crystals

MDPI

Article

Application of Serial Crystallography for Merging Incomplete Macromolecular Crystallography Datasets

Ki Hyun Nam

College of General Education, Kookmin University, Seoul 02707, Republic of Korea; structure@kookmin.ac.kr

Abstract: In macromolecular crystallography (MX), a complete diffraction dataset is essential for determining the three-dimensional structure. However, collecting a complete experimental dataset using a single crystal is frequently unsuccessful due to poor crystal quality or radiation damage, resulting in the collection of multiple incomplete datasets. This issue can be solved by merging incomplete diffraction datasets to generate a complete dataset. This study introduced a new approach for merging incomplete datasets from MX to generate a complete dataset using serial crystallography (SX). Six incomplete diffraction datasets of β-glucosidase from *Thermoanaerobacterium saccharolyticum* (TsaBgl) were processed using CrystFEL, an SX program. The statistics of the merged data, such as completeness, CC, CC*, R_{split}, R_{work}, and R_{free}, demonstrated a complete dataset, indicating improved quality compared with the incomplete datasets and enabling structural determination. Also, the merging of the incomplete datasets was processed using four different indexing algorithms, and their statistics were compared. In conclusion, this approach for generating a complete dataset using SX will provide a new opportunity for determining the crystal structure of macromolecules using multiple incomplete MX datasets.

Keywords: macromolecular crystallography; incomplete datasets; merging; data processing; serial crystallography; CrystFEL

Citation: Nam, K.H. Application of Serial Crystallography for Merging Incomplete Macromolecular Crystallography Datasets. *Crystals* **2024**, *14*, 1012. https://doi.org/10.3390/cryst14121012

Academic Editors: Eamor M. Woo and Jesús Sanmartín-Matalobos

Received: 1 November 2024
Revised: 18 November 2024
Accepted: 20 November 2024
Published: 22 November 2024

1. Introduction

X-ray crystallography provides valuable information for elucidating the structure of a macromolecule at atomic resolution [1,2]. Such structural information aids in the understanding of biomolecular functions; also, it is widely applied in drug design and enzyme engineering [3–6]. In macromolecular crystallography (MX), the three-dimensional structure is determined using diffraction data collected from a single crystal [7,8]. However, collecting a complete diffraction dataset from a single crystal is often challenging due to various factors, such as poor crystal quality and radiation damage [9,10]. When the quality of a crystal is poor or significant radiation damage occurs, the quality of X-ray diffraction will gradually decrease during data collection [11], leading to the collection of incomplete diffraction data such as low completeness (<90%), limited resolution range, or low signal-to-noise ratio.

When a complete diffraction dataset cannot be obtained, researchers usually attempt to collect another complete dataset using a different single-crystal sample. However, failure of this approach may result in multiple incomplete datasets and a lack of necessary data for structure determination. Although individual incomplete datasets cannot be directly utilized for reliable structural determination, data merging using the relative intensities can potentially generate a complete dataset with increased completeness [12].

Currently, there are several methods, such as Aimless [13], BLEND [14], and KAMO [15], XDS/XSCALE [16], nXDS [17] that merge data from multiple crystals. They are applied to generate complete datasets from poor diffraction data, as well as to determine the anomalously scattering atomic substructure for crystals with weak anomalous scatterers (e.g., S and P) [18].

Serial crystallography (SX), utilizing an X-ray free-electron laser or synchrotron X-rays, minimizes radiation damage and helps facilitate structure determination at room temperature [19–21]. SX employing pump-probe techniques with optical lasers or chemical mixing methods which enables the visualization of the molecular dynamics during specific macromolecular reactions [22–26]. In an SX experiment, many crystals are continuously delivered to the X-ray interaction point, with each crystal exposed to X-rays only once [27–30]. Then, the diffraction data obtained from numerous crystals are merged to produce a complete diffraction dataset for three-dimensional structural determination [31–35]. Thus, generating a complete dataset from multiple incomplete single-shot diffraction datasets is considered analogous to merging multi-crystal datasets with SX data processing. Accordingly, the merging of multiple incomplete datasets collected in MX using an SX processing program is theoretically possible but has not yet been experimentally demonstrated.

This study demonstrated the feasibility of generating a complete dataset by processing incomplete datasets collected from MX using an SX program. Newly collected and previously collected datasets [36,37] for TsaBgl, a β-glucosidase from *Thermoanaerobacterium saccharolyticum*, were used as model datasets. Diffraction images from six different incomplete datasets were processed using the CrystFEL program, successfully yielding a complete dataset in terms of data collection and structure refinement statistics. In addition, the statistics of the data merging of the incomplete datasets by four different indexing algorithms (CrystFEL, DirAx, Taketwo, and XGANDALF) were compared. These findings provide an opportunity to determine the structures from the incomplete datasets collected in MX.

2. Materials and Methods

2.1. Protein Preparation and Crystallization

The protein preparation of TsaBgl has been reported previously [38]. Briefly, the expression DNA vector containing the gene encoding TsaBgl was transformed into *Escherichia coli* BL21(DE3). Then, the cells were cultured in an LB broth medium containing 0.1 μg/mL ampicillin, and protein expression was induced by adding 0.5 mM isopropyl β-D-thiogalactopyranoside (IPTG). After cell lysis by sonication, the supernatant was purified using Ni-NTA affinity chromatography and size exclusion chromatography using a Sephacryl S-100 column (GH Healthcare, Chicago, IL, USA). The purified protein in 10 mM Tris–HCl, pH 8.0, and 200 mM NaCl was concentrated for crystallization. TsaBgl crystallization was performed using the hanging drop vapor diffusion method at 20 °C. TsaBgl crystals were grown under the crystallization condition with 100 mM Tris–HCl, pH 7.5, 20% (*w*/*v*) polyethylene glycol 4000, and 200 mM magnesium chloride. The TsaBgl crystals were obtained within a month.

2.2. Data Collection

The diffraction dataset of TsaBgl (Data I) was collected using a Pilatus 6M detector (Dectris, Baden-Dättwil, Switzerland) on the 11C beamline at the Pohang Light Source II (PLS-II, Pohang, Republic of Korea) [39]. The X-ray wavelength was 0.9794 Å. The TsaBgl crystals were transferred to a cryoprotectant solution containing a reservoir solution supplemented with 20% (*v*/*v*) ethylene glycol for 5 s and mounted on the goniometer under a liquid nitrogen stream at 100 K. The crystal-to-detector distance was set to 300 mm. The reflections were indexed, integrated, and scaled by the HKL2000 program [40].

2.3. Data Processing

The 50 images each from the diffraction data of TsaBgl (Data I) collected in this study and 5 other diffraction datasets of TsaBgl (Data II–VI) from a previous study (Table S1), at a total of 300 images, were used as a model of an incomplete dataset. Each set of 50 images of Data I–VI were processed using the CrystFEL program (version 0.10.1) [32] with the MOSFLM indexing algorithm [41]. The indexing of the diffraction images was analyzed using *indexamajig* in CrystFEL with default parameters. Then, all 300 images from the

6 incomplete TsaBgl datasets (Data I–VI) were merged using the CrystFEL program with the identical indexing parameters. During data processing, the detector geometry was optimized using *geoptimiser* [42] in CrystFEL. The effects of the indexing algorithms were compared by processing the 300 images from the incomplete datasets (Data I–VI) using other indexing algorithms, such as DirAx [43], Taketwo [44], and XGANDALF [45], with default indexing parameters in CrystFEL. Data processing script and information using CrystFEL were summarized in Table S2.

2.4. Structure Determination

Phase problems were solved using a molecular replacement (MR) method with the MR-phaser in PHENIX [46]. The crystal structure of TsaBgl (PDB code: 8WFT) [36] was used as search model. Structure refinement was performed using phenix.refine in the PHENIX program [46]. Water molecules were added to the model structure during refinement using the default parameters in PHENIX. The model structures of TsaBgl from Data I and the merged dataset were built using the COOT program [47]. Data collection and refinement statistics of the Data I were summarized in Table S3. The model structures were evaluated using MolProbity [48]. The structural figures were generated using PyMOL (https://pymol.org).

3. Results

3.1. Merging of the Incomplete Dataset Using the SX Program

In MX data collection, when the diffraction intensity of a single crystal is good, complete diffraction data can be collected and processed using an MX program, which can then be used for structure determination (Figure 1). However, when the diffraction intensity of the single crystal is weak or poor, it becomes impossible to collect a complete dataset with the desired completeness and resolution. In such cases, new crystal samples are typically used to collect X-ray diffraction studies again. When a complete dataset is not obtained during diffraction data collection, incomplete datasets accumulate (Figure 1). By merging these incomplete datasets, a complete dataset can be generated, enabling the determination of the crystal structure (Figure 1). Theoretically, using SX programs, incomplete datasets can be processed and merged into a single complete dataset.

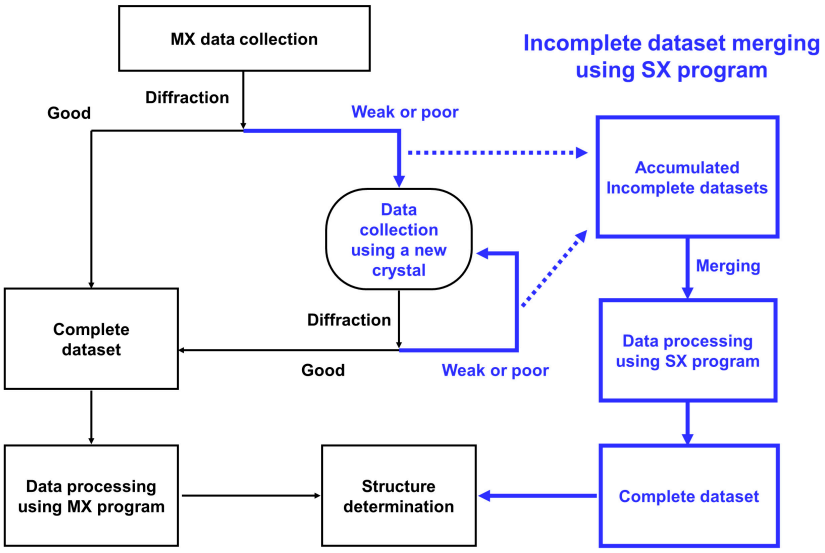

Figure 1. Process flow of the incomplete MX dataset using the SX program. The flow for generating the complete dataset from incomplete MX data using the SX program is indicated in blue.

The capability of the SX data processing techniques in merging multiple incomplete datasets collected from MX into complete datasets was examined. The diffraction datasets of

TsaBgl were used as a model sample. A new dataset of diffraction data for TsaBgl, consisting of 300 images, was collected. The processing results indicated that TsaBgl belonged to the $P2_12_12_1$ space group and that a complete dataset was processed to a resolution of 1.8 Å. However, only 50 images (Data I) were used for further processing to demonstrate the generation of a complete dataset from incomplete data. All 50 images were successfully indexed using CrystFEL, which uses the MOSFLM algorithm, achieving an indexing rate of 100%. Then, the resultant Data I was processed to a resolution of 1.55 Å, with an overall completeness of 62.93% (outer shell: 38.37%), redundancy of 3.7 (2.4), signal-to-noise ratio (SNR) of 12.43 (5.40), correlation coefficient (CC) of 0.8670 (0.0895), correlation coefficient with a correction factor (CC*) of 0.9023 (0.4054), and R_{split} of 43.15 (114.38). Although the overall completeness was less than 63%, an MR solution was successfully obtained from Data I. Structure refinement revealed R_{work} and R_{free} to be 0.2568 and 0.2958, respectively, which are relatively high because of the incomplete dataset in terms of completeness and other statistics.

Then, the generation of a complete dataset using multiple incomplete datasets was demonstrated by merging the incomplete Data I with previously collected datasets (Data II–VI). All of the datasets shared an identical $P2_12_12_1$ space group and similar unit cell dimensions (Table 1). Each of the Data II–VI datasets was originally intended to produce a complete dataset for structural determination; however, only 50 images were selected from each as incomplete data. Then, all datasets were processed using CrystFEL with the MOSFLM algorithm, resulting in the successful indexing of all 50 images and achieving a 100% indexing rate. Each dataset was processed to a resolution of 1.55 Å, with an overall completeness of 40.64–57.83% (19.75–45.32%), redundancy of 5.0–3.1 (2.9–2.3), SNR of 12.02–4.83 (5.4–0.67, CC of 0.7992–0.4844 (0.3266–0.0415), CC* of 0.9425–0.8079 (0.7017–nan), and R_{split} of 58.08–35.78 (102.45–1396.39) for Data II–VI. Although the completeness of Data II–VI was lower than 58%, an MR solution was successfully obtained from each dataset. Structure refinement revealed that the R_{work} and R_{free} values for Data II–VI ranged from 0.2472–0.3347 (0.3694–0.4770) and 0.2956–0.4345 (0.4112–0.5660), respectively, which were relatively high due to the incomplete datasets. As a result, the 50 images from each dataset were incomplete in terms of data processing and did not provide reliable three-dimensional structures with high R-values.

Table 1. Data processing statistic of the merged and the multiple TsaBgl datasets processed using MOSFLM.

Data Collection	Merge	Data I	Data II	Data III	Data IV	Data V	Data VI
Image number	300	50	50	50	50	50	50
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell [1]							
a	65.21	64.68	64.97	65.06	64.75	65.04	64.93
b	71.12	71.54	70.94	71.21	71.02	70.95	70.81
c	99.49	98.67	98.87	99.15	98.87	99.03	98.99
Resolution (Å)	100–1.55 (1.60–1.55)	100–1.55 (1.60–1.55)	65.35–1.55 (1.60–1.55)	65.35–1.55 (1.60–1.55)	70.92–1.55 (1.60–1.55)	57–80–1.55 (1.60–1.55)	99–1.55 (1.60–1.55)
Reflections	67,418 (7112)	42,719 (2564)	39,256 (3029)	31,027 (1893)	30,820 (1764)	27,583 (1320)	38,489 (2320)
Completeness (%)	99.32 (97.28)	62.93 (38.37)	57.83 (45.32)	45.71 (28.33)	45.40 (26.40)	40.64 (19.75)	56.70 (34.71)
Redundancy	14.3 (7.2)	3.7 (2.4)	5.0 (2.9)	3.7 (2.5)	4.0 (2.4)	3.1 (2.3)	3.8 (2.5)
SNR	3.42 (1.26)	12.43 (5.40)	4.83 (3.52)	9.46 (5.40)	12.02 (3.26)	10.36 (−0.67)	6.59 (1.93)
CC	0.8547 (0.2974)	0.867 −0.0895	0.4844 −0.2932	0.6183 −0.3266	0.6135 −0.2852	0.5143 (−0.0415)	0.7992 (0.0130)
CC*	0.9600 (0.6771)	0.9023 (0.4054)	0.8079 (0.6734)	0.8741 (0.7017)	0.8720 (0.6662)	0.838 (-nan)	0.9425 (0.1603)
R_{split}	28.99 −112.75	43.15 −114.38	58.08 −102.45	45.42 −120.92	48.12 −129.36	53.69 −211.49	35.78 (1396.39)

Table 1. *Cont.*

Data Collection	Merge	Data I	Data II	Data III	Data IV	Data V	Data VI
MR solution							
Top LLG	21,081.48	10,387.76	9237.271	7654.735	6965.579	4734.598	7238.53
Top TFZ	77.7	66.7	63.4	60.1	83.9	51.3	65.5
Refinement							
Resolution (Å)	49.74–1.55 (1.60–1.55)	48.06–1.55 (1.60–1.55)	48.06–1.55 (1.60–1.55)	48.06–1.55 (1.60–1.55)	48.06–1.55 (1.60–1.55)	43.28–1.55 (1.59–1.55)	49.74–1.55 (1.60–1.55)
Completeness (%)	99.1	61.99	57.38	45.1	44.7	38.14	54.48
R_{work}	0.2035 (0.3673)	0.2568 (0.4359)	0.2472 (0.3811)	0.2651 (0.3694)	0.2840 (0.3788)	0.3347 (0.4298)	0.2877 (0.4770)
R_{free}	0.2337 (0.3532)	0.2958 (0.6115)	0.2956 (0.4466)	0.3231 (0.4811)	0.3433 (0.4644)	0.4345 (0.5660)	0.3368 (0.4112)
R.M.S.D							
Bonds (Å)	0.007	0.007	0.007	0.007	0.007	0.009	0.008
Angles (°)	0.956	0.0854	0.879	0.985	0.934	1.097	0.96
B-factor (Å^2)							
Protein	20.07	20.84	12.43	17.93	18.1	22.12	26.36
Water	30.47	26.04	18.55	22.92	20.96	19.16	28.82
Ligands	22.62	21.1	15.09	18.12	18.56	18.58	25.38
Ramachandran							
Favored	97.29	97.29	97.06	95.02	94.57	94.34	96.15
Allowed	2.49	2.71	2.94	4.98	5.43	5.66	3.62
Outliers	0.23	0	0	0	0	0	
PDB code [2]	9K72	9K73	8WFV	8WFW	8XPC	8XPD	8XPE

[1] The unit cell dimensions for Data I–VI were obtained from the original structure processed using the complete dataset. [2] The PDB codes for Data I–VI correspond to complete datasets. Values for the outer shell are given in parentheses.

Next, all 300 images from Data I–VI were merged using CrystFEL with the MOSFLM algorithm (Table 1). All 300 images were successfully indexed, achieving a 100% indexing rate. The data were processed to a resolution of 1.55 Å with an overall completeness of 99.32% (97.28%), redundancy of 14.3 (7.2), SNR of 3.42 (1.26), CC of 0.8547 (0.2974), CC* of 0.9600 (0.6771), and R_{split} of 28.99 (112.75) for the merged data. This result indicated that the data processing statistics for the multiple datasets improved, with the exception of the SNR. The MR results of the merged data exhibited higher top log-likelihood gain (LLG) and top translation function Z (TFZ) values of 21081.482 and 77.7, respectively, than the individual Data I–VI. Additionally, structure refinement revealed that the merged data had lower R_{work} and R_{free} values of 0.2035 (0.3673) and 0.2337 (0.3532), respectively, than the individual Data I–VI. These results demonstrated that the merging of multiple incomplete datasets into a single complete dataset using CrystFEL was successful, enabling the determination of a crystal structure with reliable statistics for data processing and structure refinement.

3.2. Comparison of the Complete Merged and Incomplete Datasets

The statistics of the merged dataset were analyzed by comparing the data collection statistics and refinement statistics of the complete merged data with those of the incomplete Data I–VI (Figure 2). The theoretically possible number of total reflections at 1.55 Å resolution was 67,880. Merging data involves processing more images than the incomplete datasets and, therefore, containing information on a greater number of diffraction features, such as Bragg peaks, along with higher redundancy (Figure 2A–C). In addition, the number of reflections in the merged dataset was 67,418, which was 1.57–2.44 times higher than that in different incomplete datasets (Figure 2B). The completeness of Data I–VI was lower than 62.93 (45.32), whereas that of the merged data was 99.32 (97.28), indicating that the completeness was significantly improved by merging using CrystFEL (Figure 2D). Other statistical values, such as redundancy, CC, CC*, and R_{split} for the merged dataset, were improved in all processed resolution regions by merging Data I–VI, meeting the general standards in MX (Figure 2E–G). In contrast, the SNR of the merged data in all the resolution

regions was lower than that of Data I–VI, indicating that SNR was not improved by the merging of the incomplete datasets (Figure 2H).

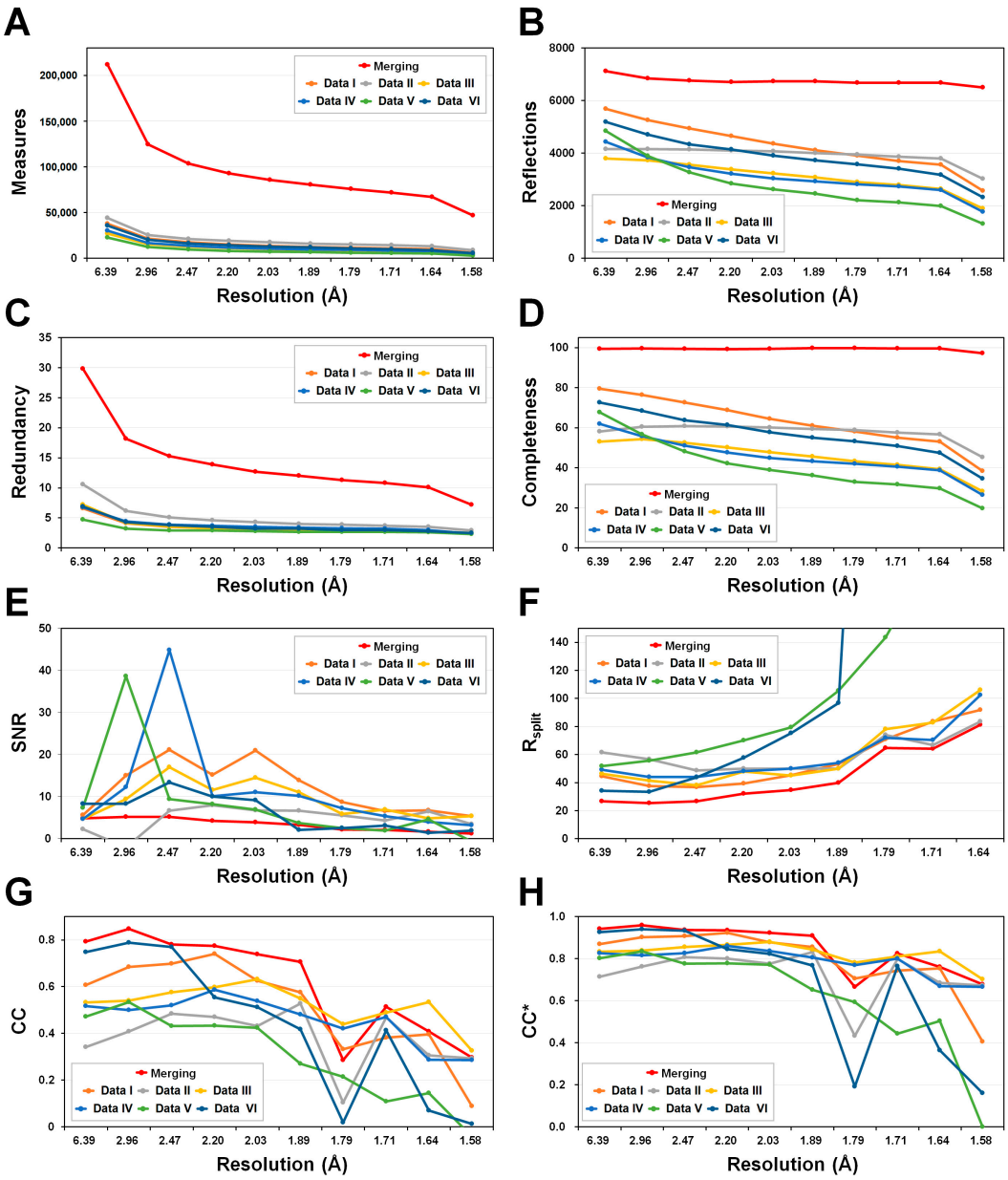

Figure 2. The profile of the data processing statistics of the merged and incomplete (Data I–VI) TsaBgl datasets. (**A**) Measures. (**B**) Reflections. (**C**) Redundancy. (**D**) Completeness. (**E**) SNR. (**F**) R_{split}. (**G**) CC. (**H**) CC*.

According to structure refinement, the electron density maps for the merged dataset had high resolution for the entire amino acid sequence of TsaBgl, whereas the electron density maps for some incomplete datasets had poor resolution due to low completeness (Figure 3). Meanwhile, the merged dataset had lower R_{work} and R_{free} values for all resolution regions, whereas the R-values for each incomplete dataset were relatively high because of lower data collection quality in terms of completeness, CC, and other factors (Table 1). This result suggested that the merged dataset processed by CrystFEL provided reliable structural information in terms of R-values. Meanwhile, the B-factor of proteins and water molecules in the merged dataset was 20.07 and 30.47, respectively, higher than those in the incomplete datasets (see Section 4).

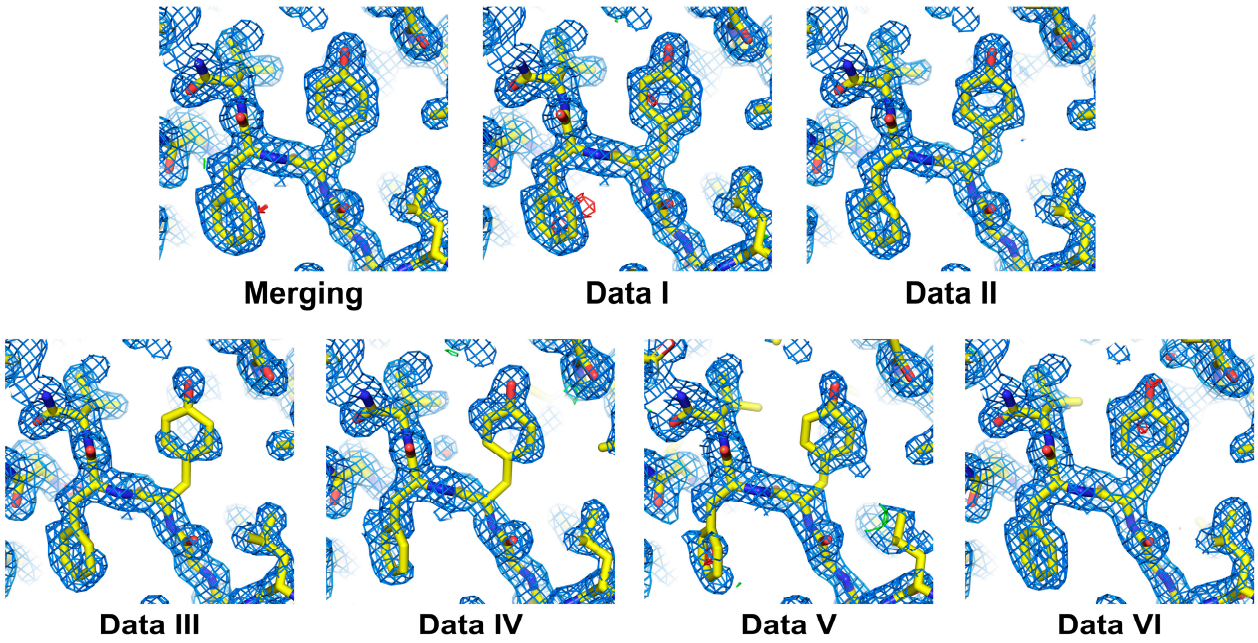

Figure 3. 2Fo-Fc (marine mesh: 1.2σ) and Fo-Fc (green mesh: +3σ; red mesh: −3σ) electron density maps of the merged complete and incomplete TsaBgl datasets from Data I–VI. The electron density maps for the aromatic rings of the tyrosine and phenylalanine residues from the incomplete datasets (Data II–V) were relatively poor.

3.3. Indexing Algorithm

The indexing efficiency of the diffraction patterns and data processing statistics, such as completeness, SNR, CC, and R_{split}, are influenced by the indexing algorithm used [49]. The impact of the indexing algorithm on the processing of multiple datasets was examined by processing 300 images from six incomplete TsaBgl datasets (Data I–VI) using different indexing algorithms, MOSFLM, DirAx, Taketwo, and XGANDALF, in CrystFEL with identical indexing parameters. The indexing results showed that both MOSFLM and XGANDALF indexed all 300 images, achieving a 100% indexing rate. In contrast, DirAx and Taketwo processed 271 and 258 images, respectively, resulting in indexing rates of 90.3% and 86.0%. Accordingly, the results of data processing by MOSFLM, Taketwo, DirAx, and XGANDAF revealed different values for the data processing statistics (Table 2 and Figure 4). The profile analysis of the data processing statistics, including completeness, CC, CC*, and R_{split}, exhibited similarities across all the resolution ranges (Figure 4). The absolute values of these data processing statistics varied among the indexing algorithms, but the overall values for completeness, CC, and R_{split} of the merged TsaBgl data processed by the four algorithms indicated successful merging to produce a complete dataset. Lastly, all datasets provided successful MR solutions, with top LLG values of 18,878–21,667 and top TFZ values of 77.0–77.8 (Table 2).

Table 2. Statistics of the merging of multiple TsaBgl datasets using various indexing algorithms.

Data Collection	MOSFLM	DirAx	Taketwo	XGANDALF
Number of images	300	300	300	300
Indexed images	300	271	258	300
Resolution (Å)	100–1.55 (1.60–1.55)	100–1.55 (1.60–1.55)	100–1.55 (1.60–1.55)	100–1.55 (1.60–1.55)
Reflections	67,418 (7112)	67,135 (6365)	67,044 (6331)	67,449 (6511)
Completeness (%)	99.32 (97.28)	98.90 (95.24)	98.77 (94.73)	99.36 (97.43)
Redundancy	14.3 (7.2)	12.7 (6.5)	13.4 (6.9)	14.4 (7.3)
SNR	3.42 (1.26)	3.39 (1.12)	3.03 (0.97)	3.64 (1.21)

Table 2. *Cont.*

Data Collection	MOSFLM	DirAx	Taketwo	XGANDALF
CC	0.8547 (0.2974)	0.8658 (0.2649)	0.8351 (0.2038)	0.8702 (0.2835)
CC*	0.9600 (0.6771)	0.9633 (0.6472)	0.9540 (0.5819)	0.9646 (0.6646)
R_{split}	28.99 (112.75)	29.09 (118.44)	31.97 (136.51)	28.56 (111.30)
MR solution				
Top LLG	21,081.482	20,510.825	18,878.869	21,667.57
Top TFZ	77.7	77.0	74.7	77.8
Refinement				
Resolution (Å)	49.74–1.55 (1.60–1.55)	49.74–1.55 (1.60–1.55)	49.74–1.55 (1.60–1.55)	49.74–1.55 (1.60–1.55)
Completeness (%)	99.1	98.68	98.31	99.18
R_{work}	0.2035 (0.3673)	0.2087 (0.3692)	0.2123 (0.3997)	0.1994 (0.3737)
R_{free}	0.2337 (0.3532)	0.2258 (0.3529)	0.2511 (0.4623)	0.2227 (0.4017)
R.M.S.D				
Bonds (Å)	0.007	0.008	0.007	0.008
Angles (°)	0.956	0.982	0.2511	0.989
B-factor (Å^2)				
Protein	20.07	20.92	20.95	20.37
Water	30.47	31.21	31.51	31.56
Ligands	22.62	22.27	22.71	21.85
Ramachandran				
Favored	97.29	97.51	97.51	97.51
Allowed	2.49	2.49	2.49	2.49
Outliers	0.23	0	0	0

Values for the outer shell are given in parentheses.

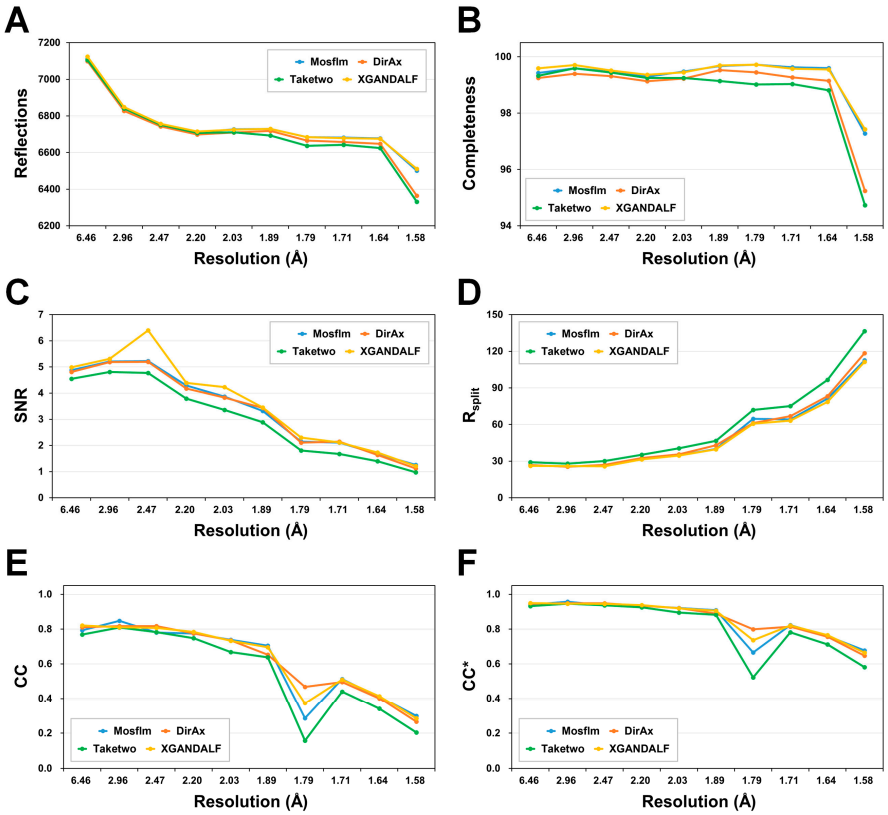

Figure 4. The profile of the data processing statistics of the merged TsaBgl datasets processed by MOSFLM, DirAx, Taketwo, and XGANDALF. (**A**) Reflections. (**B**) Completeness. (**C**) SNR. (**D**) R_{split}. (**E**) CC and (**F**) CC*.

The MR solutions were further used for structure refinement, resulting in a clear electron density map for the entire amino acid sequence of TsaBgl (Figure 5). Meanwhile, the structure refinement results from processing by the four different indexing algorithms displayed a slight variation in the R_{work} and R_{free} values, ranging from 19.94% to 21.23% and 22.27% to 25.11%, respectively (Table 2). Although all of the R-values were acceptable according to the standard criteria for structure determination in MX, the detailed quality of the final model structure could be affected by the choice of the indexing algorithm. Additionally, the B-factor values of the final model processed by the four different indexing algorithms were similar for TsaBgl and water, with ranges of 20.07–20.95 $Å^2$ and 30.47–31.56 $Å^2$, respectively. These results suggest that an indexing program can influence certain statistics for data collection and structure refinement. Therefore, optimizing the indexing algorithm is also important for obtaining a merged dataset of higher quality.

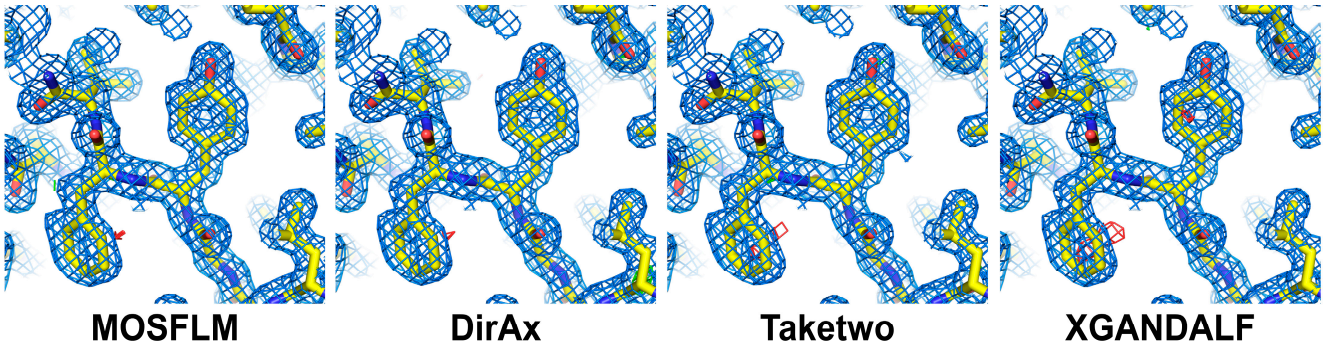

Figure 5. 2Fo-Fc (marine mesh: 1.2σ) and Fo-Fc (green mesh: +3σ; red mesh: −3σ) electron density maps of the merged TsaBgl dataset processed by MOSFLM, DirAx, Taketwo, and XGANDALF.

4. Discussion

During MX data collection, factors such as low crystal quality and radiation damage may lead to incomplete datasets with insufficient statistics, including low completeness and low CC. The merging of incomplete datasets to generate a complete dataset is useful for determining macromolecular structures. In this study, a new approach for generating a complete dataset from incomplete MX datasets using CrystFEL, an SX data processing program, was introduced.

Six incomplete TsaBgl datasets were processed and merged using CrystFEL. While processing incomplete MX diffraction datasets using CrystFEL, it is crucial to ensure the consistency of the geometry file and unit cell information. During diffraction image indexing using "*indexamajig*" in CrystFEL, it is essential to input the detector geometry information, which includes parameters such as wavelength, detector-to-crystal distance, and detector details. Since Bragg peaks are indexed based on this geometry file, if the wavelength, detector settings, or detector-to-crystal distance differs among the processed individual incomplete datasets, the dataset inconsistent with the detector geometry file information will not be indexed. Therefore, all datasets processed using CrystFEL must come from identical experimental setups related to the geometry file; otherwise, the differences in the setup will reduce indexing efficiency.

In addition, the unit cell parameters of the crystal samples must be consistent during data processing. In this study, all of the diffraction datasets of TsaBgl belonged to the same $P2_12_12_1$ space group. The unit cell parameters for the six incomplete datasets were within the range of a at 64.48–65.07 Å, b at 70.81–71.55 Å, and c at 98.67–99.16 Å. The default tolerance values for the unit cell dimensions in the SX data processing were used, with tolerances of 5% for a, b, c, and 1.5% for α, β, and γ. As a result, all unit cells fell within the indexing parameter tolerances, enabling the successful indexing of all images. However, if the differences in unit cell dimensions among the incomplete datasets exceed the range of tolerance, it is necessary to increase the range during data processing to improve the indexing efficiency. However, high tolerance in the unit cell dimensions can degrade

the quality of the processed data. Therefore, it is advisable to merge only good-quality datasets—if a sufficient number of incomplete datasets are available—to achieve better data quality. In this study, TsaBgl datasets were selected to demonstrate data merging with a sufficient volume of more than 6 datasets. Although the diffraction datasets were not abundant, datasets from lysozyme and glucose isomerase also worked well in the merging process using CrystFEL.

Four different indexing algorithms, i.e., MOSFLM, DirAx, Taketwo, and XGANDALF, were used to merge the incomplete datasets. Although the statistics for data processing and structure refinement varied according to the indexing algorithm used, all algorithms successfully performed data merging and structure determination. In this study, the default parameters provided by *indexamajig* in CrystFEL for data processing were used, and parameter optimization to enhance the efficiency of each indexing algorithm was not performed. Therefore, the most notable finding of the study is not that the indexing efficiency of the algorithms differed, but that CrystFEL could successfully merge incomplete datasets using various indexing algorithms. Additionally, during data processing with CrystFEL, other indexing algorithms, such as XDS [16], FELIX [50], and pinkIndexer [51], can also be used to merge incomplete datasets and generate a complete diffraction dataset.

The complete dataset generated through SX data processing achieved reliable statistical values, such as completeness, CC, CC*, R_{split}, and redundancy, compared to the individual incomplete datasets. In contrast, the merged dataset had lower SNR values than the incomplete datasets, indicating that SNR was not a parameter that could be improved simply by increasing the amount of data. Meanwhile, the relatively high SNR values provided by the individual incomplete datasets could not be considered reliable, as they reflected statistics from incomplete data processing.

Regarding structure refinement, the R_{work} and R_{free} values from the structure refinement of the merged dataset showed more reliable statistical values than the incomplete datasets. However, the B-factor values for TsaBgl and water molecules in the merged dataset were higher than those in the incomplete datasets. The B-factor values were not directly comparable because of the lower reliability of the individual incomplete datasets. However, theoretically, merged datasets, as the sum of multiple data, tended to have higher B-factors. Therefore, experimentally, merged datasets may provide higher B-factors than single-crystal diffraction data; thus, they may not provide entirely reliable structural information regarding protein flexibility or water molecule positions. Consequently, it would be more rational to focus on the molecular biological function of the determined structure rather than protein flexibility during structure determination using SX data merging.

During the revision, data merging was performed using XDS and XSCALE (Table 3). Several data processing statistics, such as completeness and CC1/2 values generated by XDS/XSCALE, were superior to those obtained from the CrystFEL program. As previously mentioned, these differences can arise due to variations in the indexing algorithm, making a direct comparison of data processing statistics challenging. Nevertheless, since results can vary based not only on the indexing algorithm but also on the data merging program and the user's familiarity with the program, it is important to utilize and compare a variety of data merging tools to achieve a high-quality dataset.

Table 3. Data processing and merging using XDS/XSCALE.

Data Collection	Merge	Data I	Data II	Data III	Data IV	Data V	Data VI
Image number	300	50	50	50	50	50	50
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell *							
a	65.21	64.68	64.97	65.06	64.75	65.04	64.93
b	71.12	71.54	70.94	71.21	71.02	70.95	70.81
c	99.49	98.67	98.87	99.15	98.87	99.03	98.99

Table 3. *Cont.*

Data Collection	Merge	Data I	Data II	Data III	Data IV	Data V	Data VI
Resolution (Å)	50–1.55 (1.60–1.55)	50–1.55 (1.60–1.55)	50–1.55 (1.60–1.55)	50–1.55 (1.60–1.55)	50–1.55 (1.60–1.55)	50–1.55 (1.60–1.55)	50–1.55 (1.60–1.55)
Reflections	67,344	58,833	43,668	43,561	49,882	57,921	58,353
Redundancy	10.56 (8.02)	2.00 (1.85)	2.77 (2.52)	2.77 (2.53)	2.36 (2.14)	2.04 (1.89)	1.98 (1.81)
Completeness (%)	99.3 (94.0)	87.4 (85.5)	64.8 (63.8)	64.3 (63.3)	74.7 (75.0)	85.6 (84.5)	86.9 (86.4)
I/sigma	9.57 (3.10)	12.63 (2.44)	9.00 (3.91)	8.03 (2.37)	13.12 (3.25)	6.13 (1.18)	8.49 (1.37)
R-factor	0.152 (0.623)	0.047 (0.313)	0.082 (0.204)	0.095 (0.443)	0.057 (0.325)	0.114 (0.762)	0.068 (0.551)
R-meas	0.160 (0.668)	0.060 (0.417)	0.097 (0.247)	0.111 (0.534)	0.069 (0.401)	0.147 (1.005)	0.087 (0.729)
CC1/2 (%)	99.3 (79.7)	99.8 (83.3)	99.2 (93.5)	99.4 (79.4)	99.8 (85.4)	99.1 (39.8)	99.7 (61.8)

* Merge dataset was generated by XSCALE.

In summary, a new approach for merging incomplete diffraction datasets from MX into a complete dataset using an SX program was explored and validated. The successful data merging approach using SX programs expands the applicability of SX programs in MX data processing. This approach can be beneficial for structure determination using multiple incomplete datasets from MX.

5. Conclusions

During MX data collection, incomplete diffraction datasets with low completeness may sometimes be obtained due to low diffraction intensity or poor data quality. To determine the crystal structure of macromolecules using such incomplete datasets, an approach has been introduced that merges these datasets using serial crystallography (SX) programs. This study demonstrates that SX programs can effectively merge incomplete MX datasets, enabling the successful generation of a complete dataset and determination of the crystal structure. This method expands the potential for determining crystal structures in MX when incomplete datasets are collected, along with existing data-merging tools.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cryst14121012/s1, Table S1: Resource of the diffraction images used for data processing in this study.; Table S2: Data processing script and information using CrystFEL.; Table S3: Data collection and refinement statistics for TsaBgl from the complete dataset of Data I.

Funding: This work was funded by the National Research Foundation of Korea (NRF) (NRF-2021R1I1A1A01050838).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The structure factors and coordinates have been deposited in the Protein Data Bank under the accession codes 9K72 (merged data from Data I–VI dataset) and 9K73 (Data I: 300 images). X-ray diffraction images for Data I were deposited in the ZENODO (doi: 10.5281/zenodo.13948740).

Acknowledgments: I would like to thank the beamline staff at the 11C beamline at the Pohang Accelerator Laboratory for their assistance with data collection. The author thanks the Global Science experimental Data hub Center (GSDC) at the Korea Institute of Science and Technology Information (KISTI) for providing computing resources and technical support.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Smyth, M.S.; Martin, J.H. X Ray crystallography. *Mol. Pathol.* **2000**, *53*, 8–14. [CrossRef] [PubMed]
2. Shi, Y. A Glimpse of Structural Biology through X-Ray Crystallography. *Cell* **2014**, *159*, 995–1014. [CrossRef] [PubMed]

3. Zheng, H.; Hou, J.; Zimmerman, M.D.; Wlodawer, A.; Minor, W. The future of crystallography in drug discovery. *Expert Opin. Drug Discov.* **2013**, *9*, 125–137. [CrossRef] [PubMed]
4. Cooper, D.R.; Porebski, P.J.; Chruszcz, M.; Minor, W. X-ray crystallography: Assessment and validation of protein-small molecule complexes for drug discovery. *Expert Opin. Drug Discov.* **2011**, *6*, 771–782. [CrossRef] [PubMed]
5. Roda, S.; Robles-Martín, A.; Xiang, R.; Kazemi, M.; Guallar, V. Structural-Based Modeling in Protein Engineering. A Must Do. *J. Phys. Chem. B* **2021**, *125*, 6491–6500. [CrossRef]
6. Ovchinnikov, S.; Huang, P.-S. Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **2021**, *65*, 136–144. [CrossRef]
7. Ni, T.W.; Tofanelli, M.A.; Ackerson, C.J. Structure Determination by Single Crystal X-ray Crystallography. In *Protected Metal Clusters—From Fundamentals to Applications*; Frontiers of Nanoscience; Elsevier: Amsterdam, The Netherlands, 2015; pp. 103–125.
8. Bernstein, H.J.; Förster, A.; Bhowmick, A.; Brewster, A.S.; Brockhauser, S.; Gelisio, L.; Hall, D.R.; Leonarski, F.; Mariani, V.; Santoni, G.; et al. Gold Standard for macromolecular crystallography diffraction data. *IUCrJ* **2020**, *7*, 784–792. [CrossRef]
9. Heras, B.; Martin, J.L. Post-crystallization treatments for improving diffraction quality of protein crystals. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61*, 1173–1180. [CrossRef]
10. Abe, M.; Suzuki, R.; Kojima, K.; Tachibana, M. Evaluation of crystal quality of thin protein crystals based on the dynamical theory of X-ray diffraction. *IUCrJ* **2020**, *7*, 761–766. [CrossRef]
11. Nam, K.H. Effects of Radiation Damage on Metal-Binding Sites in Thermolysin. *Crystals* **2024**, *14*, 876. [CrossRef]
12. Aller, P.; Geng, T.; Evans, G.; Foadi, J. Applications of the BLEND Software to Crystallographic Data from Membrane Proteins. In *The Next Generation in Membrane Protein Structure Determination*; Advances in Experimental Medicine and Biology; Springer: Berlin/Heidelberg, Germany, 2016; pp. 119–135.
13. Evans, P.R.; Murshudov, G.N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 1204–1214. [CrossRef] [PubMed]
14. Foadi, J.; Aller, P.; Alguel, Y.; Cameron, A.; Axford, D.; Owen, R.L.; Armour, W.; Waterman, D.G.; Iwata, S.; Evans, G. Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 1617–1632. [CrossRef] [PubMed]
15. Yamashita, K.; Hirata, K.; Yamamoto, M. KAMO: Towards automated data processing for microcrystals. *Acta Crystallogr. D Struct. Biol.* **2018**, *74*, 441–449. [CrossRef]
16. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 125–132. [CrossRef]
17. Kabsch, W. Processing of X-ray snapshots from crystals in random orientations. *Acta Crystallogr. D Biol.* **2014**, *70*, 2204–2216. [CrossRef]
18. Akey, D.L.; Terwilliger, T.C.; Smith, J.L. Efficient merging of data from multiple samples for determination of anomalous substructure. *Acta Crystallogr. D Struct. Biol.* **2016**, *72*, 296–302. [CrossRef]
19. Martin-Garcia, J.M.; Conrad, C.E.; Coe, J.; Roy-Chowdhury, S.; Fromme, P. Serial femtosecond crystallography: A revolution in structural biology. *Arch. Biochem. Biophys.* **2016**, *602*, 32–47. [CrossRef] [PubMed]
20. Nam, K.H. Guide to serial synchrotron crystallography. *Curr. Res. Struct. Biol.* **2024**, *7*, 100131. [CrossRef]
21. Mehrabi, P.; Bücker, R.; Bourenkov, G.; Ginn, H.M.; von Stetten, D.; Müller-Werkmeister, H.M.; Kuo, A.; Morizumi, T.; Eger, B.T.; Ou, W.L.; et al. Serial femtosecond and serial synchrotron crystallography can yield data of equivalent quality: A systematic comparison. *Sci. Adv.* **2021**, *7*, eabf1380. [CrossRef]
22. Hekstra, D.R. Emerging Time-Resolved X-Ray Diffraction Approaches for Protein Dynamics. *Annu. Rev. Biophys.* **2023**, *52*, 255–274. [CrossRef]
23. Westenhoff, S.; Meszaros, P.; Schmidt, M. Protein motions visualized by femtosecond time-resolved crystallography: The case of photosensory vs photosynthetic proteins. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102481. [CrossRef] [PubMed]
24. Park, J.; Nam, K.H. Recent chemical mixing devices for time-resolved serial femtosecond crystallography. *TrAC Trends Anal. Chem.* **2024**, *172*, 117554. [CrossRef]
25. Henkel, A.; Oberthür, D. A snapshot love story: What serial crystallography has done and will do for us. *Acta Crystallogr. D Struct. Biol.* **2024**, *80*, 563–579. [CrossRef] [PubMed]
26. Hough, M.A.; Owen, R.L. Serial synchrotron and XFEL crystallography for studies of metalloprotein catalysis. *Curr. Opin. Struct. Biol.* **2021**, *71*, 232–238. [CrossRef]
27. Park, J.; Nam, K.H. Sample Delivery Systems for Serial Femtosecond Crystallography at the PAL-XFEL. *Photonics* **2023**, *10*, 557. [CrossRef]
28. Zhao, F.Z.; Zhang, B.; Yan, E.K.; Sun, B.; Wang, Z.J.; He, J.H.; Yin, D.C. A guide to sample delivery systems for serial crystallography. *FEBS J.* **2019**, *286*, 4402–4417. [CrossRef]
29. Grünbein, M.L.; Nass Kovacs, G. Sample delivery for serial crystallography at free-electron lasers and synchrotrons. *Acta Crystallogr. D Biol. Crystallogr.* **2019**, *75*, 178–191. [CrossRef]
30. Martiel, I.; Muller-Werkmeister, H.M.; Cohen, A.E. Strategies for sample delivery for femtosecond crystallography. *Acta Crystallogr. D Struct. Biol.* **2019**, *75*, 160–177. [CrossRef]
31. White, T.A.; Barty, A.; Stellato, F.; Holton, J.M.; Kirian, R.A.; Zatsepin, N.A.; Chapman, H.N. Crystallographic data processing for free-electron laser sources. *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69*, 1231–1240. [CrossRef]

32. White, T.A.; Mariani, V.; Brehm, W.; Yefanov, O.; Barty, A.; Beyerlein, K.R.; Chervinskii, F.; Galli, L.; Gati, C.; Nakane, T.; et al. Recent developments in CrystFEL. *J. Appl. Crystallogr.* **2016**, *49*, 680–689. [CrossRef]
33. Lyubimov, A.Y.; Uervirojnangkoorn, M.; Zeldin, O.B.; Brewster, A.S.; Murray, T.D.; Sauter, N.K.; Berger, J.M.; Weis, W.I.; Brunger, A.T. IOTA: Integration optimization, triage and analysis tool for the processing of XFEL diffraction images. *J. Appl. Crystallogr.* **2016**, *49*, 1057–1064. [CrossRef]
34. Li, X.; Li, C.; Liu, H. ClickX: A visualization-based program for preprocessing of serial crystallography data. *J. Appl. Crystallogr.* **2019**, *52*, 674–682. [CrossRef] [PubMed]
35. Boutet, S.; Lomb, L.; Williams, G.J.; Barends, T.R.M.; Aquila, A.; Doak, R.B.; Weierstall, U.; DePonte, D.P.; Steinbrener, J.; Shoeman, R.L.; et al. High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. *Science* **2012**, *337*, 362–364. [CrossRef]
36. Nam, K.H. The Conformational Change of the L3 Loop Affects the Structural Changes in the Substrate Binding Pocket Entrance of β-Glucosidase. *Molecules* **2023**, *28*, 7807. [CrossRef] [PubMed]
37. Nam, K.H. Structural analysis of Tris binding in β-glucosidases. *Biochem. Biophys. Res. Commun.* **2024**, *700*, 149608. [CrossRef]
38. Kim, I.J.; Bornscheuer, U.T.; Nam, K.H. Biochemical and Structural Analysis of a Glucose-Tolerant β-Glucosidase from the Hemicellulose-Degrading *Thermoanaerobacterium saccharolyticum*. *Molecules* **2022**, *27*, 290. [CrossRef] [PubMed]
39. Park, S.Y.; Ha, S.C.; Kim, Y.G. The Protein Crystallography Beamlines at the Pohang Light Source II. *Biodesign* **2017**, *5*, 30–34.
40. Otwinowski, Z.; Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **1997**, *276*, 307–326. [CrossRef]
41. Battye, T.G.; Kontogiannis, L.; Johnson, O.; Powell, H.R.; Leslie, A.G. iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 271–281. [CrossRef]
42. Yefanov, O.; Mariani, V.; Gati, C.; White, T.A.; Chapman, H.N.; Barty, A. Accurate determination of segmented X-ray detector geometry. *Opt. Express* **2015**, *23*, 28459–28470. [CrossRef]
43. Duisenberg, A.J.M. Indexing in Single-Crystal Diffractometry with an Obstinate List of Reflections. *J. Appl. Crystallogr.* **1992**, *25*, 92–96. [CrossRef]
44. Ginn, H.M.; Roedig, P.; Kuo, A.; Evans, G.; Sauter, N.K.; Ernst, O.P.; Meents, A.; Mueller-Werkmeister, H.; Miller, R.J.; Stuart, D.I. TakeTwo: An indexing algorithm suited to still images with known crystal parameters. *Acta Crystallogr. D Struct. Biol.* **2016**, *72*, 956–965. [CrossRef] [PubMed]
45. Gevorkov, Y.; Yefanov, O.; Barty, A.; White, T.A.; Mariani, V.; Brehm, W.; Tolstikova, A.; Grigat, R.R.; Chapman, H.N. XGANDALF—Extended gradient descent algorithm for lattice finding. *Acta Crystallogr. A Found. Adv.* **2019**, *75*, 694–704. [CrossRef]
46. Liebschner, D.; Afonine, P.V.; Baker, M.L.; Bunkoczi, G.; Chen, V.B.; Croll, T.I.; Hintze, B.; Hung, L.W.; Jain, S.; McCoy, A.J.; et al. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr. D Struct. Biol.* **2019**, *75*, 861–877. [CrossRef] [PubMed]
47. Emsley, P.; Cowtan, K. Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60*, 2126–2132. [CrossRef] [PubMed]
48. Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.J.; Chen, V.B.; et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **2018**, *27*, 293–315. [CrossRef]
49. Nam, K.H. Processing of Multicrystal Diffraction Patterns in Macromolecular Crystallography Using Serial Crystallography Programs. *Crystals* **2022**, *12*, 103. [CrossRef]
50. Beyerlein, K.R.; White, T.A.; Yefanov, O.; Gati, C.; Kazantsev, I.G.; Nielsen, N.F.; Larsen, P.M.; Chapman, H.N.; Schmidt, S. FELIX: An algorithm for indexing multiple crystallites in X-ray free-electron laser snapshot diffraction images. *J. Appl. Crystallogr.* **2017**, *50*, 1075–1083. [CrossRef]
51. Gevorkov, Y.; Barty, A.; Brehm, W.; White, T.A.; Tolstikova, A.; Wiedorn, M.O.; Meents, A.; Grigat, R.-R.; Chapman, H.N.; Yefanov, O. pinkIndexer—A universal indexer for pink-beam X-ray and electron diffraction snapshots. *Acta Crystallogr. A Found. Adv.* **2020**, *76*, 121–131. [CrossRef]