INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 07 Issue 06 June 2024

DOI: 10.47191/ijmra/v7-i06-26, Impact Factor: 8.22

Page No. 2626-2635

Using the Autoproc Toolkit for Data Processing and Analysis

Muhammad D. Hassan
Collage of Health and Medical Techniques, Northern Technical University, Kirkuk, Iraq

ABSTRACT: A typical diffraction experiment yields several pictures and datasets from various crystals within a brief timeframe. This poses a barrier for the efficient operation of advanced synchrotron beamlines and the following data processing. Novice users, in particular, may have a sense of being inundated by the tables, graphs, and numerical information that is provided to them via different data-processing systems and software packages. Here, we outline many common obstacles that users face while processing a collection of photos to generate a processed dataset. Our main emphasis is on the challenges that often occur while transitioning from the earliest phases of converting experimental data collecting into an interpreted electron density model. Examining certain data properties during processing may frequently help solve or identify challenges such as unforeseen crystal formations, issues with crystal manipulation, and suboptimal selection of data-collection techniques. Ultimately, it is important to distinguish between concerns that are beyond of one's immediate control after the experiment has concluded and those that can be dealt with later. Presenting auto PROC, an innovative software solution that combines external processing packages with new tools and an automated workflow script. This program is designed to help users effectively handle data that is impacted by the issues outlined above. It specifically focuses on automating the processing of multi-sweep data sets produced from multi-axis gonio metrics. It offers consumers both instruction and important insights into the processing of data that is not done online.

KEYWORDS: Various Crystals, PROC, Processing of Data, Collection Techniques

I. INTRODUCTION

In an optimal scenario, a standard diffraction experiment would yield a series of images that display (i) clear and unmistakable diffraction patterns, (ii) distinct and clearly defined lines, (iii) flawless spot shapes, (iv) a single crystal lattice, (v) multiple measurements for each (hkl) value within the crystal's range of diffracting, and (vi) statistically reliable data. It is often seen that diffraction experiments do not meet the parameters outlined before. This might possibly cause the structure to become intractable, either by molecular replacement or experimental phasing, or result in significant portions of the electron density being indecipherable. It is now advantageous to examine some prevalent issues that may have been disregarded during the first stages of data processing[1].

II. CENTRE OF THE BEAM

The primary cause of failure in indexing a specific collection of photographs is most likely mistakes in precisely locating the beam center, which is dependent on the coordinate systems used by the integration program. The coordinate system that is being employed is usually not clearly mentioned, despite the fact that the values that are presented in the image header are almost always correct. Furthermore, it is not unheard of to come across picture headers that include nonstandard elements and values that are not valid. Consequently, it is of the utmost importance to provide the precise values by using a separate program-specific site-definition file or template for processing[2].

The intricacy of this work may be intimidating, especially for a person with little to no prior expertise who is attempting to examine data collected from a beamline that is unknown to them. It is essential to validate two things in order to guarantee proper indexing: (i) the precision of the beam-center coordinates in the header, and (ii) the particular convention that is used in order to describe these values.

Using the Autoproc Toolkit for Data Processing and Analysis

Under typical conditions, detector manufacturers and beamline scientists strive to ensure the accuracy and easy accessibility of this information in the necessary format for different processing packages. On the other hand, it is conceivable that the user did not capture this information in a sufficiently timely manner. Additional difficulties may arise as a consequence of the fact that, in some instrumental setups, the location of the beam center may fluctuate as a consequence of changes in the temperature of the surrounding environment. There are times when alterations to the beamline software have the potential to disturb a standard that had been established in the past for the storage of beam-center coordinates. The right axis convention may typically be determined by analyzing the concentricity of certain visual characteristics around the assumed beam center using all eight standards[3].

III. A NUMBER OF LATTICES

As can be seen in Figure 1, crystals often display many lattices during the process of data collection. The occurrence of these phenomena may be attributed to the presence of split crystals, the presence of a satellite crystal inside the loop, or a unique interaction between domains, such as the absence of merohedral twining.

The relative strength and level of overlap between two or more lattices may not always be immediately apparent in the first picture, particularly when there is a clear relationship between them. In order to ensure accurate results, it is essential to analyze a large number of photos captured at various angles relative to the rotating axis, including 45 degrees and 90 degrees from the initial position. If another fragment of the crystal (or loop/mount) is introduced into the beam while the crystal is rotating, there is a possibility that a secondary lattice may become apparent. This may occur intermittently. Adjusting the crystal's location to alter the direction of the beam onto a new area might sometimes mitigate the potential issues caused by extra lattices during indexing or integration. Recent advancements in synchrotron technology have enabled the use of grid, line, or mesh scans, which are effective in identifying the most well-organized section of a crystal[4][5].

IV. REGULAR INDEXING

Diffraction data sets consist of many sweeps, each including a series of pictures obtained by continuous rotation around a common axis. MOSFLM and XDS are used to individually process these searches, specifically for the purpose of indexing. Ensuring the constant indexing of these diverse surveys is often crucial. In order to achieve consistency, it is essential that all scans be indexed in a same way, with a specific focus on the crystal's point-group symmetry. During the process of indexing the sweeps that correspond to distinct wavelengths in multi-wavelength data sets, it is obvious that it is necessary to meet this consistency criteria. In the event that this is not done, the subsequent use of the data for MAD would be characterized by chaos and confusion

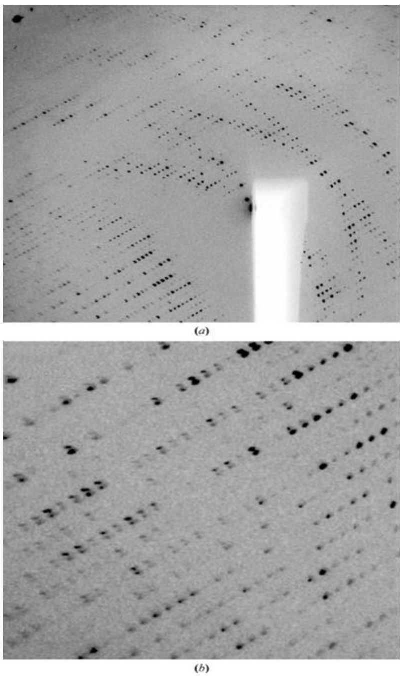

Figure 1. An illustration of numerous lattices may be seen in the above example. In orientation (a), two lattices of similar strength are represented by separate unique lunes. On the other hand, when seen from orientation (b), the reflections from the two lattices nearly completely overlap every other reflection.

using a shared absorption surface specified in the crystal frame, a higher level of consistency is necessary. In order to reach this level of consistency, it is essential to possess a comprehensive identity rather than mere equivalence up to a point-group operation. This might potentially have a substantial impact, particularly in the context of sulfur-SAD phasing at longer wavelengths. This happens when there is a lack of noticeable changes and considerable absorption[6][7].

When dealing with consistent indexing, it is necessary to consider the established connections between the crystal orientations associated with different sweeps. This demands a thorough description of the instrument, especially the gonio stat.

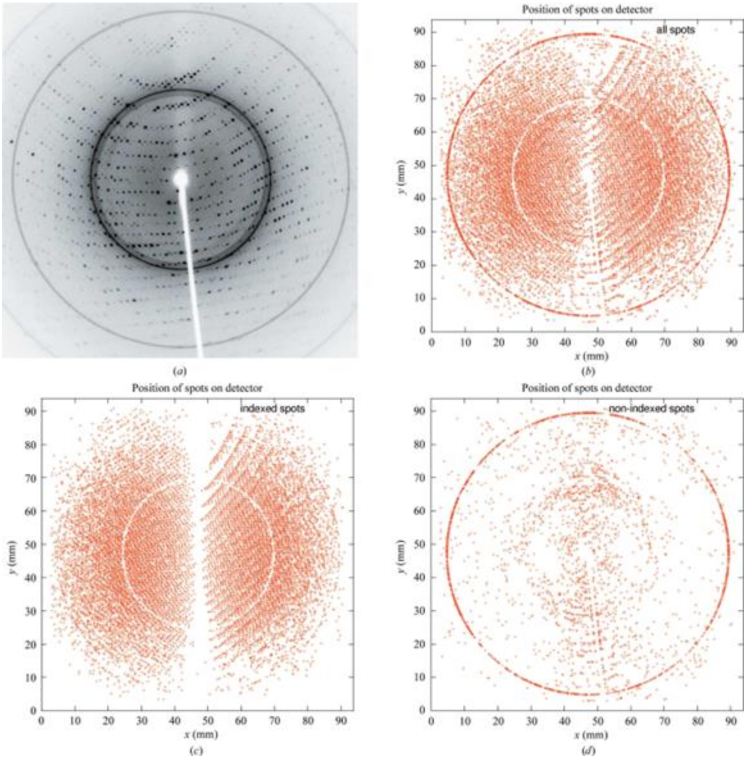

Figure 2. The taxonomy of spots for ice rings is as follows: (a) the originality of single photo; (b) spots that are represented by a red cross and collected from a series of photos; (c) the selection of spots suitable for indexing, illustrated by a 'white' circle, with a strong ice ring that prevents the detection of any spots; and (d) the remaining spots that have not been indexed.

V. RINGS OF ICE

In the event that it is not feasible to avoid them from occurring throughout the cryocooling procedure of the crystal, the resolution ranges that are affected need to be removed from all of the processing steps. During the process of indexing, all identified locations may be categorized as either part of the identified indexing solution or remaining unallocated. This enables a simple visual inspection to determine the existence of ice rings (refer to Figure 2).

VI. EXAMINING THE RECORDED REFLECTION

One must use the CCP4 software HKLVIEW to examine the reflection file(s). This program will exhibit simulated precession images of a solitary column extracted from an MTZ file, including the specified reflections. The HKLVIEW interface provides a range of features for zooming in and out of the reflection locations. The dimensions and hue of these markings, shown as square boxes in shades of gray, are directly correlated with the intensity level of the reflected light. A novel feature has been included into the PHENIX software package, enabling the generation of analogous images via the use of unprocessed diffraction data.

These images exhibit many fundamental characteristics of reflections. Figure 3 displays very intense reflections at a high level of detail. These observations align with the usual ice-ring resolutions of around 2.15 A˚. They indicate issues that occurred during data processing, such as inadequate removal of ice rings, or during merging, such as the inability to recognize outliers.

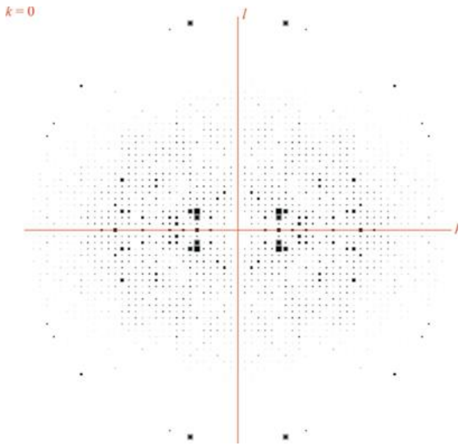

Figure 3. Using HKLVIEW, a snapshot of a pseudo-precession was taken.

A. *Limits of Anisotropy in Diffraction*

A typical scenario is illustrated by the decrease in intensity that occurs in different directions within the h00, 0k0, or 00l planes. This is because the amount of crystal that is contained within the beam may change as the crystal is rotated throughout the experiment. Additionally, the arrangement of atoms within the crystal may be more favorable in certain directions compared to other directions. The presence of anisotropy in the data results in a consistent decrease in accuracy for a specific portion of the data. This may lead to complications when using approaches that presume a more consistent behavior of the diffraction data. Some examples of these techniques are molecule replacement, substructure discovery via the use of normalized structural factors, and real-space methods such as density modification [8][9].

B. *Shells For Problematic Resolution*

It is expected that there would be a gradual decrease in the intensity values as the resolution increases. Generally, reflections at lower resolutions are stronger compared to those at higher resolutions. Anything that deviates from this course of action is especially suspect. A mismanagement of the high background that is associated with diffuse ice diffraction during the integration process may be the source of very strong reflections that are captured at a higher resolution[10]. When the resolution of these strong high-resolution reflections is examined, it may be possible to determine whether or not there is a problem with ice rings (refer to the following example displayed in Figure 3). The resolutions of the ice rings that are most often seen are as follows: 3.90, 3.67, 3.44, 2.67, 2.25, 2.07, 1.95, 1.92, 1.88, and 1.72 angstroms.

C. *Overloading Detectors and Lacking Low-Resolution Data*

When the detector is overloaded, the measurements that relate to it will be ignored during the processing of the data. they have an effect on the most intense reflections, which often become apparent at low resolution. density modification, and bulk-solvent modeling in refinement are just some of the vital approaches that need these reflections, despite the fact that their quantity may be restricted. These reflections are essential for the proper utilization of these many techniques. To address the issue of overloads, it is recommended to conduct a distinct low-intensity scan and include it into the dataset during the processing stage. To get more accurate measurements of low-resolution, it is advisable to use a weakened beam instead of reducing the exposure duration. Additionally, capturing a wider variety of angles per picture would also contribute to improved results.

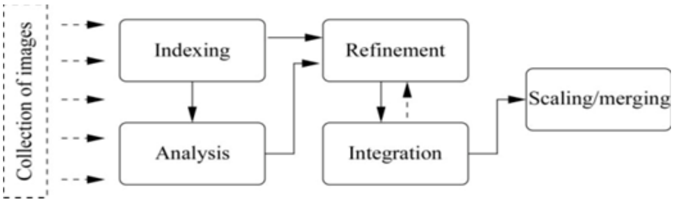

Figure 4. Standard offline data-processing procedures.

Improve the accuracy of measurements by reducing mistakes caused by intermittent crystal rotation.

VII. ANTICIPATE THE UNEXPECTED

Even a collection of excellent photos, leading to a dataset of high quality with strong statistical properties, may not be sufficient to solve the structure. This may be particularly exasperating if, for instance, the unusual signal seems to have a high level of accuracy in terms of resolution or if there is a structure in the PDB that is quite similar.

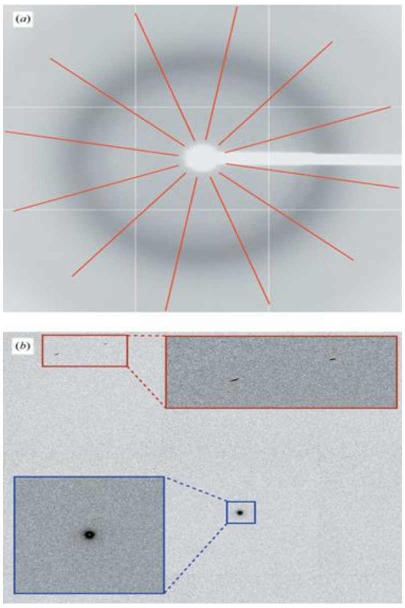

Figure 5. GETBEAM is used to establish direct-beam coordinates. This involves the utilization of a back-ground-only picture for 1vq0, where lines are employed to calculate correlations between opposing regions. The extended regions around the direct-beam position are highlighted in blue, and any rogue high-value pixels are noted in red. In addition, a section of a picture that was captured with a direct beam is analyzed.

One possibility is that the purified protein is really an expression artifact rather than the isolated protein that was intended to be obtained. One possible solution would be to go through the Protein Data Bank (PDB) for entries that have a space group and unit cell that are similar to one another. In the event that there is an existing entry, especially for a protein that is either identical to or closely resembles the one that is naturally produced by the expression system that was used for the sample preparation, molecular replacement may be used in order to either confirm or reject it as a viable answer. It has come to our attention that there have been several occasions in which data were collected from crystals of inorganic pyrophosphatase that were produced from Escherichia coli[11].

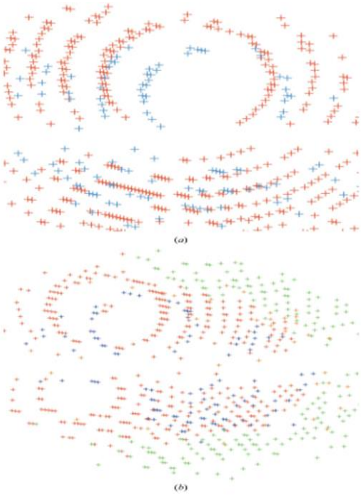

Figure 6. The autoPROC software generates visual representations of numerous lattices in the 1vk2 structure. Both images display the lattices using distinct colors. The red and blue colored lattices represent the two major lattices.

VIII. SOFTWARE CALLED AUTOPROC

Several extensive software programs have been created to assist users in navigating the several stages from photos to a complete, scaled, and integrated data set. During the last five years, we have created a collection of applications known as the autoPROC framework, which includes many third-party products. When it comes to the automated processing of diffraction photographs obtained from single-sweep or multi-sweep studies, the modules that are included in this framework are designed to be used offline on certain occasions. The multi-wavelength MAD, low-resolution and high-resolution passes, as well as the collection of inverse-beam or interleaved-wavelength data are all included in these experiments. It is common practice for the technique to consist of the following stages: There are various stages involved in the process: the first step is to examine the image, the second step is to identify spots, the third step is to classify them, the fourth step is to evaluate the quality of the diffraction and detector parameters, the fifth step is to improve the initial unit-cell parameters, orientation, and mosaicity, the sixth step is to determine the most likely space group, the seventh step is to combine all of the images, and the eighth step is to adjust and merge the integrated intensities (see Figure 4).

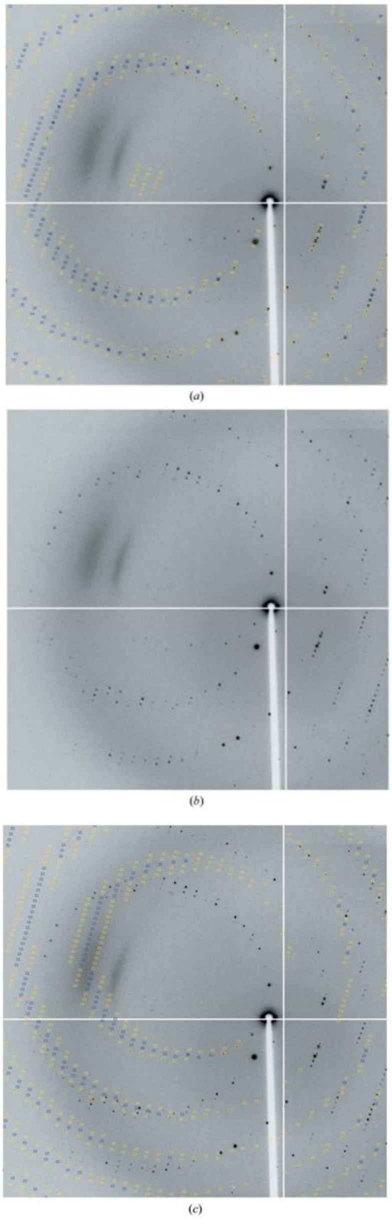

Figure 7. Calculating distinct orientation matrices for various lattices in the 1vk2 structure: (a) Forecasts for the primary lattice (complete, shown in blue; partial, indicated in yellow; excessively broad in ‘, indicated in green); (b) Diffraction picture without any forecasts; (c) Predictions for the secondary lattice.

A. Implementation

The autoPROC programming language is organized as a set of modules that correspond to the various steps shown in Figure 4. There is a clear separation between each module and the others, and each module has its own set of input and output parameters that are well stated. The MOSFLM and SCALA components of the pipeline were the ones that were used most often in the first iteration. In further developments, XDS was used as the data-processing engine, and POINTLESS was utilized for the purpose of identifying the space-group. Exclusive software components specifically designed for autoPROC are accessible to enhance usefulness and reliability. The user is given with a set of supplementary tools to assist them in the automatic processing of data. Program execution mostly relies on command-driven operations, which may be carried out with a single command with default parameters.

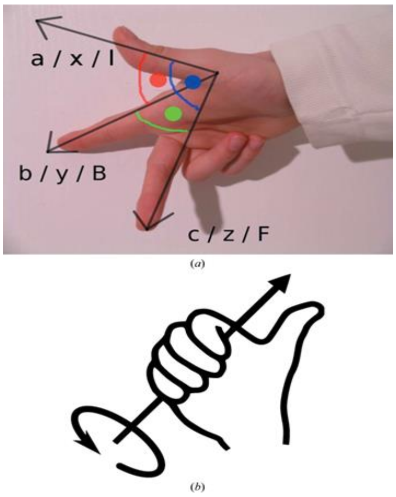

Figure 8. Illustration of a right-handed coordinate system and the right-hand rule for rotational movement around an axis Combining of different datasets. This functionality also makes it easier to integrate the program into a more comprehensive internal workflow, which may be useful in many situations, such as when conducting drug discovery projects or designing structure-based drugs.

B. Finding The Center of The Beam

The objective of the GETBEAM software is to provide the user with assistance. The maximum pixel value in the image array is selected when provided with a direct-beam shot picture. The search technique is limited to the starting beam-centre value, often derived from the picture header, to prevent the detection of anomalous pixels or outliers, as seen in Figure 5.

If a direct-beam shot picture is not accessible, a sequence of regular photographs might be used. In order to mitigate the impact of diffraction spots on these photos, an underlay image is created. This involves selecting the minimum pixel value from all photographs at each point. If many photos with sufficiently different oscillation angles are utilized, the resulting final image should not include any discernible diffraction spots. Ideally, the only surviving characteristic of this picture should be the diffused backdrop mostly originating from the solvent present in and around the crystal. If the direct beam is oriented in a direction that is perpendicular to the surface of the detector, then the distribution ought to be radially symmetric, with the coordinates of the direct beam being at the center of the distribution. Figure 5 illustrates a collection of lines that originate from the current beam center. These lines were generated in order to measure the degree of correlation between the pixel values that are shown along these lines. In order to decide.

This was the case for all of the information. In contrast, the installation of GETBEAM resulted in the average distance remaining at a paltry 5.7 pixels throughout the whole process. The benefit of utilizing this approach to verify for the coordinate convention of header data is shown in a way that cannot be disputed by this.

C. Various Lattices

AutoPROC makes it possible to recognize many lattices and to accurately index the principal lattice (see to Figure 6 for further information). It is possible to do this by picking locations in an iterative manner that correspond to the current indexing matrix. There are similarities between this method and the one that is offered by

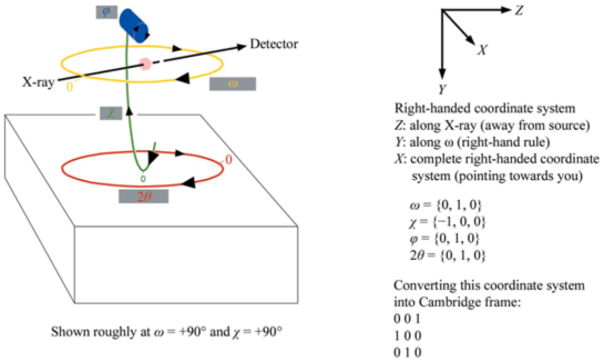
φ
Detector
X-ray
0
ω
χ
0
0
2θ
Shown roughly at ω = +90° and χ = +90°
Z
X
Y
Right-handed coordinate system
Z: along X-ray (away from source)
Y: along ω (right-hand rule)
X: complete right-handed coordinate system (pointing towards you)
ω = {0, 1, 0}
χ = {−1, 0, 0}
φ = {0, 1, 0}
2θ = {0, 1, 0}
Converting this coordinate system into Cambridge frame:
0 0 1
1 0 0
0 1 0

Figure 9. Describing the axes of a gonio stat. The Cambridge reference frame, often known as the Cambridge coordinate system, adheres to the definition of MOSFLM.

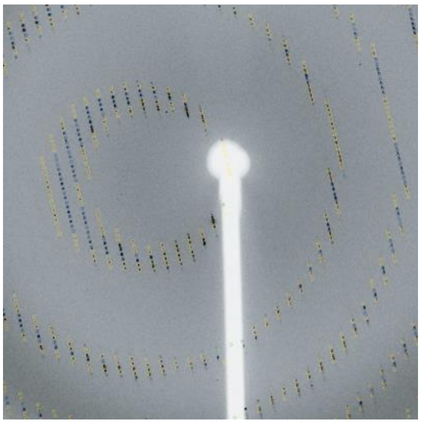

Figure 10. The XDS findings are visualized using MOSFLM by converting the orientation matrix from XDS into MOSFLM format. This conversion also includes the distance, beam center, and mosaicist. The generated descriptions may be easily imported into MOSFLM, which provides interactive capabilities for displaying forecasts, analyzing the beam center or ice rings, and modifying mosaicist settings.

During the second round of indexing, spots that are manifestly not compatible with the current orientation matrix are gathered together for further examination. Consequently, this makes it possible to do automatic identification of additional lattices and analysis of the link between other lattices and the original lattice. In addition, any regions that are not included in any of the orientation matrices that were gathered are investigated in order to detect any possible ice rings that may be present in the diffraction images (Fig. 2).

As shown in Figure 7, the processing of data is carried out by using the ideal orientation that was established for the lattice that has the highest density of particles. Nevertheless, the user also has the choice to integrate any of the lattices that have a lower population density via the usage of this option

D. Regular Indexing

For the purpose of calculating the movements of general gonio metrics (Kappa and Eulerian) and the movements of two arms, if appropriate, AutoPROC makes use of an auxiliary software known as KAPPAROT. It is necessary to carry out this action in order to guarantee accurate indexing consistency across the many sweeps of a multi-sweep data set that includes a gonio stat. With regard to right-handed coordinate systems and axis rotations, the definitions of instruments are flexible and follow to clear criteria (Fig. 8).

The description that is shown in Figure 9 indicates that the distinct connection that exists between the separate sweeps is maintained. As long as the essential gonio stat angles are provided in each picture header, the orientation matrices that are produced will precisely match to the movements that are known to be associated with the gonio stat.

E. Data Visualization

As can be seen in Figure 10, autoPROC converts the XDS orientation information into a format that is compatible with MOSFLM. This is done in order to validate the results that were obtained throughout the data processing process. Because of this, it is possible to visually examine the predictions that were produced by utilizing the present orientation matrix, unit-cell characteristics, mosaicist, and other criteria.

F. Final Product

It is just the most important facts that are shown to the user, such as the indexing solution, the identification of the space group, the merging statistics, and the automatic computation of the high-resolution limit. Additionally, any relevant annotations and warning messages are also included. The data includes many statistics and precise characteristics, which are provided either based on the resolution or the picture number. The former enables the determination of suitable resolution thresholds for decision-making, while the latter may illustrate events or patterns that occur throughout the rotation of the crystal (see to Fig. 11 for an example).

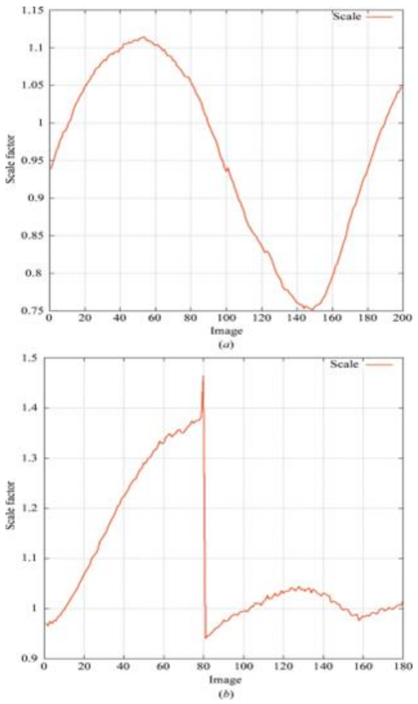

Figure 11. Calculation of the scale factor using background scatter as a function of picture number in XDS. The generation of these graphs is automated by autoPROC. (a) illustrates the distinct scattering power of a crystal as it undergoes a complete 180-degree rotation; (b) depicts an occurrence occurring between two pictures of 2eth.

IX. CONCLUSIONS

We identify many typical challenges that users encounter while processing a collection of pictures to produce a processed dataset. Our primary focus is on the difficulties that often arise during the transition from the first stages of transforming experimental data collection into an interpreted electron density model. An analysis of certain data attributes throughout the processing stage may often assist in resolving or detecting hurdles such as unexpected crystal forms, difficulties in manipulating crystals, and poor choices in data-collection approaches. It is crucial to differentiate between problems that are beyond of one's immediate control once the experiment has ended and those that can be addressed at a later time. Introducing auto PROC, a cutting-edge software solution that integrates external processing packages with advanced tools and an automated workflow script. This tool is specifically developed to assist users in efficiently managing data that is affected by the aforementioned difficulties. The main objective is to automate the processing of multi-sweep data sets generated by multi-axis gonio measurements. It provides users with both guidance and valuable knowledge on offline data processing.

REFERENCES

1) Vonrhein, C., Tickle, I. J., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., & Bricogne, G. (2018). Advances in automated data analysis and processing within autoPROC, combined with improved characterisation, mitigation and visualisation of the anisotropy of diffraction limits using STARANISO. Acta Crystallogr. A, 74, A360-A360.
2) Monleón, D., Colson, K., Moseley, H. N., Anklin, C., Oswald, R., Szyperski, T., & Montelione, G. T. (2002). Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed Linux-based computing architecture, and an integrated set of spectral analysis tools. Journal of structural and functional genomics, 2, 93-101.
3) Vonrhein, C., Flensburg, C., Keller, P., Fogh, R., Sharff, A., Tickle, I. J., & Bricogne, G. (2024). Advanced exploitation of unmerged reflection data during processing and refinement with autoPROC and BUSTER. Acta Crystallographica Section D: Structural Biology, 80(3).
4) Vonrhein, C., & Bricogne, G. (2008). AutoPROC–a framework for automated data processing. Foundations of Crystallography, 64(a1), 78-78.
5) Powell, H. R. (2017). X-ray data processing. Bioscience reports, 37(5), BSR20170227.
6) Baran, M. C., Moseley, H. N., Sahota, G., & Montelione, G. T. (2002). SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. Journal of biomolecular NMR, 24, 113-121.

7) Kavi, K. M., Aborizka, M., & Kung, D. (2002, October). A framework for designing, modeling and analyzing agent based software systems. In Fifth International Conference on Algorithms and Architectures for Parallel Processing, 2002. Proceedings. (pp. 196-200). IEEE.

8) Huang, Y. J., Moseley, H. N., Baran, M. C., Arrowsmith, C., Powers, R., Tejero, R., ... & Montelione, G. T. (2005). An integrated platform for automated analysis of protein NMR structures. In Methods in enzymology (Vol. 394, pp. 111-141). Academic Press.

9) Winter, G. (2005). Application of Automation to Data Processing & Analysis. CCP4 Newsletter on Protein Crystallography, (43).

10) Noske, G. D., Nakamura, A. M., Gawriljuk, V. O., Fernandes, R. S., Lima, G. M., Rosa, H. V. D., ... & Godoy, A. S. D. (2021). A crystallographic snapshot of SARS-CoV-2 main protease maturation process. Journal of Molecular Biology, 433(18), 167118.

11) Wollenhaupt, J., Metz, A., Barthel, T., Lima, G. M., Heine, A., Mueller, U., ... & Weiss, M. S. (2020). F2X-universal and F2X-entry: structurally diverse compound libraries for crystallographic fragment screening. Structure, 28(6), 694-706.

12) Beroza, P., Crawford, J. J., Ganichkin, O., Gendelev, L., Harris, S. F., Klein, R., ... & Lemmen, C. (2022). Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. Nature Communications, 13(1), 6447.