is the variation of diffraction quality within a crystal. It has been shown that the different crystal volumes can vary widely, leading to significantly better data sets from the more ordered regions (Bowler *et al.*, 2010; Thompson *et al.*, 2018), but how common is it for crystals to diffract heterogeneously? Previous work has defined a

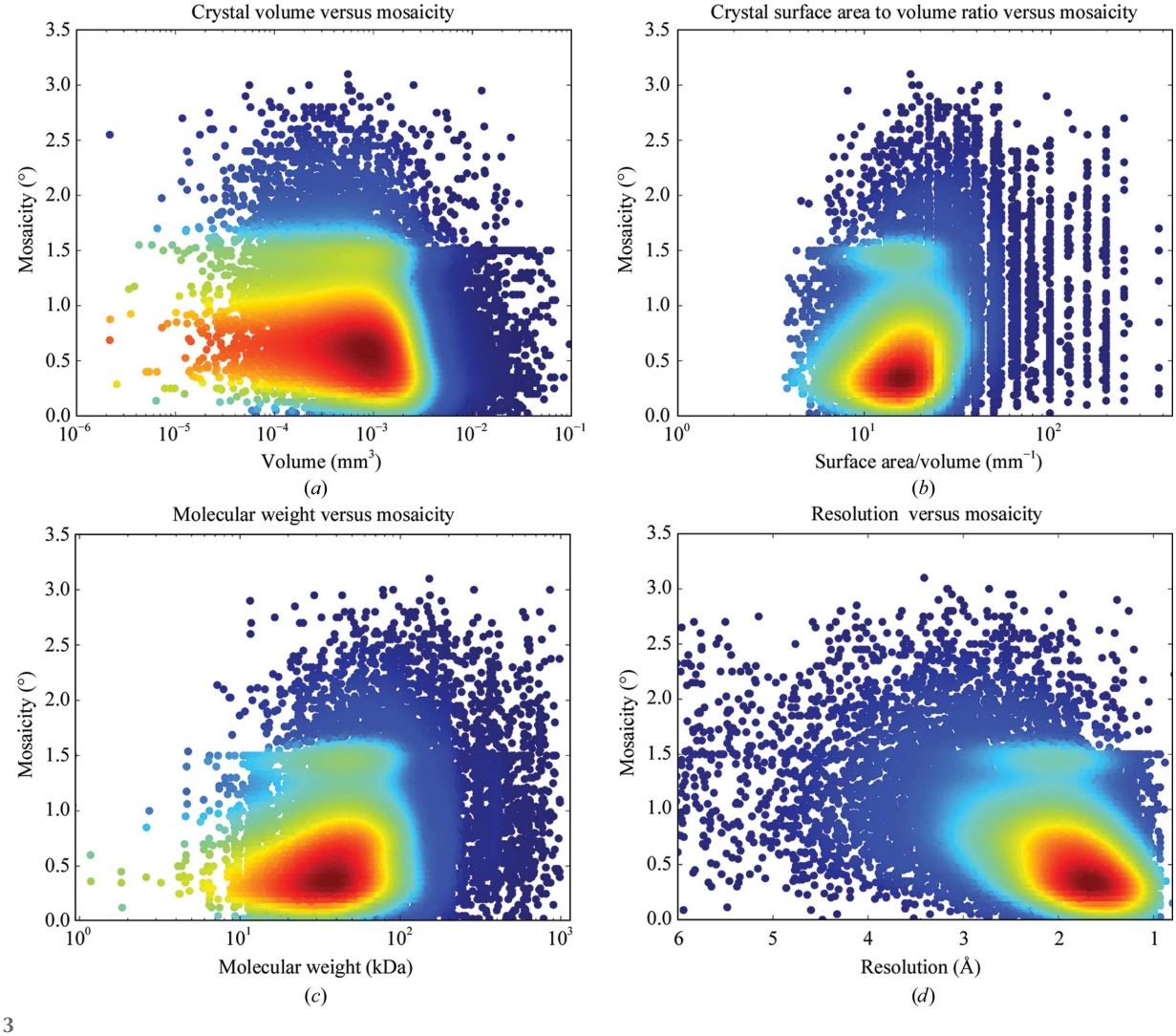


Figure 3
What parameters correlate with mosaic spread? Mosaic spread values are plotted against crystal volume, Spearman R = −0.26 (*a*), surface-area-to-volume ratio, Spearman R = 0.23 (*b*), molecular weight of the entity crystallized, Spearman R = 0.2 (*c*), and the resolution cutoff of the processed data set, Spearman R = 0.44 (*d*). Data points are coloured by KDE; N = 14 234 for each panel.

measure of diffraction variability within crystals that was demonstrated on 19 test samples (Bowler & Bowler, 2014). The measures, V_1 and V_2, define variability as the variance in diffraction quality over the mean squared and the peak value over the mean, respectively. A simple model defines the ratio N, giving an idea of the proportion of the crystal that varies and how great the difference in diffraction quality is. At the time of the initial study, the total integrated signal of the images collected during a mesh scan was used to define the diffraction quality. Since MASSIF-1 started, the measure is the *Dozor* score (Svensson *et al.*, 2015; Melnikov *et al.*, 2018). *Dozor* determines the distribution of background intensity, azimuthally averages the spot intensities and removes areas showing ice or salt diffraction. The mean intensity of Bragg spots against resolution over background is then determined and used to create a score of quality. A plot of V_1 against V_2 with various ratios N, from all mesh scans performed on MASSIF-1 using the *Dozor* score as a metric, is shown in Fig. 4(*a*). From the plot it can be seen that most crystals are quite homogenous, displaying ratios N of below 5. However, there are a large number of observations where the diffraction quality varies enormously, with peaks 4–6 times above the average (it should be noted that the *Dozor* score only varies between 10 and 20% within images of a data set from a single position). Several lines can be seen defined by data points: these describe lines of N = 0.5, 1 and 2 and arise from small crystals that have been probed at only two or three positions. Two positions can only give N = 1.0, and either N = 0.5 or 2.0 for three positions. How does the variability relate to other characteristics? When compared with mosaic spread there is no correlation [Fig. 4(*b*)], nor is there any correlation with the molecular weight of the entity crystallized [Fig. 4(*c*)]. When compared with the resolution of the final data set there is a weak correlation for higher resolution for a higher ratio N [Fig. 4(*d*)]. As the final data set is collected from a single position, with the beam diameter adapted to the best region, it

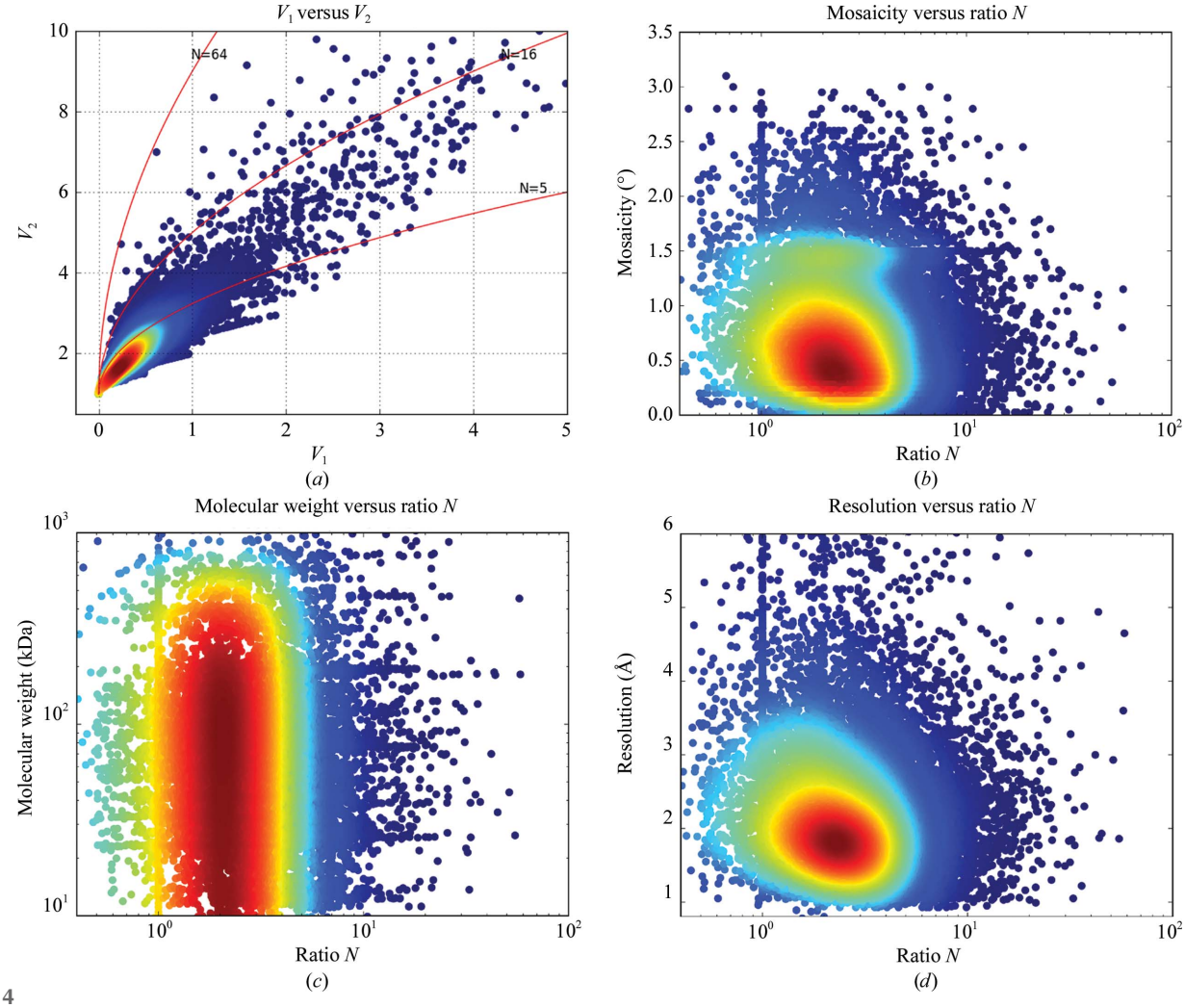


Figure 4
The variation of quality within crystals. (*a*) Comparison of variability measures of crystals. Values of V_1 and V_2 are plotted against each other and coloured by KDE. Lines show the values obtained for various ratios N between positions at increasing differences in diffraction power (for an explanation of the model, see Bowler & Bowler, 2014). The red line representing the ratio N = 5 is a reasonable cutoff between variable and homogenous diffraction within crystals. N = 15 864. (*b*) Mosaic spread values plotted against ratios N. Greater variability within a crystal is not related to higher mosaic spread values, Spearman R = −0.13, N = 13 780. (*c*) The molecular weight has no effect on the degree of variability, Spearman R = −0.06, N = 15 188. (*d*) Obtained resolution against ratios N. Higher variability is weakly correlated to higher resolution, Spearman R = −1.9, N = 15 188.

is perhaps not entirely surprising that good data can still be collected from a region of a variable crystal and it is clear that heterogeneous quality does not prevent the collection of a good-quality data set if the correct strategy is employed. Given the significant variation in quality observed, the ultimate strategy for scanning would be to use the smallest possible beam to probe variation and then adapt the diameter to match the size of the best volume determined.

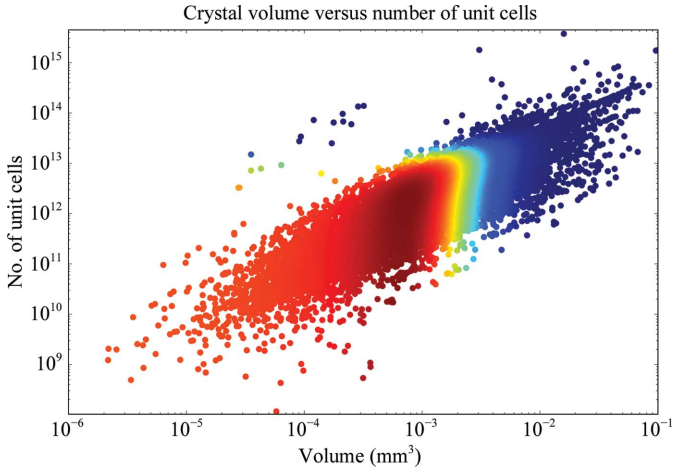


Figure 5
Number of unit cells plotted against crystal volume. There is a large spread across three orders of magnitude in the number of unit cells in the highest count bins. The smallest number of unit cells in a crystal that yielded a data set was 2.3×10^8. Data points are coloured by KDE. $N = 15\,905$.

3.3. Relationship between crystal volume, molecular weight of the protein and resolution

The scattering power of a crystal depends on the number of unit cells that can be illuminated in the beam, meaning that the volume of both the crystal and of the unit cell, as well as the properties of the molecule, are critical to a successful experiment (Holton, 2009). Practically, this means that the larger the molecule being studied, the larger its *B* factor and the higher the required resolution, the larger the crystal will have to be. How does this relate to the actual measured values of crystals that yielded data sets on MASSIF-1? Firstly, the number of unit cells can be plotted against the crystal volume (Fig. 5). While not surprising, it is informative to see how the number ranges across projects that make it through to indexing: 10^{12} unit cells is the most common (mode) and the smallest number that led to a processed data set was 2.3×10^8 unit cells from a crystal of $10 \times 12 \times 20$ µm in size with 15 µm beam diameter, space group *P*6 and unit-cell parameters $a = b = 85.94$, $c = 201.45$ Å, $\alpha = \beta = 90$, $\gamma = 120^\circ$. A thorough theoretical treatment for this relationship has been demonstrated (Holton & Frankel, 2010; Holton, 2009), defining the minimum crystal volume required under ideal conditions. This theoretical relationship has been extremely useful for predicting the requirements and limits of new facilities and for providing a target for beamlines to aim for when optimizing experiments (Grimes *et al.*, 2018). Does the relationship hold experimentally over a large number of different samples? Fig. 6 shows the resolution obtained plotted against crystal volume coloured by the molecular weight of the molecule crystallized. The theoretical relationship for the minimum required crystal volume is also plotted. The lines describe equation (16) from Holton & Frankel (2010) assuming a Nave–Hill effect (photoelectron escape) of 1, as all crystals are larger than 1 µm^3, and 100 photons per *hkl* that approximates the experimental conditions on MASSIF-1. The agreement between the theoretical curves and observed samples is remarkable (Fig. 6), with most crystals remaining above the minimum volume predicted for a given resolution and molecular weight. The curves represent specific molecular weights and *B* factors, which can vary enormously, and should be taken as a guide to where these values

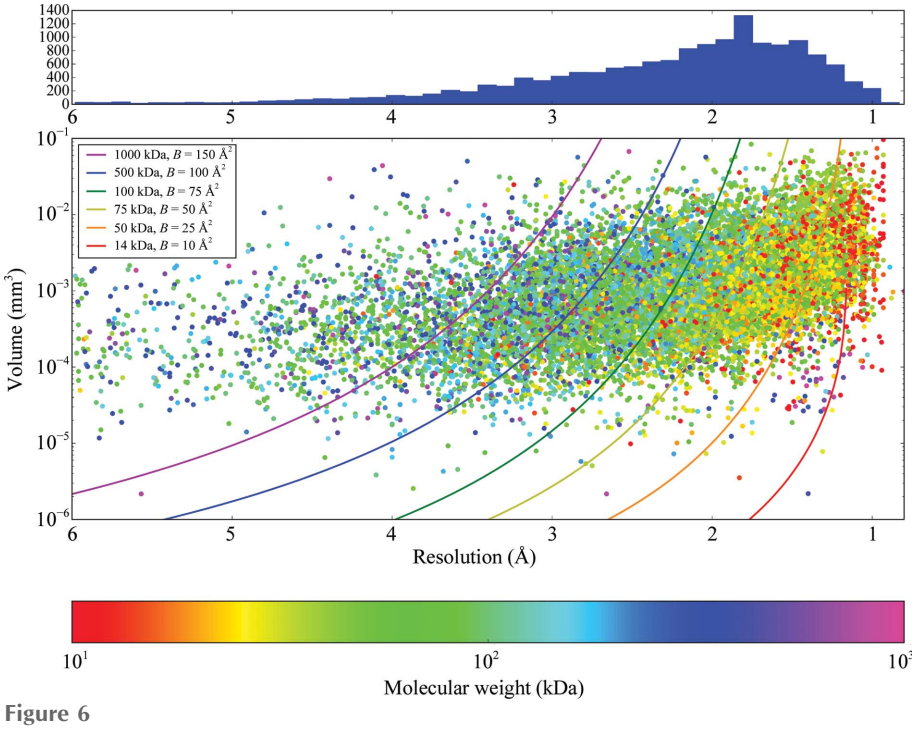


Figure 6
The relationship between crystal volume, obtained resolution and molecular weight. The crystal volume is plotted against the highest resolution cutoff from autoprocessing and is coloured by the molecular weight of the entity crystallized; note the logarithmic scales for volume and molecular weight. Lines describe the minimum crystal volume required for a certain resolution given the parameters shown assuming a Nave–Hill effect (photoelectron escape) of 1, as all crystals are larger than 1 µm^3, and 100 photons per *hkl* (equation 16 from Holton & Frankel, 2010). The lines indicate a specific molecular weight and *B* factor and should be taken as a guide to the rough range described. The wavelength on MASSIF-1 is fixed and the beam diameter is altered to match the crystal. The histogram shows the distribution of resolutions obtained in the scatter plot. $N = 15\,241$.

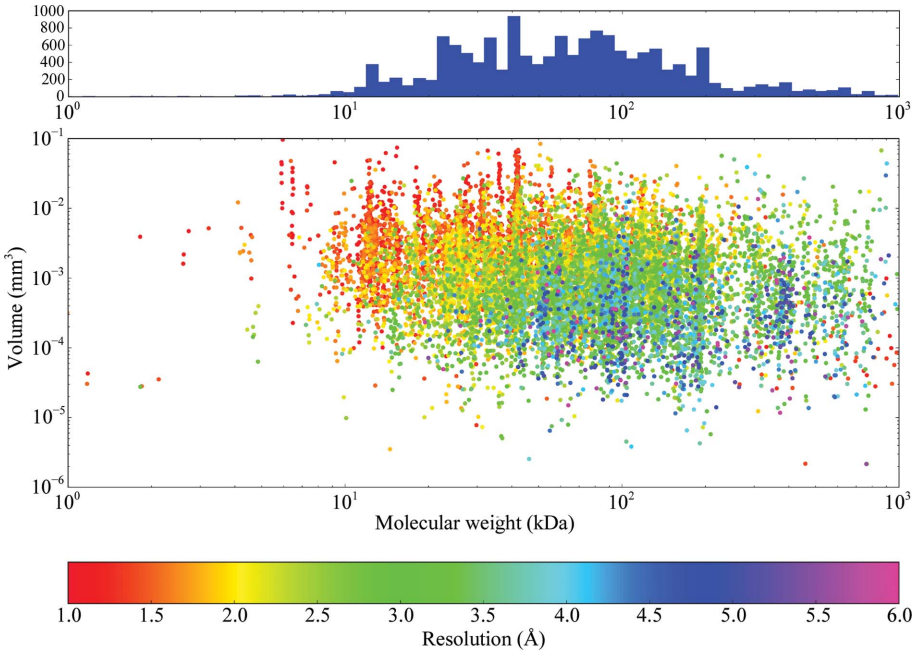


Figure 7
The relationship between crystal volume, molecular weight and obtained resolution. The crystal volume is plotted against the molecular weight of the entity crystallized coloured by the highest resolution cutoff from autoprocessing; note the logarithmic scales for volume and molecular weight. The histogram shows the distribution of molecular weight in the scatter plot. N = 15 241.

lie. The observation confirms the theoretical treatment as excellent and it clearly defines the standards that beamlines should be aiming for.

A more informative plot shows the molecular weight of the molecule crystallized plotted against the volume of the crystal coloured by the final resolution obtained from the data set (Fig. 7). The most striking observation is that, on average, the larger the molecule, the smaller the crystal. This is rather unfortunate given the dependence on volume for a given resolution. It is also interesting as a distribution of the molecules studied: there is a clear drop off after ∼200 kDa, showing the current range of samples studied on MASSIF-1. This cutoff is significant when considering the role of synchrotron beamlines in the future of structural biology. With technological advances in cryo-electron microscopy (cryo-EM) allowing structure determination at medium resolutions for very large complexes (Subramaniam *et al.*, 2016), X-ray crystallography will soon no longer be the method of choice for systems over ∼120 kDa. While analysing proteins below this molecular weight is possible by cryo-EM (Khoshouei *et al.*, 2017) it remains extremely difficult, requiring an excellent sample, and has not yet obtained the same resolutions and speeds of data acquisition as a modern synchrotron beamline. Fig. 7 demonstrates that the two techniques remain highly complementary: as molecular weight increases, crystal volume and resolution tend to decrease, making structure determination by X-ray crystallography harder. This trend is inverted for cryo-EM, meaning that X-ray crystallography can concentrate on the <100 kDa macromolecules, providing high data throughput and resolution, with cryo-EM working in the >100 kDa region where the resolution will be equivalent, or higher, and experiments less difficult. Most crystals in this molecular-weight range lie in the 10^{-4}–10^{-2} mm^3 range and would require X-ray beams from 30 to 100 µm in diameter.

4. Conclusion

The fully automatic collection of data at MASSIF-1 has allowed the study and comparison of the physical and molecular properties of a wide range of crystals and their constituent macromolecules for the first time. The results provide an overview of how size and shape is distributed over these crystals to a high level of precision, superseding assumptions based on individual systems. Coupling these data to the internal properties of these crystals from the processing of derived data sets, we can start to challenge assumptions and theoretical treatments. Have our assumptions been correct? Regarding crystal shapes and sizes, many have been wrong. These data demonstrate that an 'average' crystal on MASSIF-1 is more likely to be a plate or needle with a minimum dimension in the 30–50 µm range. The expectation that smaller or flatter crystals will cool better also seems to be incorrect and, as is so common in experimentation, sample handling and preparation is probably much more important. It is also reassuring that the theory behind diffraction and radiation damage holds extremely well in the real world.

How can these data be used? This study is limited to the samples sent to a single beamline and cannot therefore be representative of all projects. However, we believe that it can still help in directing the needs of future facilities. Globally, the data can guide the development of future facilities by demonstrating that a range of beam diameters are required, even if larger complexes will be increasingly top-sliced by cryo-EM, and that the theoretical limits should be strived for in an experimental station. New, highly intense, beamlines with submicrometre beams are being constructed across the world that will be excellent for microcrystals and time-resolved experiments (Cohen *et al.*, 2014; Sanchez-Weatherby *et al.*, 2019). For more standard experiments, the data presented here demonstrate that microfocus beamlines will still be needed and, crucially, beamlines with larger focal spots should not be neglected. Further insights could also be gained by linking the data collected here to crystallization databases from highly automated facilities (Ng *et al.*, 2016; Shaw Stewart & Mueller-Dieckmann, 2014), providing a further link between crystallogenesis and the final result. This link could inform the crystallization laboratory on the highest quality data as well as volumetric data and how these relate to the conditions that produced them. Additionally, further

information on crystallization techniques, such as cryoprotection protocols, linked to the information presented here would be very valuable (Newman *et al.*, 2012). Studies such as this have been limited to the PDB and, while highly informative (Abrahams & Newman, 2019), no data are available on the crystals or the screening required to obtain the result. More specifically, the data gathered can help individual projects by informing on the spread of volumes and how these relate to data quality, potentially improving data-collection strategies. The analysis of these data has already improved the operation of MASSIF-1 (Svensson *et al.*, 2015, 2018), but could this go further? Recent developments in machine learning could be applied to all of the data collected and may help to improve data-collection strategies. Looking more closely with more data than has previously been available, questions such as 'when is a helical or multi-position data collection better than a single-position strategy?' and 'can specific strategies, such as SAD, be improved?' could be answered. The analysis presented here has only started to delve into the data and we hope that modern data-science techniques could help further improve the measurement of diffraction data from protein crystals.

Acknowledgements

We would like to thank all of the users of MASSIF-1 over the years for trusting the beamline to physically handle and make data-collection decisions for their precious samples. These samples have made the beamline the success it is and have provided the data on which this study is based.

References

Abad-Zapatero, C. (2012). *Acta Cryst.* D**68**, 613–617.
Abrahams, G. J. & Newman, J. (2019). *Acta Cryst.* F**75**, 184–192.
Amunts, A., Drory, O. & Nelson, N. (2007). *Nature (London)*, **447**, 58–63.
Andrews, S. J., Hails, J. E., Harding, M. M. & Cruickshank, D. W. J. (1987). *Acta Cryst.* A**43**, 70–73.
Andrews, S. J., Papiz, M. Z., McMeeking, R., Blake, A. J., Lowe, B. M., Franklin, K. R., Helliwell, J. R. & Harding, M. M. (1988). *Acta Cryst.* B**44**, 73–77.
Bagaria, A., Jaravine, V. & Güntert, P. (2013). *Comput. Biol. Chem.* **46**, 8–15.
Berman, H. M., Coimbatore Narayanan, B., Di Costanzo, L., Dutta, S., Ghosh, S., Hudson, B. P., Lawson, C. L., Peisach, E., Prlić, A., Rose, P. W., Shao, C., Yang, H., Young, J. & Zardecki, C. (2013). *FEBS Lett.* **587**, 1036–1045.
Berman, H. M., Gabanyi, M. J., Groom, C. R., Johnson, J. E., Murshudov, G. N., Nicholls, R. A., Reddy, V., Schwede, T., Zimmerman, M. D., Westbrook, J. & Minor, W. (2015). *IUCrJ*, **2**, 45–58.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bourenkov, G. P. & Popov, A. N. (2010). *Acta Cryst.* D**66**, 409–419.
Bowler, M. G. & Bowler, M. W. (2014). *Acta Cryst.* F**70**, 127–132.
Bowler, M. W., Guijarro, M., Petitdemange, S., Baker, I., Svensson, O., Burghammer, M., Mueller-Dieckmann, C., Gordon, E. J., Flot, D., McSweeney, S. M. & Leonard, G. A. (2010). *Acta Cryst.* D**66**, 855–864.
Bowler, M. W., Montgomery, M. G., Leslie, A. G. W. & Walker, J. E. (2006). *Acta Cryst.* D**62**, 991–995.
Bowler, M. W., Nurizzo, D., Barrett, R., Beteva, A., Bodin, M., Caserotto, H., Delagenière, S., Dobias, F., Flot, D., Giraud, T., Guichard, N., Guijarro, M., Lentini, M., Leonard, G. A., McSweeney, S., Oskarsson, M., Schmidt, W., Snigirev, A., von Stetten, D., Surr, J., Svensson, O., Theveneau, P. & Mueller-Dieckmann, C. (2015). *J. Synchrotron Rad.* **22**, 1540–1547.
Bowler, M. W., Svensson, O. & Nurizzo, D. (2016). *Crystallogr. Rev.* **22**, 233–249.
Brockhauser, S., Svensson, O., Bowler, M. W., Nanao, M., Gordon, E., Leal, R. M. F., Popov, A., Gerring, M., McCarthy, A. A. & Gotz, A. (2012). *Acta Cryst.* D**68**, 975–984.
Cheeseman, M. D., Chessum, N. E. A., Rye, C. S., Pasqua, A. E., Tucker, M. J., Wilding, B., Evans, L. E., Lepri, S., Richards, M., Sharp, S. Y., Ali, S., Rowlands, M., O'Fee, L., Miah, A., Hayes, A., Henley, A. T., Powers, M., te Poele, R., De Billy, E., Pellegrino, L., Raynaud, F., Burke, R., van Montfort, R. L. M., Eccles, S. A., Workman, P. & Jones, K. (2017). *J. Med. Chem.* **60**, 180–201.
Cohen, A. E., Soltis, S. M., González, A., Aguila, L., Alonso-Mori, R., Barnes, C. O., Baxter, E. L., Brehmer, W., Brewster, A. S., Brunger, A. T., Calero, G., Chang, J. F., Chollet, M., Ehrensberger, P., Eriksson, T. L., Feng, Y., Hattne, J., Hedman, B., Hollenbeck, M., Holton, J. M., Keable, S., Kobilka, B. K., Kovaleva, E. G., Kruse, A. C., Lemke, H. T., Lin, G., Lyubimov, A. Y., Manglik, A., Mathews, I. I., McPhillips, S. E., Nelson, S., Peters, J. W., Sauter, N. K., Smith, C. A., Song, J., Stevenson, H. P., Tsai, Y., Uervirojnangkoorn, M., Vinetsky, V., Wakatsuki, S., Weis, W. I., Zadvornyy, O. A., Zeldin, O. B., Zhu, D. & Hodgson, K. O. (2014). *Proc. Natl Acad. Sci. USA*, **111**, 17122–17127.
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.
Evans, G., Axford, D. & Owen, R. L. (2011). *Acta Cryst.* D**67**, 261–270.
Fischer, M., Shoichet, B. K. & Fraser, J. S. (2015). *Chembiochem*, **16**, 1560–1564.
Frey, M., Genovesio-Taverne, J. C. & Fontecilla-Camps, J. C. (1991). *J. Phys. D*, **24**, 105–110.
Garman, E. (1999). *Acta Cryst.* D**55**, 1641–1653.
Grimes, J. M., Hall, D. R., Ashton, A. W., Evans, G., Owen, R. L., Wagner, A., McAuley, K. E., von Delft, F., Orville, A. M., Sorensen, T., Walsh, M. A., Ginn, H. M. & Stuart, D. I. (2018). *Acta Cryst.* D**74**, 152–166.
Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.
Helliwell, J. R. (1984). *Rep. Prog. Phys.* **47**, 1403–1497.
Hiruma, Y., Koch, A., Hazraty, N., Tsakou, F., Medema, R. H., Joosten, R. P. & Perrakis, A. (2017). *J. Biol. Chem.* **292**, 14496–14504.
Holton, J. M. (2009). *J. Synchrotron Rad.* **16**, 133–142.
Holton, J. M. & Frankel, K. A. (2010). *Acta Cryst.* D**66**, 393–408.
Hunter, J. D. (2007). *Comput. Sci. Eng.* **9**, 90–95.
Incardona, M.-F., Bourenkov, G. P., Levik, K., Pieritz, R. A., Popov, A. N. & Svensson, O. (2009). *J. Synchrotron Rad.* **16**, 872–879.
Joachim, U. & Markus, P. (2015). *Cryst. Res. Technol.* **50**, 560–565.
Juers, D. H. & Matthews, B. W. (2001). *J. Mol. Biol.* **311**, 851–862.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kharde, S., Calviño, F. R., Gumiero, A., Wild, K. & Sinning, I. (2015). *Nucleic Acids Res.* **43**, 7083–7095.
Khoshouei, M., Radjainia, M., Baumeister, W. & Danev, R. (2017). *Nat. Commun.* **8**, 16099.
Kiefersauer, R., Than, M. E., Dobbek, H., Gremer, L., Melero, M., Strobl, S., Dias, J. M., Soulimane, T. & Huber, R. (2000). *J. Appl. Cryst.* **33**, 1223–1230.
Kriminski, S., Caylor, C. L., Nonato, M. C., Finkelstein, K. D. & Thorne, R. E. (2002). *Acta Cryst.* D**58**, 459–471.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.

Li, Y., Muir, K., Bowler, M. W., Metz, J., Haering, C. H. & Panne, D. (2018). *Elife*, **7**, e38356.
Liu, J. J., Hu, Y. D. & Wang, X. Z. (2013). *Comput. Chem. Eng.* **57**, 133–140.
Low, B. W., Chen, C. C. H., Berger, J. E., Singman, L. & Pletcher, J. F. (1966). *Proc. Natl Acad. Sci. USA*, **56**, 1746–1750.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Mayans, O. & Wilmanns, M. (1999). *J. Synchrotron Rad.* **6**, 1016–1020.
Melnikov, I., Svensson, O., Bourenkov, G., Leonard, G. & Popov, A. (2018). *Acta Cryst.* D**74**, 355–365.
Mitchell, E. P. & Garman, E. F. (1994). *J. Appl. Cryst.* **27**, 1070–1074.
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.
Na, Z., Yeo, S. P., Bharath, S. R., Bowler, M. W., Balikçi, E., Wang, C.-I. & Song, H. (2017). *Cell Res.* **27**, 147–150.
Naschberger, A., Orry, A., Lechner, S., Bowler, M. W., Nurizzo, D., Novokmet, M., Keller, M. A., Oemer, G., Seppi, D., Haslbeck, M., Pansi, K., Dieplinger, H. & Rupp, B. (2017). *Structure*, **25**, 1907–1915.
Newman, J., Bolton, E. E., Müller-Dieckmann, J., Fazio, V. J., Gallagher, D. T., Lovell, D., Luft, J. R., Peat, T. S., Ratcliffe, D., Sayle, R. A., Snell, E. H., Taylor, K., Vallotton, P., Velanker, S. & von Delft, F. (2012). *Acta Cryst.* F**68**, 253–258.
Ng, J. T., Dekker, C., Reardon, P. & von Delft, F. (2016). *Acta Cryst.* D**72**, 224–235.
Nurizzo, D., Bowler, M. W., Caserotto, H., Dobias, F., Giraud, T., Surr, J., Guichard, N., Papp, G., Guijarro, M., Mueller-Dieckmann, C., Flot, D., McSweeney, S., Cipriani, F., Theveneau, P. & Leonard, G. A. (2016). *Acta Cryst.* D**72**, 966–975.
Oliphant, T. E. (2007). *Comput. Sci. Eng.* **9**, 10–20.
Papiz, M. Z., Andrews, S. J., Damas, A. M., Harding, M. M. & Highcock, R. M. (1990). *Acta Cryst.* C**46**, 172–173.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.
Powell, H. R., Battye, T. G. G., Kontogiannis, L., Johnson, O. & Leslie, A. G. W. (2017). *Nat. Protoc.* **12**, 1310.
Robert, X., Kassis-Sahyoun, J., Ceres, N., Martin, J., Sawaya, M. R., Read, R. J., Gouet, P., Falson, P. & Chaptal, V. (2017). *Sci. Rep.* **7**, 17013.
Russi, S., Juers, D. H., Sanchez-Weatherby, J., Pellegrini, E., Mossou, E., Forsyth, V. T., Huet, J., Gobbo, A., Felisaz, F., Moya, R., McSweeney, S. M., Cusack, S., Cipriani, F. & Bowler, M. W. (2011). *J. Struct. Biol.* **175**, 236–243.
Sanchez-Weatherby, J., Bowler, M. W., Huet, J., Gobbo, A., Felisaz, F., Lavault, B., Moya, R., Kadlec, J., Ravelli, R. B. G. & Cipriani, F. (2009). *Acta Cryst.* D**65**, 1237–1246.
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.
Schulze, W. M., Stein, F., Rettel, M., Nanao, M. & Cusack, S. (2018). *Nat. Commun.* **9**, 1701.
Shaw Stewart, P. & Mueller-Dieckmann, J. (2014). *Acta Cryst.* F**70**, 686–696.
Sorigué, D., Légeret, B., Cuiné, S., Blangy, S., Moulin, S., Billon, E., Richaud, P., Brugière, S., Couté, Y., Nurizzo, D., Müller, P., Brettel, K., Pignol, D., Arnoux, P., Li-Beisson, Y., Peltier, G. & Beisson, F. (2017). *Science*, **357**, 903–907.
Subramaniam, S., Kühlbrandt, W. & Henderson, R. (2016). *IUCrJ*, **3**, 3–7.
Svensson, O., Gilski, M., Nurizzo, D. & Bowler, M. W. (2018). *Acta Cryst.* D**74**, 433–440.
Svensson, O., Gilski, M., Nurizzo, D. & Bowler, M. W. (2019). *A Catalogue Of Characteristics From All Samples Processed On The Fully Autonomous ESRF Beamline MASSIF-1 Between 2014 And 2018*. https://doi.esrf.fr/10.15151/ESRF-DC-186715792.
Svensson, O., Malbet-Monaco, S., Popov, A., Nurizzo, D. & Bowler, M. W. (2015). *Acta Cryst.* D**71**, 1757–1767.
Teng, T.-Y. & Moffat, K. (1998). *J. Appl. Cryst.* **31**, 252–257.
Thompson, M. C., Cascio, D. & Yeates, T. O. (2018). *Acta Cryst.* D**74**, 411–421.
Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.
Warkentin, M., Berejnov, V., Husseini, N. S. & Thorne, R. E. (2006). *J. Appl. Cryst.* **39**, 805–811.
Weichenberger, C. X. & Rupp, B. (2014). *Acta Cryst.* D**70**, 1579–1588.
Xu, X., Li, Y., Bharath, S. R., Ozturk, M. B., Bowler, M. W., Loo, B. Z. L., Tergaonkar, V. & Song, H. (2018). *Nat. Commun.* **9**, 3183.
Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. M. & Garman, E. F. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 20551–20556.