



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



# **Advancements in Diffraction Analysis Methods and Data Reduction Techniques for Serial Crystallography**

## **Dissertation**

zur Erlangung des Doktorgrades  
an der Fakultät für Mathematik, Informatik und Naturwissenschaften  
Fachbereich Physik  
der Universität Hamburg

vorgelegt von

**Marina Galchenkova**

Hamburg

2024



Gutachter der Dissertation:	Prof. Dr. Henry N. Chapman Dr. Oleksandr Yefanov
Zusammensetzung der Prüfungskommission:	Prof. Dr. Henry N. Chapman Dr. Oleksandr Yefanov Prof. Dr. Arwen Pearson Dr. Thomas A. White Prof. Dr. Nina Rohringer
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Arwen Pearson
Datum der Disputation:	05.03.2024
Vorsitzender Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Markus Drescher
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Wolfgang J. Parak
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

## Abstract

Proteins play a crucial role in living cells. Their functions are determined by their three-dimensional (3D) structure. This atomic-scale structure is usually investigated by crystallography using X-ray sources such as an X-ray tube, synchrotron or Free Electron Laser (FEL). The conventional approach to macromolecular crystallography (MX) is to acquire diffraction patterns from a crystal as it is rotated about one or more axes to get the full 3D diffraction volume of the studied crystal. The total X-ray exposure of the crystal is limited by the accumulation of damage to the protein structure and crystal lattice by ionising radiation. Cryogenic cooling reduces the processes of radiolysis and extends the dose that can be tolerated. However, such cooling may alter the macromolecular structure and prevent the ability to measure dynamical processes by time-resolved methods.

For efficient measurement at room temperature (RT) and investigating fast protein dynamics, serial crystallography (SX) comes into play. In this method, the studied crystals' 3D diffraction volume (reciprocal space) is merged from still diffraction patterns collected from small randomly oriented crystals exposed by X-rays. This technique must be capable of assembling a complete three-dimensional dataset of structure factor moduli using a large number of individual still diffraction patterns. SX enables a wide range of experiments, including measurements at room temperature, time-resolved studies on biological crystals, measuring sub-micron-sized crystals, and obtaining structures of radiation-sensitive proteins. Known problems in serial crystallography are the high threshold to enter the field, the lack of a user-friendly data processing pipeline, and the huge amount of data that must be processed and reduced to get the structure of the studied protein. This dissertation is dedicated to developing solutions for addressing the issues mentioned above.

Recent advancements in X-ray facilities, including 4th generation synchrotrons and FELs, in combination with state-of-the-art X-ray detectors, have enabled conducting SX experiments at a remarkable rate, capturing more than 1000 images per second. However, this increased acquisition rate comes with a trade-off - an enormous volume of data, with some experiments already yielding up to 5 PB of measured data. As a result, novel data reduction strategies need to be developed and implemented to handle this vast amount of information efficiently. The most common method to reduce the size of the measured data is the usage of lossless compression. The compression rate and speed of different compression algorithms available for the HDF5 library were checked using different datasets. This extensive evaluation demonstrated that lossless compression methods maintain the original data without any alteration but cannot achieve a high compression ratio. Thus, some lossy compression and data reduction are needed. For this reason, the following approaches were successfully tested on different datasets: binning, quantisation (including quantisation using a non-uniform step), and non-hits rejection. Also, it was shown that such approaches as measuring less data or storing data within the area of identified Bragg peaks in a diffraction pattern may lead to data quality degradation and, therefore, are not recommended for general use.

A set of data metrics capable of assessing the loss of information due to applying various compression schemes is used to evaluate the effect of any lossy compression schemes. Different data quality metrics are described and used for testing various data reduction schemes. A proper way to use each quality metric is also described in detail.

Notably, non-hits rejection and binning process automation have been successfully implemented into the routine data processing pipeline and tested on data collected with the TapeDrive sample-delivery method at the P11 beamline, PETRA III. Furthermore, the presented non-uniform quantisation compression technique holds potential for application in other datasets, including electron or neutron diffraction.

The enormous amount of measured data poses another challenge: it cannot be processed manually. Instead, an auto-processing pipeline has to be developed. Considering how the crystals are measured in MX and SX, the data analysis techniques differ for those two methods. Therefore, the existing pipelines used for MX are hardly applicable to the SX data. Despite significant progress in this field for SX over the past decade, establishing a universal, reliable processing pipeline compatible with different sample delivery



systems remains a complex challenge. This dissertation aims to develop a well-established, robust and universally applicable data processing pipeline for SX, which constitutes the generation of various figures of merit and compiling overall statistics for proper data evaluation at each stage of data processing and for publishing purposes. Multiple experiments at FELs and synchrotrons were processed during the work on the dissertation, and some of the results are presented to illustrate the benefits of using the developed algorithms. This dissertation emphasised data with observable undesirable features, such as the presence of ice rings and salt reflections. To address these issues, a special software package was developed and used as a part of the developed data processing pipeline. This automatic data processing pipeline has been implemented in the control system of a drug-screening P09 beamline, PETRA III. This dissertation outlines a strategy to optimise SSX beamtimes using fixed-target sample delivery methods like chips. The approach involves two key steps: initially, a rapid raster scan of the chip identifies crystal positions via diffraction, followed by measuring a rotational series at these positions within a small range of angles. This method efficiently avoids empty positions during data acquisition, saving precious beam time and reducing data volume. It is particularly effective when the chip has few crystals, common with challenging-to-crystallise proteins. This approach is critical for maximising crystal utilisation and enhancing the likelihood of successfully determining protein structures.

The dissertation contributes to the advancement of serial crystallography by establishing a reliable data processing and reduction framework, ensuring the reproducibility and reliability of obtained final results. Developed strategies open up new possibilities for carrying out the experiments in an efficient way and overcoming the problem with data storage.

## Zusammenfassung

Proteine spielen eine entscheidende Rolle in lebenden Zellen. Ihre Funktionen werden durch ihre dreidimensionale (3D) Struktur bestimmt. Diese Struktur im atomaren Maßstab wird üblicherweise durch Kristallographie unter Verwendung von Röntgenquellen wie Röntgenröhren, Synchrotron oder Freie-Elektronen-Laser (FEL) untersucht. Der konventionelle Ansatz zur makromolekularen Kristallographie (MX) besteht darin, Beugungsmuster von einem Kristall zu erfassen, während er um eine oder mehrere Achsen gedreht wird, um das vollständige 3D-Beugungsvolumen des untersuchten Kristalls zu erhalten. Die Gesamtbelichtung des Kristalls durch Röntgenstrahlen ist durch die Anhäufung von Schäden an der Protein- und Kristallstruktur durch ionisierende Strahlung begrenzt. Die kryogene Kühlung reduziert die Prozesse der Radiolyse und erhöht die tolerierbare Dosis. Allerdings kann eine solche Kühlung die makromolekulare Struktur verändern und die Möglichkeit zur Messung dynamischer Prozesse durch zeitaufgelöste Methoden verhindern.

Für effiziente Messungen bei Raumtemperatur (RT) und die Untersuchung schneller Proteindynamiken kommt die Serielle Kristallographie (SX) ins Spiel. Bei dieser Methode wird das 3D-Beugungsvolumen (reziproker Raum) der untersuchten Kristalle aus Beugungsmustern einzelner, zufällig orientierter Kristalle zusammengeführt, die, im Gegensatz zur konventionellen MX, dabei nicht rotiert wurden. Diese Technik muss in der Lage sein, einen vollständigen dreidimensionalen Datensatz der Strukturformfaktormoduli unter Verwendung einer großen Anzahl einzelner unbewegter Beugungsmuster zu erstellen. SX ermöglicht eine Vielzahl von Experimenten, einschließlich Messungen bei Raumtemperatur, zeitaufgelöste Studien an biologischen Kristallen, Messungen von submikrometergroßen Kristallen und Bestimmung von Strukturen strahlungsempfindlicher Proteine. Bekannte Probleme in der Seriellen Kristallographie sind die hohe Einstiegshürde in das Feld, das Fehlen einer benutzerfreundlichen Datenverarbeitungspipeline und die große Menge an Daten, die verarbeitet und reduziert werden müssen, um die Struktur des untersuchten Proteins zu erhalten. Diese Dissertation widmet sich der Entwicklung von Lösungen zur Bewältigung der oben genannten Herausforderungen.

Die jüngsten Fortschritte bei Röntgenquellen, einschließlich Synchrotronanlagen der vierten Generation und FELs, in Kombination mit hochmodernen Röntgendetektoren, haben die Durchführung von SX-Experimenten in bemerkenswertem Tempo ermöglicht, wobei mehr als 1000 Bilder pro Sekunde aufgenommen werden. Allerdings geht diese erhöhte Aufnahmerate mit einem Kompromiss einher - einer enormen Datenmenge, wobei einige Experimente bereits bis zu 5 PB gemessene Daten ergeben. Daher müssen neuartige Strategien zur Datenreduktion entwickelt und implementiert werden, um diese große Informationsmenge effizient zu verarbeiten. Die gängigste Methode zur Verringerung der Größe der gemessenen Daten ist die Verwendung von verlustfreier Kompression. Die Kompressionsrate und Geschwindigkeit verschiedener Kompressionsalgorithmen, die für die HDF5-Bibliothek verfügbar sind, wurden anhand verschiedener Datensätze überprüft. Diese umfassende Bewertung zeigte, dass verlustfreie Kompressionsmethoden die Originaldaten ohne Veränderung erhalten können, jedoch keinen hohen Kompressionsgrad erreichen können. Folglich sind einige Verluste bei Kompression und Datenreduktion erforderlich. Aus diesem Grund wurden die folgenden Ansätze erfolgreich an verschiedenen Datensätzen getestet: Binning, Quantisierung (einschließlich Quantisierung mit einem nicht-uniformen Schritt) und Abweisung von Nicht-Treffern. Es wurde auch gezeigt, dass Ansätze wie das Messen weniger Daten oder das selektive Speichern der Daten innerhalb eines identifizierten Bragg-Peaks eines Beugungsmusters zu einer Qualitätsverschlechterung der Daten führen können und daher nicht für den allgemeinen Gebrauch empfohlen werden.

Zur Bewertung der Auswirkungen von Verlustkomprimierungsschemata wird eine Reihe von Datenmetriken verwendet, mit denen der Informationsverlust durch die Anwendung verschiedener Komprimierungsschemata bewertet werden kann. Unterschiedliche Qualitätsmetriken für Daten werden beschrieben und zur Prüfung verschiedener Datenaufbereitungsschemata verwendet. Eine genaue Anleitung zur Verwendung jeder Qualitätsmetrik wird ebenfalls detailliert beschrieben.

Insbesondere wurden die Prozesse der Abweisung von Nicht-Treffern und der Binning-Automatisierung erfolgreich in eine routinemäßige Datenverarbeitungspipeline implementiert und an Daten getestet, die mit

der TapeDrive-Probenzufuhrmethode am P11-Strahlrohr bei PETRA III gesammelt wurden. Darüber hinaus hat die vorgestellte nicht-uniforme Quantisierungskompressionstechnik Potenzial für die Anwendung bei anderen Datensätzen, einschließlich Elektronen- oder Neutronenbeugung.

Die enorme Menge an gemessenen Daten stellt eine weitere Herausforderung dar: Sie kann nicht manuell verarbeitet werden. Stattdessen muss eine automatische Verarbeitungspipeline entwickelt werden. Angesichts der Unterschiede in der Messung von Kristallen in MX und SX unterscheiden sich die Datenanalysetechniken für diese beiden Methoden. Daher sind die bestehenden Pipelines, die für MX verwendet werden, kaum auf die SX-Daten anwendbar. Trotz erheblicher Fortschritte auf diesem Gebiet in den letzten zehn Jahren bleibt die Entwicklung einer universellen, zuverlässigen Verarbeitungspipeline, die mit verschiedenen Probenzuführungssystemen kompatibel ist, eine komplexe Herausforderung. Diese Dissertation zielt darauf ab, eine robuste und universell anwendbare Datenverarbeitungspipeline für SX zu entwickeln, die die Erzeugung verschiedener Leistungskennzahlen und die Zusammenstellung von Gesamtstatistiken für eine ordnungsgemäße Datenbewertung in jedem Stadium der Datenverarbeitung und für Veröffentlichungszwecke umfasst. Während der Arbeit an der Dissertation wurden mehrere Experimente an Freie-Elektronen-Lasern und Synchrotronquellen mithilfe dieser Pipeline analysiert, und einige der Ergebnisse werden präsentiert, um die Vorteile der entwickelten Algorithmen zu veranschaulichen. Die Beispieldaten in dieser Dissertation verwendeten Beispieldaten zeigen Artefakte, die häufig die Qualität von Messdaten verringern, wie das Vorhandensein von Eiskegeln und Salzreflexionen. Um diese Probleme zu lösen, wurde ein spezielles Softwarepaket entwickelt und als Teil der entwickelten Datenverarbeitungspipeline verwendet. Diese automatische Datenverarbeitungspipeline wurde im Steuerungssystem der Drug-Screening-Strahllinie P09 an PETRA III implementiert. Diese Dissertation skizziert eine Strategie zur Optimierung von SSX-Strahlzeiten unter Verwendung von Fix-Target Probenzuführungsmethoden wie Chips. Der Ansatz umfasst zwei Schlüsselschritte: zunächst identifiziert ein schneller Rasterscan des Chips Kristallpositionen mittels Beugung, gefolgt von der Messung einer Rotationsreihe an identifizierten Positionen innerhalb eines kleinen Winkelbereichs. Diese Methode vermeidet effizient das Messen leere Positionen während der Datenaufnahme, spart wertvolle Strahlzeit und reduziert das Datenvolumen. Sie ist besonders effektiv, wenn der Chip nur wenige Kristalle enthält, was bei schwer zu kristallisierenden Proteinen häufig der Fall ist. Dieser Ansatz ist entscheidend, um die Kristallnutzung zu maximieren und die Wahrscheinlichkeit der erfolgreichen Bestimmung von Proteinstrukturen zu erhöhen.

Diese Dissertation zielt darauf ab, eine etablierte, robuste und universell anwendbare Datenverarbeitungspipeline für SX zu entwickeln, die die Erzeugung verschiedener Gütekriterien ermöglicht und Gesamtstatistiken für eine ordnungsgemäße Datenbewertung in jedem Stadium der Datenverarbeitung und für spätere Veröffentlichungszwecke.



---

# Contents

<b>1</b>	<b>Motivation</b>	<b>13</b>
<b>2</b>	<b>Introduction to X-rays</b>	<b>17</b>
2.1	Why do we need X-rays? . . . . .	17
2.2	Brief history of X-rays and their general properties . . . . .	18
2.3	Scattering of X-rays . . . . .	21
2.4	Scattering by an assembly of atoms . . . . .	23
2.5	Description of Crystals . . . . .	24
2.6	Temperature factor . . . . .	25
2.7	Diffraction by a crystal . . . . .	25
2.8	Laue condition and Bragg's Law . . . . .	26
2.9	The Ewald sphere construction . . . . .	28
2.10	Anomalous signal and Friedel's law . . . . .	28
<b>3</b>	<b>X-ray sources</b>	<b>33</b>
3.1	Radiation characteristics . . . . .	33
3.1.1	Flux, Brightness, Brilliance . . . . .	33
3.1.2	Emittance . . . . .	34
3.1.3	Coherence . . . . .	35
3.2	The standard X-ray tube . . . . .	36
3.3	Synchrotrons . . . . .	37
3.3.1	An advanced High-Throughput Pharmaceutical X-ray screening instrument (HiPhaX) . . . . .	38
3.4	Free-electron lasers . . . . .	40
3.5	Detectors . . . . .	42
<b>4</b>	<b>Data collection and analysis in protein crystallography</b>	<b>45</b>
4.1	Data collection techniques in X-ray crystallography . . . . .	45
4.1.1	Laue crystallography . . . . .	45
4.1.2	Single crystal rotation . . . . .	46
4.1.3	Powder diffraction . . . . .	46
4.1.4	Serial crystallography . . . . .	47

4.1.5	Time-resolved crystallography . . . . .	48
4.1.6	Reflection partiality problem . . . . .	51
4.1.7	Crystals . . . . .	52
4.1.8	Sample delivery systems . . . . .	53
4.2	Data analysis of conventional crystallography . . . . .	56
4.3	Data analysis in serial crystallography . . . . .	57
4.3.1	Pre-processing and hit-finding . . . . .	57
4.3.2	Indexing . . . . .	58
4.3.3	Integration and merging of intensities . . . . .	59
4.4	Initial phase estimate . . . . .	60
4.4.1	Molecular replacement . . . . .	60
4.4.2	Direct methods . . . . .	60
4.4.3	Single and multiple isomorphous replacements (SIR, MIR) . . . . .	61
4.4.4	Single and multiple anomalous dispersion (SAD, MAD) . . . . .	61
4.5	Radiation damage . . . . .	62
4.6	Existing auto-processing data pipelines at different X-ray facilities . . . . .	64
4.7	Modern problems in developing data processing pipelines for serial crystallography . . . . .	67
<b>5</b>	<b>Improving data processing in protein crystallography</b>	<b>69</b>
5.1	Data quality metrics and some hints for data processing . . . . .	70
5.2	The calibration of detector distance and origin using conventional crystallography . . . . .	76
5.3	Efficient data collection using chips . . . . .	78
5.3.1	Experimental setup and data collection . . . . .	78
5.3.2	Data analysis . . . . .	79
5.3.3	Results and discussion . . . . .	82
5.4	Offline data processing pipeline for serial crystallography . . . . .	85
5.5	Tool for generating per pattern mask for salt or ice reflections . . . . .	89
5.6	Auto-processing pipeline for HiPhaX - a drug screening beamline P09, Petra III . . . . .	92
5.6.1	Specification of configuration file for current experimental setup . . . . .	92
5.6.2	Data collection and auto-processing strategy . . . . .	94
5.6.3	Google Sheets as an optimal database for monitoring results and saving metadata . . . . .	94
5.6.4	Outlook . . . . .	95
<b>6</b>	<b>Enhancing data quality through modern data processing pipelines</b>	<b>97</b>
6.1	Re-processing previously collected data . . . . .	97
6.1.1	Anomalous dataset . . . . .	98
6.1.1.1	Reprocessing previously collected SAD data at P11 beamline, PETRA III . . . . .	98
6.1.2	Reprocessing previously collected data at LCLS in 2011 . . . . .	100
6.1.3	X-ray Diffraction Analysis of Hemoglobin Samples at LCLS MFX . . . . .	105
6.2	Conclusion . . . . .	110
<b>7</b>	<b>Compression and data reduction in serial crystallography</b>	<b>113</b>
7.1	Introduction . . . . .	113
7.2	Review of existing data reduction methods in science . . . . .	115

7.3	Selection of test datasets . . . . .	117
7.4	Applying different lossless and lossy compressions . . . . .	118
7.4.1	Existing lossless compressions and its evaluation . . . . .	118
7.4.2	Lossy compression . . . . .	119
7.4.2.1	Non-hits rejection . . . . .	119
7.4.2.2	Measuring less data . . . . .	119
7.4.2.3	Storing only detectable Bragg peaks . . . . .	123
7.4.2.4	Binning to lower the number of detector pixels . . . . .	123
7.4.2.5	Quantization of detector output . . . . .	124
7.4.2.6	Non-uniform quantisation . . . . .	126
7.5	Discussion . . . . .	129
7.6	Conclusion . . . . .	130
<b>8</b>	<b>Summary and outlook</b>	<b>133</b>
<b>A</b>	<b>Introduction to X-rays</b>	<b>139</b>
A.1	Correction Terms for the Atomic Scattering Factor . . . . .	139
A.2	Refraction, reflection and absorption . . . . .	140
<b>B</b>	<b>X-ray sources</b>	<b>147</b>
B.1	Synchrotron radiation . . . . .	147
B.1.1	Synchrotron radiation from a circular arc . . . . .	147
B.1.2	The natural opening angle of synchrotron radiation . . . . .	148
B.1.3	Characteristic frequency of synchrotron radiation . . . . .	149
B.1.4	Emitted power . . . . .	149
B.2	Equipment for modern X-ray sources . . . . .	151
B.2.1	The magnet lattice . . . . .	151
B.2.1.1	Bending Magnets and Superbends . . . . .	151
B.2.1.2	Quadrupole and Sextupole Magnets . . . . .	151
B.2.2	Insertion devices . . . . .	152
B.2.2.1	Wiggler . . . . .	154
B.2.2.2	Undulator . . . . .	154
B.3	X-ray monochromators . . . . .	157
<b>C</b>	<b>Improving data processing in protein crystallography</b>	<b>159</b>
C.1	Offline data processing pipeline for serial crystallography . . . . .	159
C.2	Tool for generating per pattern mask for salt or ice reflections . . . . .	161
C.3	Auto-processing pipeline for HiPhaX - a drug screening beamline P09, Petra III . . . . .	163
C.3.1	Google Sheets as an optimal database for monitoring results and saving metadata . . . . .	163
<b>D</b>	<b>Enhancing data quality through modern data processing pipelines</b>	<b>167</b>
D.1	Re-processing old data, LCLS datasets . . . . .	167
D.1.1	Reprocessing previously collected data at LCLS in 2011 . . . . .	167
D.1.2	X-ray Diffraction Analysis of Hemoglobin Samples at LCLS MFX . . . . .	194

D.1.2.1	Crystal preparation . . . . .	195
<b>E</b>	<b>Compression and data reduction in serial crystallography</b>	<b>197</b>
E.1	Existing lossless compressions and its evaluation . . . . .	197
E.2	Lossy compression . . . . .	200
E.2.1	Binning to lower the number of detector pixels . . . . .	200
E.2.2	Quantization of detector output . . . . .	200
E.2.3	Non-uniform quantisation . . . . .	201
E.2.3.1	Chunk summation of diffraction patterns . . . . .	204
E.3	Different samples used for the tests . . . . .	211
	<b>List of Publications</b>	<b>217</b>
	<b>Bibliography</b>	<b>219</b>



---

# Motivation

Proteins play a pivotal role within living cells, and their versatile applications have been demonstrated across an extensive array of different fields [1–5]. The 3D structures of proteins can be observed at the atomic scale by the method of X-ray crystallography. In macromolecular crystallography, the conventional approach involves acquiring diffraction patterns from a crystal while it undergoes rotation along one or more axes. The total tolerable X-ray exposure of the crystal is limited by the accumulation of damage to the protein structure by ionising radiation [6, 7]. Cryogenic cooling reduces the processes of radiolysis and extends the exposure that can be tolerated. However, such cooling may alter the macromolecular structure and prevent the ability to measure dynamic processes by time-resolved methods [8].

Recently, the method of serial crystallography (serial crystallography (SX)) has been developed at synchrotron radiation facilities and X-ray free electron lasers (Free electron laser (FEL)s) to overcome this limitation [8–12]. In contrast to the conventional rotation method, in SX, many randomly oriented crystals are sequentially exposed to an X-ray beam, one at a time. This random orientation of crystals leads to collecting numerous snapshot diffraction patterns to obtain a complete set of 3D structure factors in a stochastic manner.

Serial crystallography avoids the need for cryogenic cooling. It enables the collection of data close to physiological conditions, such as room temperature and measuring tiny crystals, which gives access to faster dynamics. It is possible due to the ability to apply the total tolerable X-ray exposure to each crystal instead of distributing it across a rotation series of a single crystal. Thus, we associate the SX technique with the principle of "diffraction before destruction" [13–15].

Serial crystallography conducted at synchrotron or free-electron laser (FEL) sources offers a powerful approach to unravelling the dynamics of structural fluctuations and investigating the mechanisms of macromolecules. By obtaining protein structures at multiple time points, this technique provides valuable insights into the dynamic behaviour of biomolecules [16, 17]. First, SX experiments were demonstrated with short-pulses X-ray of FELs [9] and are usually called Serial Femtosecond Crystallography (serial femtosecond crystallography (SFX)). A later similar approach was implemented at synchrotrons [10] and is often called Serial Synchrotron Crystallography (synchrotron serial crystallography (SSX)).

In the last decades, the technology of X-ray crystallography has developed rapidly. New generations of X-ray sources generate highly intense and coherent X-ray beams, coupled with improved focusing optics that enhance the flux density at the sample. As a result, the exposure time needed to capture measurable signals has significantly decreased. Modern detectors can capture thousands of images per second. The development of these detectors, coupled with the aforementioned high-intensity photon sources, enables the collection of valuable images at a kilohertz frame rate [18]. Combined with the tiling of detector modules to increase the

number of pixels (up to 16 million pixels for Eiger or JUNGFRÄU), this leads to very high overall data rates. This equates to a potential accumulation of close to 4 PB/day by each detector, which alleviates the burden on data storage systems and imposes a substantial burden on operational budgets.

Various sample delivery systems have been employed or investigated for delivering micro- to nano-scale crystalline samples into the X-ray beam [19, 20]. The choice of the most suitable sample delivery method depends on the experimental goals, the required environment, and the characteristics of the crystals (such as size or quantity). The idea behind crystal injection methods is to obtain a fine stream of crystals by ejecting a suspension of crystals through a small nozzle. The crystal stream is usually orthogonal to the X-ray beam, and a diffraction pattern is acquired at each X-ray pulse or exposure, whether the beam intersects a crystal or not. The central concept of fixed-target sample delivery systems (for example, chips or tape delivery systems) is to fix the crystal on a solid support, which is then raster scanned. The fixed-target approach allows efficient protein usage since each crystal can be measured, dramatically reducing sample consumption. Thus, it is an appropriate method for delivering protein samples that are not expensive to produce. This sample delivery technique also allows on-the-chip crystallisation [21]. Thus, it is more suitable for brittle crystals to prevent damage. This is a massive advantage over the injection methods, which can create risks in the filtering, transfer, and loading stages and pressures and forces associated with the injection process. Recent advances have paved the way for studying a broader range of possible protein samples and various dynamic states. Having in hand electron microscopy and new achievements in the application of artificial intelligence, such as AlphaFold [22–24], significantly enhance our understanding of the determination of protein structures with X-rays.

Even though serial crystallography offers a potential solution to overcome the limitations mentioned above for macromolecular crystallography, SX introduces its own set of demands and challenges regarding data processing. The main problems in data processing for serial crystallography are the lack of a convenient pipeline for the user, a high entry threshold (knowledge is required that can only be obtained as a result of participation in real experiments at different facilities), data volume and non-standard settings, which usually change from experiment to experiment. In terms of online data processing, long experiments spanning days heighten pressure and error risks during data processing. Manual script execution and result inspection become unfeasible. Challenges with beam drift, detector repositioning for large unit cell parameters, and limited computational resources impede prompt data processing team feedback. Furthermore, achieving full automation for both conventional and serial crystallographic experiments can be problematic due to the emergence of new beamlines, detectors and X-ray sources. Existing solutions for automating data processing pipelines at beamlines may not meet the requirements of these new experimental setups and collection strategies. Thus, the development of data processing pipelines should consider the possibility of adaptability without rebuilding the entire concept and the ability to adjust to different installed control systems at various beamlines.

Data processing software has significantly improved in the following years. There are several reasons for the improvement in result quality: new algorithms for indexing (including indexing multiple crystals per pattern) and integration implemented in CrystFEL [25], better knowledge of the detector geometry [26] and a different strategy for the background subtraction. In this study, we evaluate how these more advanced data processing algorithms can help to get better results from previously collected data from different experiments and X-ray sources. Despite the evident progress in data processing, there remains a need for further enhancement in user-friendliness to establish the widespread utility of serial crystallography.

This dissertation is dedicated to solving the existing problems of serial crystallography to make it generally applicable: ease of use and data handling. The developed data processing pipeline can be easily integrated into the existing control system of the macromolecular beamlines at different facilities and triggered from the

control system per each collected run or executed as a demon in parallel to the control system and even used later for offline re-processing. This pipeline provides transparency in each step of data processing and, in the end, generates the necessary figure of merits and overall statistics for sequential structure refinement processes. A notable advantage of this data processing strategy is that it is a universal, reliable, well-established pipeline with the generation of intermediate files needed for proper data-quality checks. We also comprehensively discuss applying data quality metrics while evaluating processing results.

As mentioned, up to 5 PB of data per experiment can be easily obtained under efficient operating conditions. The combined costs associated with storing data from multiple experiments provide a compelling incentive to develop strategies that effectively reduce the amount of data stored on disk while maintaining the quality of scientific outcomes. This thesis assesses various approaches to applying lossless and lossy compression techniques to SX data. Lossless data compression methods are designed to preserve the information content of the data but often struggle to achieve a high compression ratio when applied to experimental data that contains noise. Conversely, lossy compression methods offer the potential to reduce the data volume significantly. Accurately assessing the potential loss of scientific content resulting from the application of lossy data compression necessitates a comprehensive set of metrics specifically designed to measure the extent of information loss. It is necessary to quantify not only whether the quality of the final molecular structure is affected but also whether the ability to perform any of the many intermediate analysis steps, for example, peak finding or estimating background signal, is compromised due to loss of data quality. Such metrics as data quality and reproducibility ( $SNR$ ,  $R_{split}$ ,  $CC^*$ ), the quality of the reconstructed structure factors ( $R_{free}/R_{work}$ ) and the possibility of using the anomalous signal for ab-initio structure reconstruction (single anomalous dispersion (SAD) phasing) must be properly employed for data quality evaluation. It was shown that saving the raw detector frames containing strong crystal diffraction is highly effective for reproducing results at a later stage. Discarding blank frames has little effect on data quality, even if some of those 'blank' frames may be found to contain weak diffraction using more advanced algorithms developed at a later time. Conversely, retaining information from only locations of found Bragg peaks in each pattern has a significantly detrimental effect on data quality. Lowering the number of pixels in the detector obviously saves data space, provided it is compatible with the crystal being studied. Where data is saved in floating point ADU units, quantization to integer numbers of photons is highly effective in reducing data volumes as the additional precision of sub-single-photon counting accuracy is not required for SX measurements. Also, compression of the dynamic range of measurements in a non-linear manner following statistical noise is highly effective so that weak reflections are still accurately measured, but there is less precision in the measurement of strong intensities. This is achieved in practice by saving only several of the most significant bits of the values measured by each pixel. In this way, the low signal data is saved almost without losing the precision, which is very important for data measured at high resolution close to the detector edges. In principle, such a scheme is similar to the multiple-gain mode used in modern detectors for capturing high dynamic range signals while keeping high sensitivity for low signals but applied with many more levels in software after the measurement is made. This data reduction approach is very computationally efficient. Therefore, it might be implemented inside new detectors.

By combining the above data reduction methods, including real-time hit finding, binning, quantization to photons, and non-linear reducing the dynamical range, it should be possible to continue retaining individual detector frames for later study while also reducing the volume of data which must be permanently retained at facilities or user groups worldwide. It is known that software improves over time, and careful reprocessing of previously collected data might deliver much better results at a later point in time. Thus, this dissertation showed that reprocessing the previously collected dataset measured at different facilities leads to a much higher resolution

than originally obtained, provided the raw frames containing crystal diffraction were available. Even some structural features not observed during the initial analysis were resolved after the reprocessing. It showcases the necessity to preserve raw data to evaluate newly appeared data processing pipelines or data reduction schemes. However, the question of how much data should be stored and for how long is undoubtedly a matter of debate that will continue for quite some time.

Chapters 3 and 2 of this thesis give a general introduction to crystallography, covering topics related to the X-ray properties and sources and the theoretical background, respectively. Chapter 4 gives an overview of experimental aspects of X-ray crystallography and delves into data analysis methods for conventional and serial crystallography. Chapter 5 first introduces the established quality metrics and other figures of merits for proper data evaluation and then focuses on improving data processing for conventional and serial crystallography. Also, Chapter 5 discusses the hit optimisation approach for chip measurement and describes a data processing pipeline introduced at a drug-screening beamline P09, Petra III. Chapter 6 demonstrates how reprocessing previously collected data can lead to much better results. Chapter 7 is dedicated to the topic related to data compression and reduction for serial crystallography. In Chapter 7, various lossy compression strategies were tested on different datasets and carefully evaluated. Chapter 8 summarises the thesis results and gives an outlook on future research.

# Introduction to X-rays

## 2.1 Why do we need X-rays?

X-rays - like visible light, radio waves, microwaves, ultraviolet radiation, etc. -are a form of electromagnetic radiation distinguished only by the magnitude of their wavelengths. The first reason to use X-rays is their wavelength and the ultimate resolution of an image one can obtain when using a given wavelength. X-rays offer superior resolution due to their shorter wavelength. In the visible light range, the ultimate resolution is limited by the Abbe limit ( $d = \frac{\lambda}{2n\sin\theta}$ ,  $n$ —the refractive index of the lens,  $\theta$ —is the half angle of the radiation collected by that lens), which is around  $\lambda/2$  or approximately 250 nm. In contrast, with their weak refraction, X-rays provide ultimate resolution depending on wavelength and lens characteristics, typically requiring objects to be separated by a few times  $\lambda$  for proper resolution.

The second argument favouring X-rays is that we can derive the real space information from encoded diffraction patterns. This is based on the Bragg law:

$$m\lambda = 2d\sin\theta \quad (2.1)$$

Where  $\theta$  describes the scattering angle of X-radiation of an object, let us contemplate a planar incident radiation wave upon a diminutive object. This object scatters the radiation in various directions. Now, envision two objects and gradually increase their spatial separation. As this separation widens, the scattered waves start to superpose and interfere with each other, giving rise to distinct diffraction maxima designated by the integer  $m$ . Higher values of  $m$  correspond to enhanced spatial resolution, enabling us to extract more detailed real-space information from the encoded diffraction pattern.

Thirdly, the condensed matter is partially transparent to X-rays. How transparent an object is depends on the X-ray photon's wavelength or energy, the object's size, and the density and types of atoms that make up the object being irradiated. Once the correct X-ray energy has been chosen, a different degree of transparency of the various components of an object can be used to examine the internal architecture of that object. As a rule of thumb, the denser the material, the less transparent it will be to x-rays of a given energy. The transparency of objects to X-rays implies that X-rays do not interact very strongly with matter. But if we want to use them to probe matter's structure and physical properties, we want them to interact to a certain degree. The interaction of, for example, protons,  $\alpha$  particles, or optical light with the material is so strong that analysis of this interaction can become problematic. On the contrary, since neutrons have no charge, they pass through matter with annoying ease, limiting the ability to study the smallest objects. On the other hand, X-rays have cross sections that are

large enough to provide easily measurable effects while being sufficiently small that they pass through significant volumes of material, which is only mildly impacted by the weak interactions and thus not rapidly destroyed.

Lastly, most naturally occurring elements have electrons with binding energies within the X-ray range. When X-ray energy is adjusted to be close to a specific electron's binding energy, it's called resonance, significantly boosting their interaction. This resonance allows for the probing of specific electron types, and depending on the interacting electron, various physical and chemical properties, including chemical binding, elemental composition, magnetic characteristics, and electronic structure, can be studied.

## 2.2 Brief history of X-rays and their general properties

X-rays were discovered by Wilhelm Conrad Röntgen in 1895. Since then, it has become an invaluable probe of the structure of matter. In 1912, Max von Laue proposed using crystals as natural lattices for diffraction. He was also the first one who observed the diffraction of X-rays and revealed the wave nature of X-rays. In 1913, Braggs (father and son) conducted experiments and showed that X-ray diffraction could be used to determine the atomic structure of matter. They interpreted diffraction as a reflection from the planes of a crystal and solved the crystal structures of *NaCl* and *KCl*, and introduced Fourier analysis of the X-ray measurements. The relationship between a Fourier synthesis and a Fourier analysis demonstrated that phase is the central problem in structural crystallography. Later, in 1917 P. P. Ewald introduced the 'reciprocal lattice' construction, a graphical method to express the geometrical conditions for crystal diffraction.

The gradual development of X-ray crystallography demanded a systematic understanding and tabulation of space groups. In 1935, the crystallographic community assembled the first set of Internationale Tabellen [27], containing diagrams and information on about 230 space groups. After World War II, the International Tables Volume I form [28] appeared as an extension of the previous tables, combined with Kathleen Lonsdale's structure factor formulas [29]. Later it was revised and extended again in Volume A [30].

At the beginning of the development of crystallography, simple organic compounds have been studied since the 1920s. One important example is the structure of hexamethylbenzene by Kathleen Lonsdale [31], who showed that benzene had a planar hexagonal structure. In the early 1930s J. D. Bernal could distinguish several possible structures of steroids by studying their arrangement in different unit cells [32]. In the mid-1930s, J. Monteath Robertson and I. Woodward determined the structure of nickel phthalocyanine using the heavy-atom method [33].

Bijvoet's groundbreaking work on sodium tartrate [34, 35] played a pivotal role in determining the absolute hand of the asymmetric carbon atom. By indexing the X-ray reflections using a right-handed system, Bijvoet demonstrated the breakdown of Friedel's law in the presence of an anomalous scatterer. This observation proved the asymmetric carbon atom conformed to Fischer's convention, establishing its absolute hand. As a result, the absolute structure of other molecules containing asymmetric carbon atoms, such as naturally occurring amino acids and riboses, could be confidently determined.

In the mid-1950s, structure determinations were primarily based on projection data. Using such data reduced the effort required for manual indexing and making visual estimates of intensity measurements. It made the calculation process more manageable, especially in the absence of computing machines. For example, the structure determination of penicillin during World War II by Dorothy Hodgkin and Charles Bunn utilized three-dimensional data [36, 37]. Another significant breakthrough was determining the three-dimensional structure of vitamin B12 by Hodgkin and her colleagues in the 1950s [38]. Due to the presence of the heavy metal cobalt in the vitamin B12 fragment, the team was able to identify the "corrin" ring structure. The remaining B12 structure

was determined through a collaboration between Hodgkin's group in Oxford and Kenneth Trueblood's group at UCLA, with the latter performing the computing on the early electronic Standard Western Automatic Computer and additional help from J. G. White at Princeton University in New Jersey.

The impact of advancements in data-collection devices has been significant. Before the mid-1950s, intensity measurements relied heavily on visual comparisons of reflection "spots" on films with a standard scale. However, the use of counters, which were used by Bragg in 1912, gradually became automated and became the preferred technique in the 1960s. Additionally, semi-automatic methods, which involved measuring the optical densities along reciprocal lines on precession photographs, were extensively utilised for early protein-structure determinations in the 1950s and 1960s.

Macromolecular crystallography (MX) is a powerful technique to determine biological macromolecules' three-dimensional (3D) structures, such as proteins, nucleic acids, and large complexes. It is a fundamental tool for structural biology, allowing researchers to investigate the relationships between molecular structure, function, and dynamics.

MX has undergone significant advancements since its inception in the early 1900s. The first macromolecular structure to be determined was that of hemoglobin in 1959, using X-ray crystallography. This breakthrough was followed by the developing of new methods and technologies that have revolutionised the field. One of the most significant advancements in MX was the introduction of computers in the 1960s. Using computers for data processing and analysis significantly improved the speed and accuracy of structure determination. In the 1970s, synchrotron radiation became prevalent, allowing for the collection of high-quality diffraction data and increasing the resolution of the determined structures.

In the 1980s and 1990s, automated crystallography systems were developed, allowing for the rapid screening and optimisation of crystallization conditions. This, along with developing new phasing methods such as molecular replacement and multi-wavelength anomalous dispersion, enabled the determination of structures of ever-increasing complexity and size. More recent developments in MX include serial femtosecond crystallography, which allows for the determination of structures of dynamic or radiation-sensitive samples, and cryo-electron microscopy, which enables the visualisation of structures of large complexes at near-atomic resolution. Overall, the development of MX has been a key factor in advancing our understanding of the structure and function of biological macromolecules, providing insights into biological processes and aiding in the design of new drugs and therapies. Here we can see the list of some Nobel Prizes awarded for works related to X-ray crystallography:

1901	<b>W.C. Röntgen</b>	the discovery of remarkable rays, later named after him
1914	<b>Max von Laue</b>	the discovery of the diffraction of X-rays by crystals
1915	<b>W.H. Bragg and W.L. Bragg</b>	the analysis of crystal structure by means of X-rays
1946	<b>J.B. Sumner</b>	the discovery that enzymes can be crystallised
1958	<b>F. Sanger</b>	the structure of proteins, especially of insulin
1962	<b>M. Perutz and J. Kendrew</b>	the studies of the structures of globular proteins
1962	<b>F. Crick, J. Watson, M. Wilkins</b>	the studies of molecular structure of nucleic acids
1964	<b>D.C. Hodgkin</b>	the structures determinations of vitamin B12 and insulin
1988	<b>J. Deisenhofer, R. Huber, H. Michel</b>	the structure determination of a photo-reaction centre
1997	<b>P.D. Boyer, J.E. Walker</b>	the structure determination of ATP synthase
2003	<b>R. MacKinnon</b>	structural and mechanistic studies of ion channels

2009	<b>V. Ramakrishnan, T.A. Steitz, A.E. Yonath</b>	the structure and function of the ribosome
2012	<b>R.J. Lefkowitz, B.K. Kobilka</b>	functional and structural studies of GPCR proteins

Now, we are going to talk about the general properties of X-rays. X-rays are electromagnetic waves with wavelengths in the region of an Ångström ( $10^{-10}$  m). Thus, it has all the properties of electromagnetic radiation (EM), such as:

- EM propagates in a vacuum with the velocity equals the speed of light  $c = 3 \cdot 10^8 \text{ m s}^{-1}$ .
- The vectors of the electric field  $E$  and magnetic field  $B$  are mutually orthogonal and both perpendicular to the direction of the wave propagation.
- The X-ray wavelength range is between the ultraviolet region and the region of  $\gamma$ -rays emitted by radioactive substances.
- Optical components for X-rays are quite different than those used for visible light because the refractive index of X-rays is very close to unity. Moreover, soft X-rays are strongly absorbed by condensed matter and can even be strongly attenuated by gas. Thus, this makes it difficult to bend or redirect X-rays with conventional lenses using visible light or electrons. The phenomena of total external reflection, refraction, and diffraction are all used in X-ray optics to achieve these ends.

There are several ways in which X-ray interacts with matter:

- They could be scattered elastically (without loss of energy, Thomson scattering) or inelastically (with loss of energy, Compton scattering)
- The atoms could absorb the photons to eject electrons (photoelectric effect). The ionised atoms then emit photons (fluorescence) or Auger electrons.

Electrons scatter X-rays. Atomic electron clouds have the same length scale as the wavelength of radiation, and atoms are represented as extended objects, so we cannot consider atoms as points. The X-ray scattering does not distinguish isotopes due to the same number of electrons, but its power increases simply with the atomic number. Thus, heavy atoms dominate in the diffraction pattern, which could be used for labelling in crystallography.

We could think of X-ray scattering as oscillating fields of electromagnetic radiation that create oscillating dipoles in the electron cloud, emitting radiation. As mentioned earlier, an atomic electron cloud behaves like an extended object in space, and spherical waves scattered from each point of this object interfere, which leads to a scattering intensity distribution that depends on the scattering angle (the form factor). The diffraction from a single molecule is too weak to observe. An ordered three-dimensional array of molecules (crystal) is used to magnify the signal. If the crystal is imperfect, the X-rays will not be directed to high angles, and the data will not have a detailed structure. On the other hand, in a situation with a well-ordered crystal, diffraction will be measured at large angles or with high resolution, and the result should be a detailed structure. In addition, the resulting resolution of a diffraction experiment depends on how well the particles are ordered with respect to each other in the sample.



## 2.3 Scattering of X-rays

In this section, we will consider the coherent scattering of an X-ray in detail. When X-rays interact with an object, the scatterers are electrons. J.J. Thomson derived the expression for X-ray scattering by a single electron, and the calculated scattering factor, known as the Thomson factor, is  $f_{el} = e^2/mc^2 = 2.8 \times 10^{-15}$  m, where  $e$  and  $m$  are respectively the charge and mass of the electron and  $c$  is the speed of light. Based on classical electrodynamics, the periodic electromagnetic field of the incident wave exerts on the charged particle a periodic force, and thus, the particle will oscillate with the same frequency as the electric field. Therefore, the charged particle becomes the source of electromagnetic radiation of the same frequency. The total intensity of the radiation is

$$I_{Th} = I_0 \left( \frac{f_{el}}{r} \right)^2 (\sin^2 \mu + \cos^2 \mu \cos^2 2\Theta) \quad (2.2)$$

Where  $I_0$  is the intensity of incident wave,  $r$  is the distance from the particle (or in other words,  $1/r^2$  is a measure of solid angle),  $\mu$  is polarisation angle and  $2\Theta$  is the angle between the scattered and incident beam (scattering angle).

If the incident beam is non-polarised Eqn. 2.2 becomes

$$I_{Th} = I_0 \left( \frac{f_{el}}{r} \right)^2 \left( \frac{1 + \cos^2 2\Theta}{2} \right) \quad (2.3)$$

where  $P = \frac{1 + \cos^2 2\Theta}{2}$  is called polarisation factor. The scattered radiation will be partially polarised even if the incident radiation is not.

Coherent X-ray scattering is a result of contribution only from electrons because neutrons do not have electric charge, while protons are much heavier than electrons, which makes their contribution negligible. The scattering is coherent because there is a defined phase relation between the incident and scattered beam:  $\Delta\phi = \pi$  for electrons, therefore, the scattered waves will interfere.

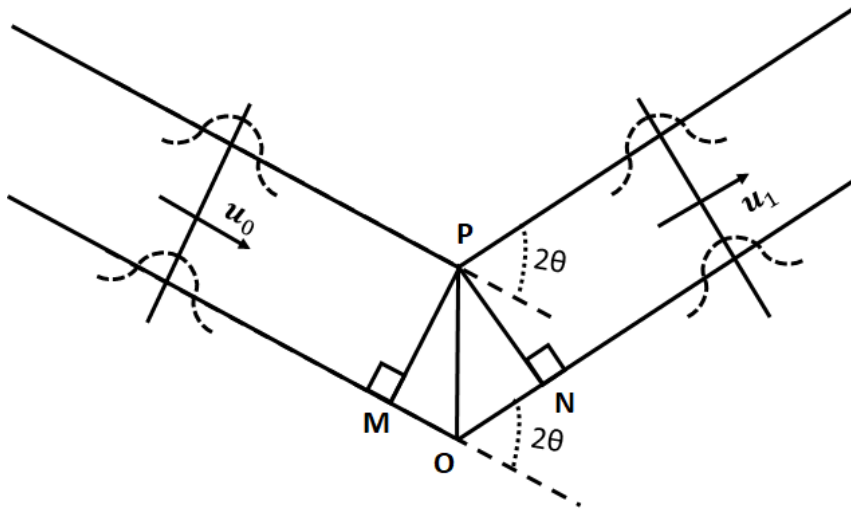


Figure 2.1: The path difference between the waves scattered from P and O is  $\Delta = OM + ON = r(\mathbf{u}_1 - \mathbf{u}_0)$ , where  $r$  is the vector  $OP$  and  $\mathbf{u}_1$  and  $\mathbf{u}_0$  are unit vectors parallel to the incident and scattered waves respectively. The incident and scattered waves have the same energy:  $\|\mathbf{k}_1\| = \|\mathbf{k}_0\| \Rightarrow$  the phase difference is  $\delta\phi = \frac{2\pi}{\lambda} \Delta = (\mathbf{k}_1 - \mathbf{k}_0)r$

Now, we are going to consider two electrons illuminated with a monochromatic X-ray beam and that the elastically scattered radiation is observed along a direction  $\mathbf{k}_0$  (Fig. 2.1). We assume that the Fraunhofer

conditions are satisfied: the source and detector are sufficiently far from the origin that the incident and scattered X-rays may be represented as plane waves. The phase difference between the incident and scattered X-rays in the direction  $\mathbf{u}_1$  is equal to  $(\mathbf{k}_1 - \mathbf{k}_0)\mathbf{r} = \frac{2\pi}{\lambda}(\mathbf{u}_1 - \mathbf{u}_0)\mathbf{r}$ . Therefore, we could describe the scattered wave from P by

$$f \exp \frac{2\pi}{\lambda}(\mathbf{u}_1 - \mathbf{u}_0)\mathbf{r} = f \exp i(\mathbf{k}_1 - \mathbf{k}_0)\mathbf{r} = f \exp i\mathbf{r}^*\mathbf{r} \quad (2.4)$$

Where we introduce wave vectors  $\mathbf{k}_0$  and  $\mathbf{k}_1$  of the magnitude  $\frac{2\pi}{\lambda}$  in the directions of the incident and scattered wave respectively, and  $\mathbf{r}^* = \mathbf{k}_1 - \mathbf{k}_0$  with magnitude equals to  $\frac{4\pi}{\lambda} \sin \theta$  (Fig. 2.2).

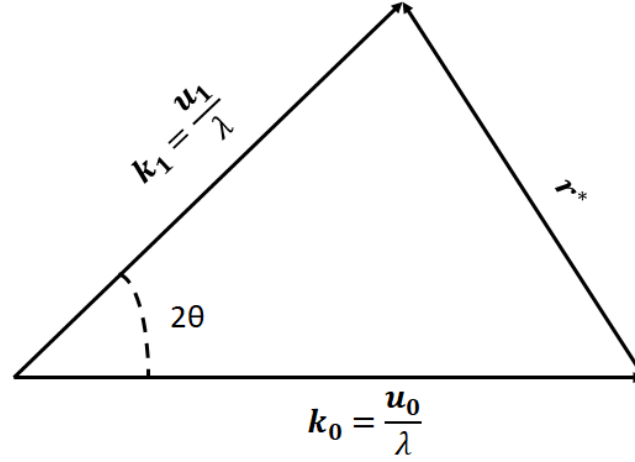


Figure 2.2: Definition of the scattering vector  $\mathbf{r}^* = \mathbf{k}_1 - \mathbf{k}_0$

If there are  $n$  scatterers with a scattering amplitude  $f_j$ , we get

$$F(\mathbf{r}^*) = \sum_{j=1}^n f_j \exp i\mathbf{r}^*\mathbf{r} \quad (2.5)$$

In the case that the electrons are continuously distributed with a continuous electron density  $\rho(\mathbf{r})$  the sum is from Eqn. 2.5 replaced by an integral, and the structure factor becomes

$$f(\mathbf{r}^*) = \int_V \rho(\mathbf{r}) \exp i\mathbf{r}^*\mathbf{r} d\mathbf{r} = \mathcal{F}(\rho(\mathbf{r})) \quad (2.6)$$

where  $\mathcal{F}(\mathbf{r})$  represents Fourier transform operator. The space of  $\mathbf{r}^*$  vectors is called reciprocal space: for smaller values  $\mathbf{r}$ , larger values of  $\mathbf{r}^*$  are needed to obtain the same path difference.

Generally, it is assumed that the electron cloud around the atom is spherically symmetric. In other words, it ignores the spatial heterogeneity of the electron orbitals, which is often a good simplification, especially for atoms with large atomic numbers, which means having many electrons. But for some atoms, this is not a very good approximation. For example diamond, each carbon atom in a diamond is covalently bound to four neighbouring carbon atoms in a tetrahedral configuration. To attain the diamond structure, the four outermost electrons within the  $N = 2$  shells coalesce to generate tetrahedral  $sp^3$  hybrid orbitals possessing requisite symmetry. But carbon only contains 6 electrons, so 4 out of the 6, two-thirds form an electron cloud that is most certainly not spherically symmetric. Certain diffraction peaks in diamond-like structures, including diamond itself, silicon, and germanium, are curiously missing in their diffraction patterns. One of these is the so-called 222 diffraction peak. Such missing peaks or so-called systematic absences are only predicted to have a mathematically 0 intensity by diffraction theory when one assumes that the electron cloud is spherically symmetric.

## 2.4 Scattering by an assembly of atoms

Based on the two-atom case, we can build a picture of scattering by assembling atoms. Thus, it is necessary to consider the resulting wave in the direction determined by the given scattering vector as the result of the interference of waves from all possible pairs of atoms in the ensemble. If the scattered waves interfere with each other and give a single resultant wave in a given direction, as in (Fig. 2.1), we are dealing with coherent scattering. For the correct summation of the wave amplitudes  $f_j$ , it is necessary to consider the phase relations between the waves from each atom  $j$ . The intensity when the waves are in phase is  $(\sum f_j)^2$ . If the scattering is incoherent, the resulting intensity equals  $\sum f_j^2$ , the sum of the intensities scattered by each of the atoms individually as if there were no other atoms. However, the intermediate case is often observed when scattered waves contain coherent and incoherent components.

As briefly mentioned in 2.2, we need very large numbers of atoms to amplify a weak signal from scattering by a single atom to observe the diffraction. Thus, we need to work with a coherent structure, which means that the sample must contain many pairs, all of which are made up of the same two atom types separated by the same vector. Otherwise, all deflections from a coherent structure (variations in either atom type or the vector separating the atoms) will contribute to incoherence.

To obtain coherent scattering from a group of atoms, both the incident beam and the properties of the sample must be coherent. The coherence length of the source should be longer than the sample size, and a constant amplitude and phase relationship with the incident wave must be maintained for each equivalent atom in the sample found in the same spatial environment. If not all equivalent positions in the crystal structure are occupied by atoms of the same type, the scattered radiation has an incoherent component. Incoherent scattering in a diffraction experiment does not contain structural information and contributes to the background noise.

The scattering process could be divided not only into coherent or incoherent but also into elastic and inelastic. If the scattered beam has the same energy as the incident beam, it is elastic scattering. In an inelastic case, the sample either loses energy to the radiation or gains energy from it. Standard diffraction experiments that provide information about the spatial arrangement of atoms are based on coherent elastic scattering (Thomson scattering). Meanwhile, Compton scattering of X-rays is incoherent, inelastic and related to absorption.

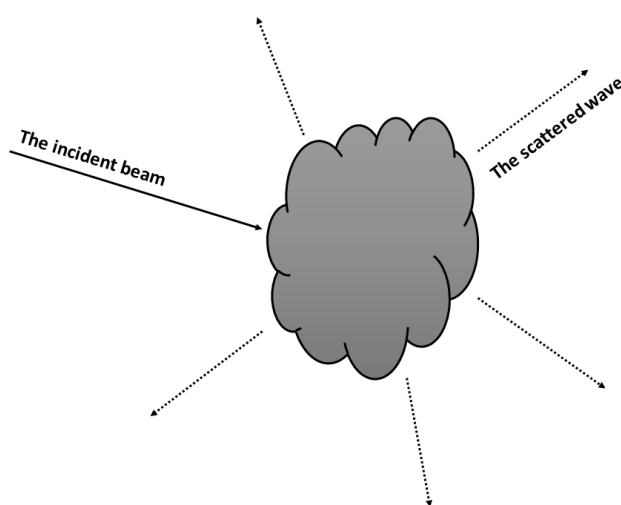


Figure 2.3: Scattering by an assembly of atoms in a particle.

A model of the diffraction pattern of a sample can be built up gradually, first considering scattering from all electrons in an atom, then from all atoms in a molecule, and so on, until we arrive at the description of the

scattering from the material of interest. The scattering from an atom at a point P with respect to an arbitrary origin O (Fig. 2.1) is described by Eqn. 2.4. In the case of a molecule, the scattered wave is determined simply by the sum of the waves scattered by each atom individually with respect to an arbitrary origin (Fig. 2.3):

$$F_M(\mathbf{r}^*) = \sum_{j=1}^n f_{aj} \exp i\mathbf{r}^* \cdot \mathbf{r} \quad (2.7)$$

where  $f_{aj}$  - the atomic scattering factor of  $j^{th}$  atom.

## 2.5 Description of Crystals

X-ray scattering from a single molecule would be incredibly weak and difficult to detect above the noise level because of the scattering from air and water. A crystal arranges many molecules in the same orientation so that the scattered waves can add up in phase and boost the signal to a measurable level. In a sense, a crystal acts as an amplifier.

A crystal is a periodic arrangement in space of a repeated motif, which is called a unit cell. A unit cell may contain one or several macromolecules organised symmetrically. An ideal crystal consists of an infinite array of identical units separated evenly from each other in three directions in space. We could describe an ideal crystal in terms of an infinite regular lattice of points in space (Bravais lattice). The Bravais lattice represents the periodicity of the crystal, and a group of atoms called the ‘basis’ that is anchored to each Bravais-lattice point. The lattice and basis form the crystal structure. Nevertheless, in reality, we are dealing with crystals of finite size containing some crystallographic defects.

Three-dimensional lattice can be specified by a set of vectors  $\mathbf{R}_{n,k,l}$ :

$$\mathbf{R}_{n,k,l} = n\mathbf{a} + k\mathbf{b} + l\mathbf{c} \quad (2.8)$$

where  $n, k, l$  are integers and  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  are noncoplanar vectors called lattice vectors. A given lattice has characteristic symmetries such as translations, rotations, reflections, and compound symmetries formed by combining a translation with rotations and/or reflections.

The diffraction of X-rays by a crystalline material is associated with scattering by atoms, which can be considered within families of planes. A plane can be defined by three non-collinear points in a volume. The Miller indices are the most convenient way to specify families of planes within a crystal. For a given family of planes, the Miller indices  $(h, k, l)$  are defined such that the plane closest to the origin (but not including the origin) has intercepts  $(a/h, b/k, c/l)$  on the axes  $(a, b, c)$ .

When the possible symmetries of the basis (known as the point group) are combined with those of the lattice, it turns out that all crystal structures can be classified into one of 230 possible space groups, as described in standard books on crystallography.

The lattice is a construct comprised of a series of infinitely sharp points that could be written as

$$L(\mathbf{R}) = \sum_{n,k,l=-\infty}^{+\infty} \delta(\mathbf{R} - \mathbf{R}_{n,k,l}) \quad (2.9)$$

where  $\delta(\mathbf{R})$  is the Dirac delta function. Knowing the electron density  $\rho_M(\mathbf{R})$  in the unit cell, to calculate the electron density of the infinite crystal, we need to perform a convolution of  $L(\mathbf{R})$  with  $\rho_M(\mathbf{R})$

$$\rho_\infty(\mathbf{R}) = \rho_M(\mathbf{R}) * L(\mathbf{R}) \quad (2.10)$$

## 2.6 Temperature factor

Since the period of thermal vibrations of atoms is much shorter than the time scale of the scattering experiment, the electron density of the atom, which determines the scattering, is the electron density averaged over time. In XFEL experiments, the period of thermal vibrations of atoms can be comparable to or even longer than the time scale of the scattering experiment.

The probability of finding an atom in the position  $\mathbf{r}$  for spherically symmetric vibrations is determined by the Gaussian with the average shift of the atom  $\sqrt{\overline{u^2}}$ :

$$p(\mathbf{r}) = \frac{1}{(2\pi\overline{u^2})^{3/2}} \exp(-\mathbf{r}^2/2\overline{u^2}) \quad (2.11)$$

The electron density corresponding to a thermally excited atom can be described by

$$\rho_{at} = \int \rho(\mathbf{r} - \mathbf{r}')p(\mathbf{r}')d\mathbf{r}' = \rho(\mathbf{r}) * p(\mathbf{r}) \quad (2.12)$$

The structure factor becomes:

$$f_{at}(\mathbf{r}^*) = \mathcal{F}(\rho(\mathbf{r}) * p(\mathbf{r})) = \mathcal{F}(\rho(\mathbf{r}))\mathcal{F}(p(\mathbf{r})) = f_a(\mathbf{r}^*) \exp(-B_{iso}\mathbf{r}^{*2}/4) \quad (2.13)$$

where  $B_{iso} = 8\pi^2\overline{u^2}$  is usually known as the atomic temperature factor. Based on Eqn. 2.13, it is necessary to notice the following: when  $\mathbf{r}^* = 0$ ,  $f_{at} = f_a = Z$  where  $Z$  is the atomic number; for any other value of  $\mathbf{r}^*$  we have  $f_{at} < f_a$ , thus, the thermal vibrations reduce the coherent scattering.

Generally, atoms in the crystal lattice will not be free to vibrate equally in all directions, and the thermal factor in this anisotropic case is represented by an ellipsoid centred on each atom.

## 2.7 Diffraction by a crystal

Having introduced the mathematical description of the crystal lattice and its electron density in 2.5, we can now calculate the scattering amplitude of the whole crystal from Eqn. 2.6 and get the following result

$$\begin{aligned} f_{\infty}(\mathbf{r}) &= \mathcal{F}(\rho_M(\mathbf{r}) * L(\mathbf{r})) = \mathcal{F}(\rho_M(\mathbf{r}))\mathcal{F}(L(\mathbf{r})) \\ &= f_M(\mathbf{r}^*) \frac{1}{V} \sum_{h,k,l=-\infty}^{+\infty} \delta(\mathbf{r}^* - \mathbf{r}_{h,k,l}^*) \\ &= \frac{1}{V} \sum_{h,k,l=-\infty}^{+\infty} F_{hkl} \delta(\mathbf{r}^* - \mathbf{H}_{hkl}) \end{aligned} \quad (2.14)$$

where  $V$  is the volume of the unit cell,  $\mathbf{H}_{hkl}$  is the generic lattice vector of the reciprocal lattice that is defined as

$$\mathbf{H}_{hkl} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*, \quad \mathbf{a}^* = 2\pi \frac{\mathbf{b} \times \mathbf{c}}{V}, \quad \mathbf{b}^* = 2\pi \frac{\mathbf{c} \times \mathbf{a}}{V}, \quad \mathbf{c}^* = 2\pi \frac{\mathbf{a} \times \mathbf{b}}{V} \quad (2.15)$$

$\mathbf{a}^*$ ,  $\mathbf{b}^*$  and  $\mathbf{c}^*$  are the reciprocal lattice vectors forming the reciprocal lattice and they fulfil

$$\begin{aligned} \mathbf{a} \cdot \mathbf{a}^* &= 2\pi; \mathbf{a} \cdot \mathbf{b}^* = 0; \mathbf{a} \cdot \mathbf{c}^* = 0; \\ \mathbf{b} \cdot \mathbf{a}^* &= 0; \mathbf{b} \cdot \mathbf{b}^* = 2\pi; \mathbf{b} \cdot \mathbf{c}^* = 0; \\ \mathbf{c} \cdot \mathbf{a}^* &= 0; \mathbf{c} \cdot \mathbf{b}^* = 0; \mathbf{c} \cdot \mathbf{c}^* = 2\pi; \end{aligned} \quad (2.16)$$

Due to the presence of a periodic scattering object, the amplitude of the scattered wave from Eqn. 2.14 is non-zero only when  $\mathbf{r}^*$  coincides with the reciprocal lattice point. The structure factor is a scattering factor of the unit cell in the reciprocal space at  $\mathbf{r}^* = \mathbf{H}_{hkl}$ , known as the Laue condition; thus, it is simply  $F_{hkl} = f_M(\mathbf{H}_{hkl})$ .

As we mentioned in 2.5, we are dealing with crystals of finite size, which must be taken into account when introducing the form function of the crystal:

$$\Phi(\mathbf{r}) = \begin{cases} 1 & \text{inside the crystal} \\ 0 & \text{outside the crystal} \end{cases} \quad (2.17)$$

Thus, we can rewrite the electron density of the finite crystal, based on Eqn. 2.10, as

$$\rho_{cr} = \rho_{\infty}(\mathbf{r})\Phi(\mathbf{r}) \quad (2.18)$$

and the amplitude of the diffracted wave from Eqn. 2.14 becomes

$$f_{cr}(\mathbf{r}^*) = \frac{1}{V} \sum_{h,k,l=-\infty}^{+\infty} F_{hkl} D(\mathbf{r}^* - \mathbf{H}_{hkl}) \quad (2.19)$$

Where

$$D(\mathbf{r}^*) = \mathcal{F}(\Phi(\mathbf{r})) = \int_V \exp i\mathbf{r}^* \cdot \mathbf{r} d\mathbf{r} \quad (2.20)$$

As a result, the delta function in Eqn. 2.14, corresponding to each point of the reciprocal lattice, in the case of a finite crystal, becomes the distribution function  $D(\mathbf{r}^*)$ , which is the same for all reciprocal lattice points.

## 2.8 Laue condition and Bragg's Law

The Laue condition  $\mathbf{r}^* = \mathbf{H}_{hkl}$  must be satisfied to observe X-ray diffraction. As we have already discussed in 2.7, only if  $\mathbf{r}^*$  coincides with a reciprocal lattice vector the scattered amplitude, described by Eqn. 2.14, from a crystallite is non-vanishing. Thus, coming back to 2.3, the definition of  $\mathbf{r}^*$  is  $\mathbf{r}^* = \mathbf{k}_1 - \mathbf{k}_0$ , therefore, the Laue condition would be represented in the following vector equation:

$$\mathbf{k}_1 - \mathbf{k}_0 = \mathbf{H}_{hkl} \quad (2.21)$$

where  $|k_1| = |k_0| = \frac{2\pi}{\lambda}$ ,  $\lambda$  is the wavelength of radiation.

In 1912 W. L. Bragg and his father, W. H. Bragg, explained the peak positions of an X-ray diffraction pattern by the famous equation named after them

$$m\lambda = 2d \sin \theta, \quad (2.22)$$

Where  $\lambda$  is the wavelength of the X-ray light,  $d$  is the interplanar spacing of the  $(hkl)$  planes and  $\theta$  is the angle of incidence above the plane surface, and  $m$  is an integer. Bragg's law describes the difference in the optical path length between reflections from adjacent crystal planes, which must be an integer multiple of wavelengths for constructive interference to occur. This also means that the phase difference between scattering from neighbouring planes is a multiple of  $2\pi$ .

It can be shown that the Laue condition is exactly equivalent to Bragg's Law. We are going to prove it by considering a two-dimensional square lattice. We have atomic planes with distance  $d$ , from which X-rays are elastically reflected. Constructive interference requires that the path length difference be an integer multiple

of the wavelength, in other words, Bragg's law must be satisfied  $m\lambda = 2d \sin \theta$ . For simplicity, let's consider  $m = 1$ . If we look at the same event in reciprocal space, we know that the Laue condition works  $\mathbf{r}^* = \mathbf{H}_{hkl}$ , discussed in 2.7. In the case of a two-dimensional square lattice, the reciprocal lattice is also a square with a lattice spacing equal to  $2\pi/d$ . We set the coordinates as is shown in (Fig. 2.4), thus,  $\mathbf{r}^* = \frac{2\pi}{d}(0, 1)$ . From the geometry we could see that  $r^* = 2k \sin \theta$  since  $|k_1| = |k_0| = k = \frac{2\pi}{\lambda}$ , therefore, we obtain  $\frac{2\pi}{d} = 2k \sin \theta$  which can be rearranged to Bragg's Law.

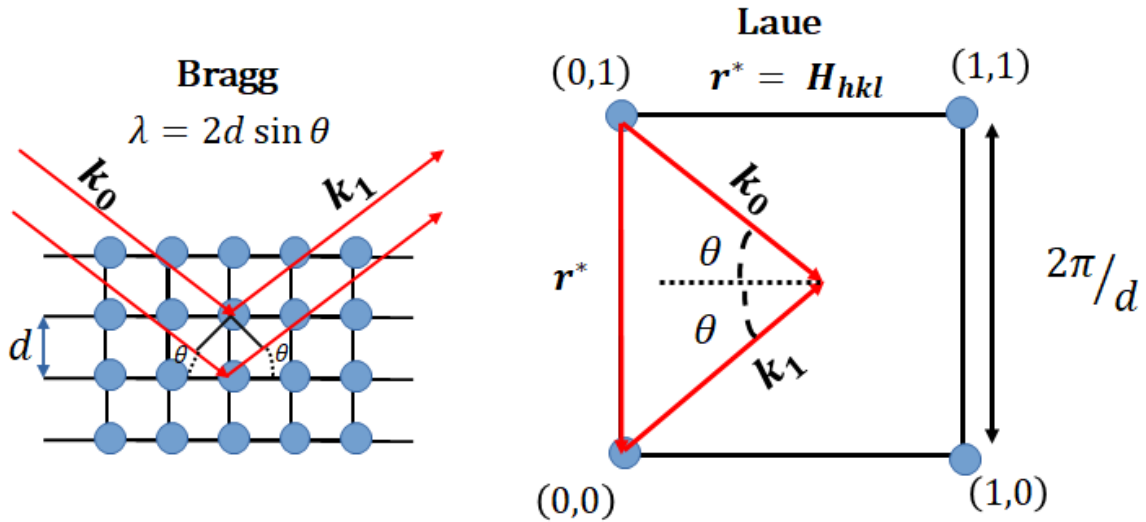


Figure 2.4: The equivalence of Bragg's Law and the Laue condition for the particular case of the 2D square lattice

It is necessary to mention a relationship between points in reciprocal space and planes in the direct lattice. For each reciprocal lattice point given by Eqn. 2.15, there is a set of planes in the direct lattice such that:

1.  $\mathbf{H}_{hkl}$  is orthogonal to the planes with Miller indices  $(h, k, l)$ ;
2.  $|\mathbf{H}_{hkl}| = \frac{2\pi}{d_{hkl}}$ , where  $d_{hkl}$  is the lattice spacing of the  $(h, k, l)$  planes.

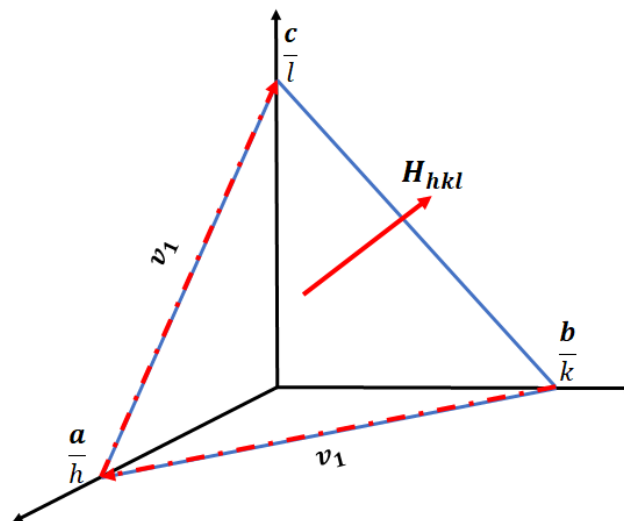


Figure 2.5: The figure to prove that the reciprocal lattice vector  $\mathbf{H}_{hkl}$  is orthogonal to the  $(h, k, l)$  planes, and its magnitude equals to  $\frac{2\pi}{d_{hkl}}$ .

From (Fig. 2.5), we can easily establish the following relationship between two vectors  $v_1$  and  $v_2$  in the plane with Miller indices  $(h, k, l)$

$$\mathbf{v}_1 = \frac{\mathbf{c}}{l} - \frac{\mathbf{a}}{h}; \mathbf{v}_2 = \frac{\mathbf{a}}{h} - \frac{\mathbf{b}}{k} \quad (2.23)$$

We could define any point in the plane as  $\mathbf{v} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2$ , where  $\alpha_1$  and  $\alpha_2$  are scalars. Based on Eqn. 2.16, the scalar product of  $\mathbf{H}_{hkl}$  and  $\mathbf{v}$  is

$$\mathbf{H}_{hkl} \cdot \mathbf{v} = (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*) \cdot \left( (\alpha_2 - \alpha_1) \frac{\mathbf{a}}{h} - \alpha_2 \frac{\mathbf{b}}{k} + \alpha_1 \frac{\mathbf{c}}{l} \right) = 2\pi(\alpha_2 - \alpha_1 - \alpha_2 + \alpha_1) = 0 \quad (2.24)$$

Eqn. 2.24 proves that  $\mathbf{H}_{hkl}$  is orthogonal to the planes with Miller indices  $(h, k, l)$ . To prove the statement (ii), we need to take the scalar product of  $\hat{\mathbf{H}}_{hkl} = \frac{\mathbf{H}_{hkl}}{|\mathbf{H}_{hkl}|}$  with such a vector connecting the origin to the plane as  $\frac{\mathbf{a}}{h}$ , thus, as required  $d_{hkl}$  is

$$d_{hkl} = \frac{\mathbf{a}}{h} \cdot \frac{\mathbf{H}_{hkl}}{|\mathbf{H}_{hkl}|} = \frac{2\pi}{|\mathbf{H}_{hkl}|} \quad (2.25)$$

We will complete the general proof before we prove that Bragg's Law is equivalent to the Laue condition. We could rewrite the Laue condition as  $\mathbf{k}_1 = \mathbf{k}_0 + \mathbf{H}_{hkl}$ . Then we take the square of both sides and using the fact that the scattering is elastic, and if  $\mathbf{H}_{hkl}$  is a reciprocal lattice vector, then  $-\mathbf{H}_{hkl}$ , we will get the following result

$$\mathbf{H}_{hkl}^2 = 2\mathbf{H}_{hkl} \cdot \mathbf{k}_0 \quad (2.26)$$

From (Fig. 2.5) it is obvious that  $\mathbf{H}_{hkl} \cdot \mathbf{k}_0 = H_{hkl} \frac{2\pi}{\lambda} \sin \theta$ . Since we have already shown that  $|\mathbf{H}_{hkl}| = \frac{2\pi}{d_{hkl}}$ , Eqn. 2.26 can be rearranged as  $\lambda = 2d \sin \theta$ .

## 2.9 The Ewald sphere construction

The Ewald sphere provides the visualisation of diffraction events in reciprocal space. It is a geometric representation of the condition  $\mathbf{r}^* = \mathbf{H}_{hkl}$ . Each point of the reciprocal lattice has  $(hkl)$  coordinates satisfied by the Laue condition. The origin ( $h = 0, k = 0, l = 0$ ) is placed at the point where the incident beam intersects the Ewald sphere, see (Fig. 2.6).

In the case of conventional crystallography, the lattice  $(hkl)$  will rotate around the origin of the reciprocal space coordinates due to the rotation of the crystal. As a rule, diffraction can only be observed when the lattice point intersects with the Ewald sphere.

It is also possible to observe multiple scattering when more than one point of the reciprocal lattice falls on the Ewald sphere simultaneously, which leads to the simultaneous observation of several reflections. In the case of a not completely monochromatic beam, the Ewald sphere will have a finite width, and observed reflections will be within the spheres of radius equal to maximum and minimum  $\mathbf{k}$  vector in the beam.

## 2.10 Anomalous signal and Friedel's law

The structure factor  $F_{hkl}$  is the Fourier transform of the electron density  $\rho(\mathbf{r})$ . Considering the electron density  $\rho(\mathbf{r})$  as an approximately real-valued function and taking into account that the Fourier transform of a real function is centrosymmetric, we can formulate Friedel's law: in the reciprocal space, two reflections that are centrosymmetric to each other, the so-called Friedel-pair, have the same amplitude and opposite angles:



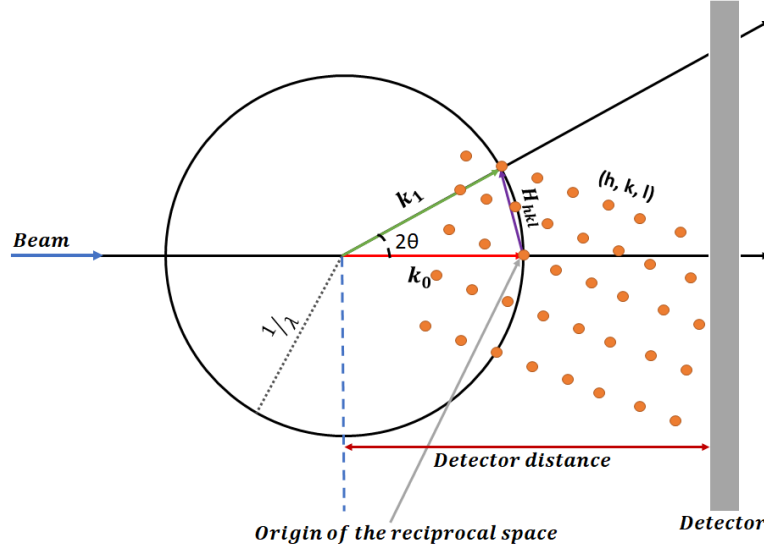


Figure 2.6: The figure demonstrates the principle of Ewald sphere construction in the case of a monochromatic beam incident on a sample: when reciprocal lattice point of the vector  $H_{hkl}$  lies on the Ewald sphere, diffraction condition  $k_1 - k_0 = H_{hkl}$ .

$$|F_{hkl}| = |F_{-h-k-l}|, \phi_{hkl} = \phi_{-h-k-l} \quad (2.27)$$

However, in the case of anomalous scattering, Friedel's law is broken. Anomalous scattering, or resonant scattering, occurs when the incident X-ray energy is close to a transition energy, bringing the atom into an excited state. The X-ray wavelength is, in this case, close to what is called an absorption edge of a particular atom (i.e., sudden change in the value of the linear absorption coefficient). For wavelength longer than the edge, the absorption is low. When the wavelength is shortened, the X-ray energy increases and a transition can occur, leading to a rise in absorption. The X-rays from the "anomalous" atom are out of phase with the incident beam. As mentioned before, in Section A.1, the total form factor contains more terms considered dispersive. Here, we will formalise statements made in Section A.1. Because the electron is associated with atoms in each orbital according to the laws of quantum mechanics, we could consider them oscillators with a characteristic orbital frequency. The resonance phenomenon occurs when the frequency of the incident wave is close to the natural frequency, which results in so-called anomalous scattering.

The following equation describes the motion of the electron placed in the electric field  $E_0$  of the incident wave with the frequency  $\omega/2\pi$

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \omega_B^2 x = \frac{eE_0}{m} \exp i\omega t \quad (2.28)$$

where  $\omega_B$  is the natural angular frequency of the electron and  $\gamma$  is the damping coefficient.

The solution of the Eqn. 2.28 is

$$x(t) = \frac{eE_0}{m} \frac{\exp i\omega t}{\omega_B^2 - \omega^2 + i\gamma\omega} \quad (2.29)$$

The dipole moment of this oscillating electron can be obtained by multiplication of the displacement  $x(t)$  by  $e$ , and the magnitude of the produced electric field by the dipole oscillator is

$$E = \frac{e^2 E_0 P}{mc^2 r} \frac{\omega^2}{\omega_B^2 - \omega^2 + i\gamma\omega} \quad (2.30)$$

Where  $P$  is the polarisation coefficient, therefore, keeping in mind the Thomson formula (Eqn. 2.3), the scattering amplitude of the electron has a frequency-dependant factor.

$$\frac{E}{E_{Th}} = \frac{\omega^2}{\omega_B^2 - \omega^2 + i\gamma\omega} \quad (2.31)$$

The scattering factor of an atom, in this case, will be a complex number and can be described as

$$f_a = f_0 + f' + if'' \quad (2.32)$$

Where  $f_0$  is the 'normal' scattering factor in the absence of anomalous scattering,  $f'$  and  $f''$  are called the real and imaginary dispersion corrections, respectively, which was more detailed described in Section A.1.  $f'$  and  $f''$  are wavelength-dependent but almost independent of the scattering angle. For example, the Fig. 2.7 shows the variation of the coefficients  $f'$  and  $f''$  with energy for the bromine (Br) atom.  $f'$  is negative. It decreases the effective number of scattering electrons. The imaginary part  $f''$  is positive and corresponds to a scattered wave exactly  $\pi/2$  out of phase with the incident beam. Using Eqn. A.2 or its short version ( $f = f' + if''$ ), we can rewrite the structure factor as follows:

$$F(hkl) = \sum_{j=1}^N f'_j \exp i2\pi(hx_j + ky_j + lz_j) + i \sum_{j=1}^N f''_j \exp i2\pi(hx_j + ky_j + lz_j) \quad (2.33)$$

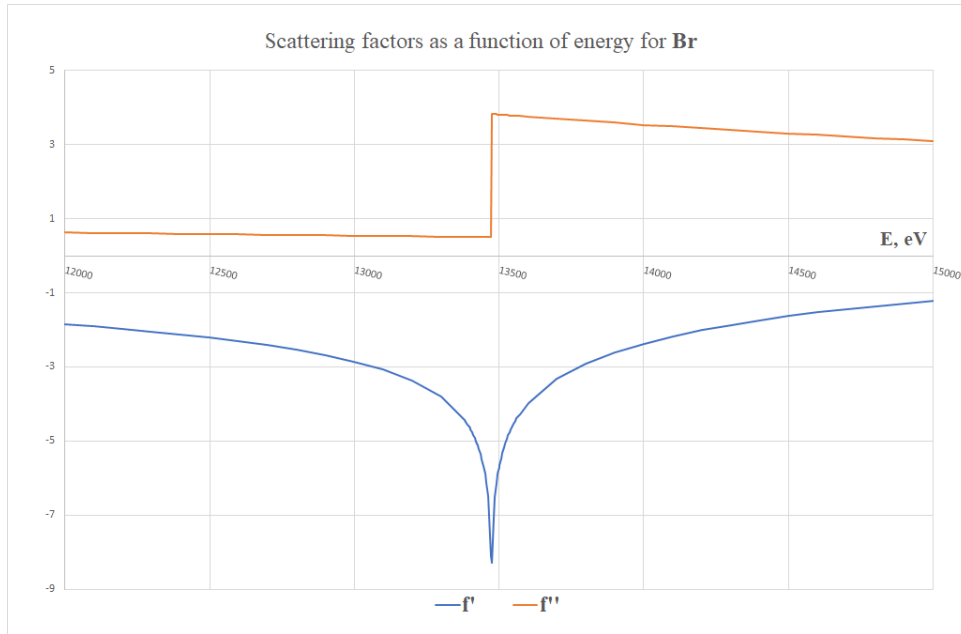


Figure 2.7: Scattering factors as a function of energy for Br (<http://skuld.bmsc.washington.edu/scatter/data/Br.dat>)

If we imply Euler's formula ( $\exp ix = \cos x + i \sin x$ ), the Eqn. 2.33 can be represented in the following way:

$$F(hkl) = (A_{hkl} - D_{hkl}) + i(B_{hkl} + C_{hkl}) \quad (2.34)$$

where

$$A_{hkl} = \sum_j f'_j \cos 2\pi((hx_j + ky_j + lz_j)) \quad (2.35)$$

$$B_{hkl} = \sum_j f'_j \sin 2\pi((hx_j + ky_j + lz_j)) \quad (2.36)$$

$$C_{hkl} = \sum_j f''_j \cos 2\pi((hx_j + ky_j + lz_j)) \quad (2.37)$$

$$D_{hkl} = \sum_j f''_j \sin 2\pi((hx_j + ky_j + lz_j)) \quad (2.38)$$

The structure factor of the reflection  $(-h - k - l)$  can be expressed as:

$$F(-h - k - l) = (A_{hkl} + D_{hkl}) + i(C_{hkl} - B_{hkl}) \quad (2.39)$$

It is easy to show that:

$$|F(hkl)|^2 - |F(-h - k - l)|^2 = -4A_{hkl}D_{hkl} + 4B_{hkl}C_{hkl} \quad (2.40)$$

From Eqn. 2.40, it can be seen that Friedel's law is not verified anymore.

As a result, the atomic scattering factors display strong deviation from the Thomson scattering when the incident beam energy is close to the atomic absorption edges, and in general, due to the imaginary term  $f''$ , Friedel's law is not valid in the presence of the anomalous scattering.



# X-ray sources

In 1895, Wilhelm Roentgen discovered X-rays; since then, they have been considered an invaluable tool for studying the structure of matter. By the mid-1970s, due to the limitations of the X-ray source invented in 1912, our theoretical understanding of the interaction of X-rays with matter and our knowledge of how to use them experimentally had progressed. In the 1970s, it became clear that the synchrotron radiation emitted by charged particles circulating in storage rings designed for high-energy nuclear physics experiments had the potential to be a much more intense and versatile source of X-rays. Thus, many storage rings have been constructed around the world dedicated solely to the production of X-rays. This has led to the emergence of so-called third-generation synchrotron sources, brighter than early laboratory sources by about  $10^{12}$  times. With the advent of synchrotron sources, innovation in X-ray science has accelerated markedly. More new facilities with different principles and components have appeared, such as X-ray free electron lasers (FELs) and 4<sup>th</sup> generation of synchrotrons. These modern X-ray sources generate highly intense and coherent X-ray beams, coupled with improved focusing optics that enhance the flux density at the sample. In parallel, equally significant has been the concurrent advancement in detector technology. In this chapter, the major focus lies on the X-ray sources, their main characteristics and detector technologies adjusted to specific needs and constraints. Discussing the key principles of operating modern facilities will elucidate the limitations of some data collection techniques and corresponding proper data analysis pipelines observed in Chapter 4. In the meantime, delving into different detectors will ease the understanding of the first steps of preprocessing raw data, known as detector calibration, detector geometry refinement (see Chapter 5), and how various detectors contribute to collected diffraction patterns which will significantly influence on employing lossless compression algorithms described in Chapter 7.

## 3.1 Radiation characteristics

### 3.1.1 Flux, Brightness, Brilliance

Defining the figures of merit used to describe and compare X-ray sources is necessary. The flux of a source is the number of photons per second per unit area:

$$\Phi = \frac{dN_{ph}}{dSdt} \quad (3.1)$$

Brightness is the flux per unit solid angle:

$$B = \frac{d\Phi}{d\Omega} \quad (3.2)$$

Brilliance is the number of photons within a bandwidth of 0.1% centred around a certain frequency per second per unit area per unit solid angle, or in other words, it states how the flux is distributed in space and angular range:

$$Brilliance = \frac{\text{photons/second}}{(\text{mm}^2 \text{ source area})(\text{mrad})^2(0.1\% \text{ bandwidth})} = \frac{d^2\Phi}{d\omega d\Omega} \quad (3.3)$$

And is, therefore, the flux per unit source area and unit solid angle or flux per total emittance.

### 3.1.2 Emittance

From Eqn. 3.3, we can see that the brilliance is inversely proportional to both the source size and the X-ray beam divergence. The emittance  $\epsilon$  is the product of the linear source size  $\sigma$  and the beam divergence  $\sigma'$  in the same plane:

$$\epsilon_x = \sigma_x \sigma'_x, \epsilon_y = \sigma_y \sigma'_y, \quad (3.4)$$

Hence, achieving a low emittance, where an extremely small source emits nearly perfectly parallel X-rays, becomes highly desirable. The emittance in a given transverse direction remains a constant, primarily determined by the magnet lattice, for each synchrotron storage ring. Consequently, the objective is to minimise this constant.

The total source size  $\sigma_i$  and the beam divergence  $\sigma'_i$ , where  $i = (x, y)$ , in  $i$ -plane perpendicular to the propagation of a given storage ring, are convolutions of contributions from the electron and photon beams:

$$\begin{aligned} \sigma_i &= [(\sigma_i^e)^2 + (\sigma^p)^2]^{1/2} \\ \sigma'_i &= [(\sigma_i'^e)^2 + (\sigma'^p)^2]^{1/2} \end{aligned} \quad (3.5)$$

The fundamental lower limit to the total emittance, given by Heisenberg's uncertainty principle, is

$$\epsilon^p = \sigma^p \sigma'^p = \frac{\lambda}{4\pi} = \frac{98.66[\text{pm rad}]}{E \text{keV}} \quad (3.6)$$

In third-generation storage rings, electron beam contributions ( $\sigma^e$  and  $\sigma'^e$ ) dominate, significantly causing the total emittance to exceed the Heisenberg limit in the horizontal plane. In fourth-generation DLSRs, which set the modern standard for storage ring design, electron emittance has been lowered to levels approaching or falling below the intrinsic photon emittance, especially for soft X-rays.

Another important relationship between the total source size  $\sigma_i$ ,  $i = (x, y)$  and the full-widths at half-maximum ( $FWHM_i$ ) is

$$FWHM_i = \sqrt{8 \ln 2} \sigma_i = 2.355 \sigma_i \quad (3.7)$$

To maintain an electron in a closed orbit within a storage ring, dipole bending magnets alone cannot suffice when the electron deviates from the ideal reference orbit and possesses nonzero transverse momentum. Pairs of quadrupole magnets are employed with alternating vertical and horizontal focusing to correct and restore the electrons to their ideal path. Consequently, while the emittances remain constant, the beam shape characterised by  $\sigma_x$  and  $\sigma_y$  changes the magnet lattice. The following formulas for the "beta function" aid in quantifying these variations within the storage ring.

$$\begin{aligned} \beta_x &= \sigma_x / \sigma'_x \\ \beta_y &= \sigma_y / \sigma'_y \end{aligned} \quad (3.8)$$

If we insert Eqn. 3.8 into Eqn. 3.4 equations, we will obtain

$$\begin{aligned}\sigma_x &= \sqrt{\epsilon_x \beta_x} \\ \sigma_y &= \sqrt{\epsilon_y \beta_y}\end{aligned}\tag{3.9}$$

As mentioned before, the emittance is a constant for any given storage ring, and focusing the electron beam results in a small beam size with a significant divergence (low  $\beta$ ). Meanwhile, a broader beam will be more parallel (high  $\beta$ ). The preferred combination of  $\beta_x$  and  $\beta_y$  varies around the ring depending on the elements of the magnet lattice through which the electron beam passes.

The storage ring's emittance is determined by the interplay of two opposing factors - radiation damping and quantum excitation. Radiation damping, driven by axial acceleration from the RF cavity, reduces the electron's angular deviation from the ideal orbit, minimising transverse momenta. To maximise radiation damping, high-field bending magnets and/or damping wigglers can be introduced.

The opposite effect of quantum excitation directs electrons into oscillatory paths within an ideal orbital due to the emission of a photon and the subsequent loss of energy by the electron. This leads to a stochastic transverse electron momenta distribution, significantly increasing emittance. Quantum excitation can be minimised by designing the magnetic array so that the dispersion of electron energy is minimised at the main locations of radiation, namely the bending magnets. This can be achieved by horizontal focusing at the bends and using many small-deflection-angle bends in multi-bend-achromat lattices to limit dispersion growth.

### 3.1.3 Coherence

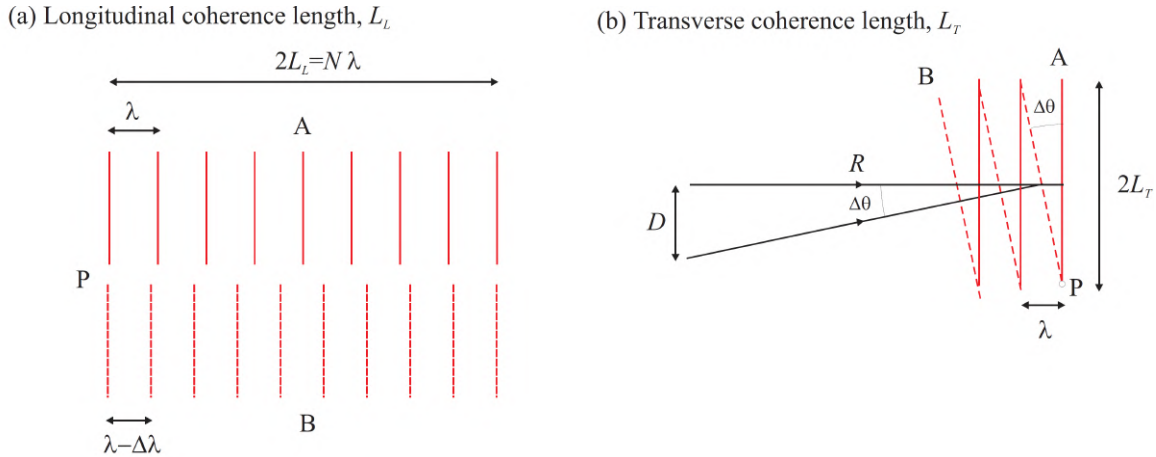


Figure 3.1: It is an original figure taken from [39] that demonstrates longitudinal and transverse coherence lengths.

In this section, we will discuss the concept of beam coherence length and its relation to the source and the monochromator, see Fig. 3.1. In reality, no X-ray source has an infinitely narrow bandwidth. The different frequencies within the bandwidth of the beam will sooner or later drift out of phase with one another.

Consider two wave planes, A and B, which have slightly different wavelengths  $\lambda$  and  $\lambda - \Delta\lambda$ , respectively, and propagate in the same direction. The two waves are exactly in phase at wavefront P. But they will be out of phase after passing the so-called longitudinal (or temporal) coherence length  $L_L$  and will be in phase again after passing  $2L_L$ . The formula for  $L_L$  is

$$L_L = \frac{1}{2} \frac{\lambda^2}{\Delta\lambda}\tag{3.10}$$

The longitudinal coherence after a monochromator is usually determined by the quality of the crystal or grating used in the monochromator, which defines  $\frac{\lambda^2}{\Delta\lambda}$ .

If we have two waves, A and B, of the same wavelength but propagate in a slightly different direction, say by an angle of  $\Delta\theta$ . Their wave-fronts coincide at point P and will be out of the phase after travelling the distance equal to the transverse coherence length (also called the spatial coherence length)

$$L_T = \frac{\lambda R}{2 D} \quad (3.11)$$

$R$  - is the distance from the observation point P to the source, and  $D$  is the lateral extent of the finite source. This arises because all sources have a finite size  $D$  (for example, the size of the pinhole) and a nonzero divergence  $\Delta\theta$ . If we assume the source has a Gaussian profile, to determine  $D$ , we need to integrate interference contributions across the entire source's intensity distribution. It turns out that  $L_T$  refers to the standard deviation of the beam  $\sigma_i$  by

$$L_T[\mu\text{m}] = 28.21 \frac{\lambda[\text{\AA}]R[\text{m}]}{\sigma_i[\mu\text{m}]} \quad (3.12)$$

A finite coherence length imposes an upper limit on how far apart two objects can be for interference effects to occur. For instance, in the case of two scattering electrons, if their separation projected onto the wave-vector transfer greatly exceeds the coherence length, the total scattered intensity results from the sum of individual electron scattering intensities rather than the squared modulus of the sum of amplitudes.

### 3.2 The standard X-ray tube

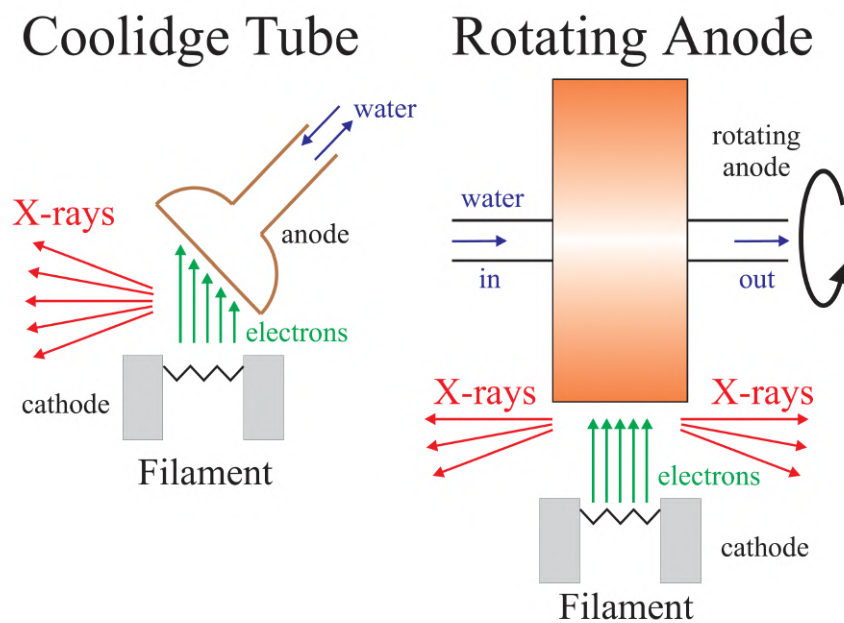


Figure 3.2: The picture is originally from [39]. The left plot is the standard X-ray tube developed by Coolidge around 1912. The right plot demonstrates the rotating anode.

In November 1895, Röntgen discovered X-rays in his University of Würzburg laboratory in Germany using vacuum tubes, leading to the invention of an X-ray tube. In 1912, W. D. Coolidge of the General Electric Research Laboratory in New York developed a more efficient tube, allowing independent voltage and current



control, with a maximum power of about 1 kW (the left plot in Fig. 3.2). This Coolidge tube served as the standard X-ray tube for many decades, but in the 1960s, rotating anode generators were introduced, dissipating heat more effectively and increasing total power (see the right plot in Fig. 3.2).

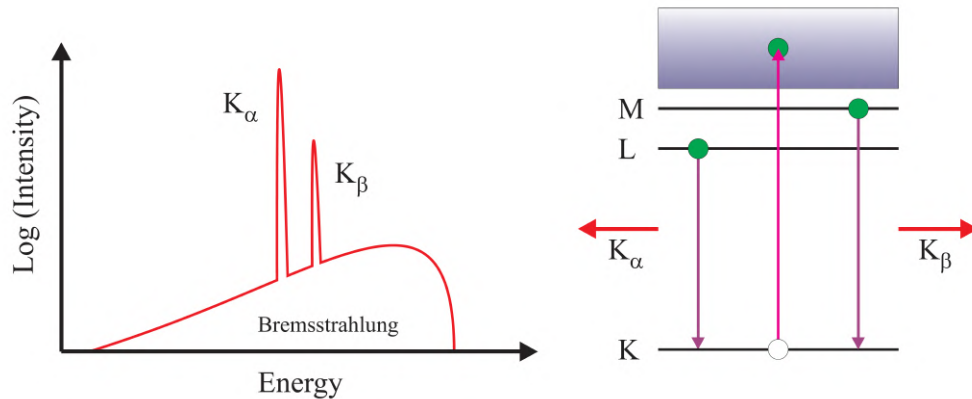


Figure 3.3: The left plot shows the spectrum from an X-ray tube. This spectrum has discrete fluorescent lines superimposed on the continuous bremsstrahlung radiation. On the right plot, we can see the Schematic atomic energy level diagram: the  $K_{\alpha}$  line results from transitions between an L and K shell, whereas the  $K_{\beta}$  - from an M to K transition.

Two distinct components are observed in the X-ray spectrum produced when electrons interact with a metal anode, as shown in Fig. 3.3. The continuous component results from electron deceleration, known as bremsstrahlung, with its maximum energy determined by the applied high voltage. Superimposed on this broad spectrum is a sharp line spectrum. When electrons collide with atoms, they can eject inner-shell atomic electrons, creating vacancies. Subsequent electron relaxation from outer shells generates X-rays with characteristic energies corresponding to the shell energy differences, known as fluorescent radiation. The  $K_{\alpha}$  line is commonly used for monochromatic beams due to its high intensity compared to the bremsstrahlung spectrum. However, it has limitations, such as a limited angular divergence and the inability to tune the wavelength continuously. In contrast, synchrotron-generated X-rays offer a solution, overcoming these limitations and providing significantly higher brightness than standard laboratory sources.

### 3.3 Synchrotrons

Synchrotron radiation results from charged particles moving at relativistic speeds within applied magnetic fields, causing them to follow curved trajectories. Electrons in synchrotrons (see Fig. 3.4) are first generated in an electron gun, then accelerated and injected into a booster ring. In this smaller ring, they attain the energy level required for the storage ring, where they are subsequently transferred. Besides synchrotrons, synchrotron radiation is also generated in storage rings where electrons or positrons circulate at a constant energy level. Storage ring sizes range from tens of meters to over two kilometres. Within a storage ring, synchrotron radiation is produced in bending magnets that maintain electron orbits or in insertion devices like wigglers or undulators placed in the straight sections. These devices create oscillating electron paths due to alternating magnetic fields. The oscillations have a significant amplitude in wigglers, resulting in incoherent radiation addition from various oscillations. In contrast, undulators produce coherent radiation by combining the small amplitude oscillations from individual electrons' passages.

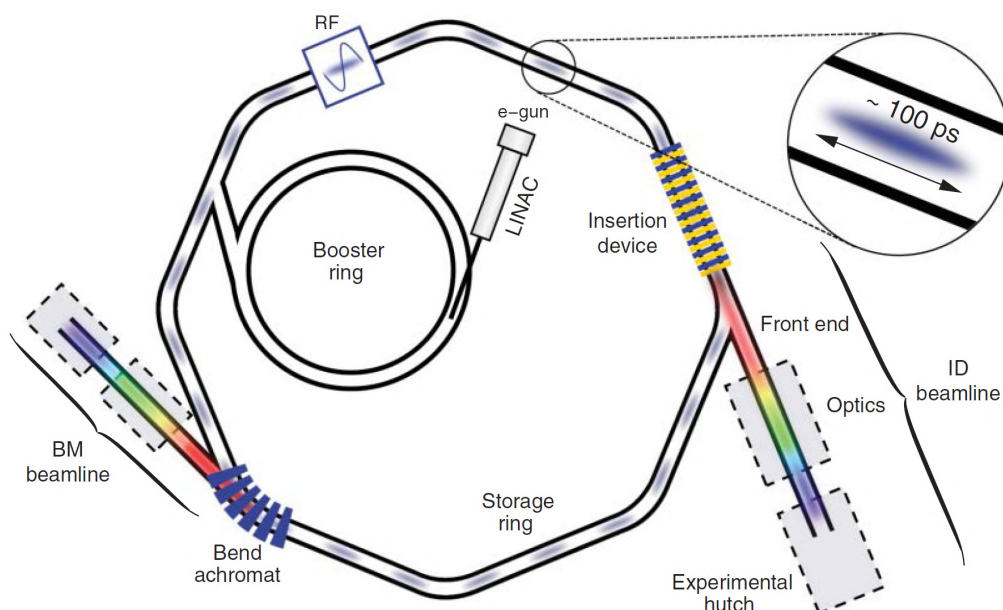


Figure 3.4: The image is sourced from [40] and illustrates the key components of a modern synchrotron facility. An electron gun with a heated filament generates electrons, which undergo initial acceleration in a linear accelerator (LINAC) before entering an evacuated booster ring for further acceleration. These accelerated electrons are injected into a storage ring, where bending-magnet achromats in arc sections maintain their closed path. Straight sections between arcs feature insertion devices (IDs) like wigglers/undulators. The emitted radiation from these insertion devices and bending magnets are utilised at beamlines along the emission axes. A radio frequency (RF) supply replenishes the electrons' energy to compensate for energy loss due to synchrotron light radiation.

Third-generation storage rings are designed to optimise the photon flux and brightness required for modern experiments. Fourth-generation 'diffraction-limited storage rings' (DLSRs) take this a step further by significantly reducing the total emittance, dominated by the electron beam in third-generation facilities, by up to two orders of magnitude.

The high brilliance of synchrotrons is attributed to several factors. Firstly, the radiation source size results from convoluting the photon source size with the electron beam's transverse size. This value is approximately 100 micrometres in third-generation facilities, but it is an order of magnitude smaller in fourth-generation DLSRs. Secondly, synchrotrons emit an extraordinary amount of light. The emitted flux is directly proportional to the square of the electron's acceleration. Since centripetal acceleration in the storage ring is proportional to  $\gamma^2$ , the flux increases proportionally to  $\gamma^4$  for high-energy synchrotron storage rings. This results in exceptionally high brilliance.

### 3.3.1 An advanced High-Throughput Pharmaceutical X-ray screening instrument (HiPhaX)

For rational drug design, protein structures play a crucial role, and major pharmaceutical companies possess well-developed in-house structure-based drug design capabilities. One increasingly popular method involves X-ray screening of fragment libraries and, more recently, complex compound libraries. This approach aims to identify potential compounds as starting points for drug discovery, serving as an alternative or complement to biochemical and biophysical high-throughput screening methods [41]. X-ray screening offers a distinct advantage by providing detailed 3D structural information on atomic interactions and binding modes. This information

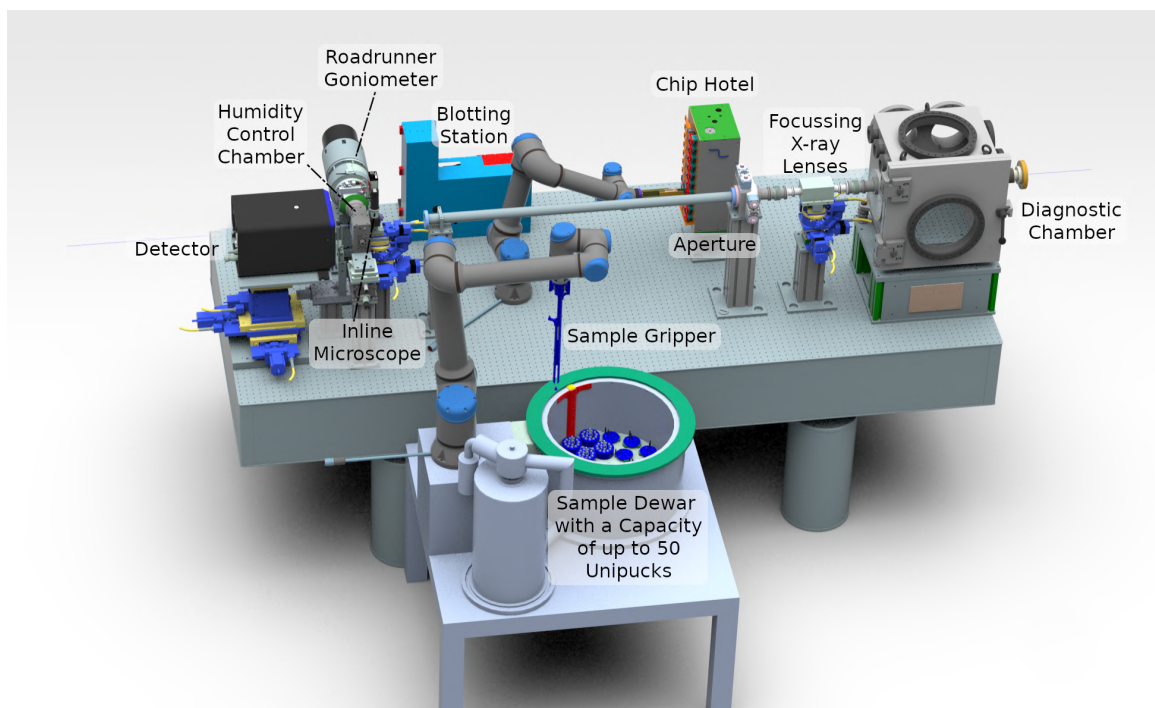


Figure 3.5: HiPhaX experimental hutch overview

can be used for subsequent computational fragment extension and compound optimisation procedures. Despite its benefits, three main limitations are hindering the full exploitation of X-ray screening in drug discovery: the manual intervention required in the handling of crystals and X-ray data collection procedures, which severely limit the throughput, and the challenges associated with expanding initial hits into highly potent and selective lead compounds.

The High-Throughput Pharmaceutical X-ray screening instrument (HiPhaX) (P09 beamline at the PETRA III) is a new instrument for macromolecular crystallography exclusively dedicated to high-throughput X-ray fragment and compound screening. The goal of this beamline is to set up a fully automated end station with protein structures without any human intervention.

The beamline operates at a fixed 16 keV X-ray energy with a 15-20  $\mu\text{m}$  spot at the sample position. A key focus is data collection at non-cryogenic temperatures and controlled humidity. The beamline has a high-precision Roadrunner goniometer, allowing for conventional single-crystal rotation data collection and high-speed fixed-target serial crystallography experiments.

A special emphasis on HiPhaX lies in experiments at non-cryogenic temperatures in a humidity-controlled atmosphere. The environmental chamber adjusts the temperature between 0-60 °C and the relative humidity between 30–100% r.h. An overview of the HiPhaX experimental hutch and its installations is represented in Fig. 3.5.

This state-of-the-art beamline is designed to run protein crystallography experiments using a variety of sample delivery systems such as loops, chips, and tape drives while supporting multiple detectors, including Lambda (1.5M), Pilatus CdTe (2M), Eiger 4M, and Pilatus 6M. Researchers utilising the P09 beamline have the flexibility to choose between rotational or serial crystallography techniques based on factors like sample availability and their specific scientific objectives, including time-resolved studies and T or pH jump investigations. For efficient serial data collection, crystals are presented on so-called ‘fixed-target’ sample holders to the X-ray beam. Different designs and sizes of these sample holders are available and will be provided for free to the instrument

users. For screening applications, compartmented chips, which provide space for up to 12 different compounds on a single chip, are recommended. Compounds can be either applied to pre-grown crystals by soaking or by co-crystallization. For efficient automated fixed-target serial data collection, the instrument is equipped with a chip storage hotel with capacity for 10 large Roadrunner II chips and a robotic arm for fully automated chip exchange.

In addition to pure screening applications, the setup at HiPhaX allows systematically exploring the influence of relative humidity and temperature on protein structures and their conformational flexibility, which is valuable additional insight for computational ligand binding predictions.

For experiments at cryogenic temperatures, the instrument is equipped with a large LN2 storage Dewar with a capacity for 868 samples mounted on uni-pucks. The sample changer is equipped with a novel magnetic sample gripper, providing extremely high reliability as it does not involve any mechanically moving parts. Samples exchange times are below 20 seconds. The instrument further offers automatic X-ray centring of the crystals in the beam, based on the work described in Section 5.3.

The mentioned further advancements in diffraction analysis methods and data reduction techniques for SX found their application at the P09 beamline and were excessively tested and later integrated into the software control system, which is discussed in detail in Section 5.6. Thus, during the measurements, the beamline provides real-time feedback about auto-processing of the diffraction data.

### 3.4 Free-electron lasers

Fourth-generation DLSR radiation shares similarities with laser radiation, including high intensity, parallel propagation, and partial monochromaticity, particularly in undulator sources. However, the potential of undulators at synchrotrons for radiation could be significantly enhanced. The radiation from different electrons crossing the undulator in a bunch is incoherent, resembling an electron gas due to the lack of positional order. To overcome this, electrons within the bunch could be organised into smaller micro-bunches, each containing an average of  $N_q$  electrons, where  $N_q \gg 1$ , with separations equal to the X-ray wavelength. This arrangement would ensure that the radiation from one micro-bunch is in phase with subsequent ones. Consequently, the charge within a single micro-bunch,  $eN_q$ , is much larger than  $e$ , and as the micro-bunch is confined within a distance shorter than the emitted wavelength, this charge can be treated as point-like. In this scenario, the brightness would increase by  $N_q^2$  compared to a conventional undulator.

In an undulator, the radiation field starts at zero at the entrance and reaches its full intensity at the exit. As an electron moves through the undulator, it encounters magnetic forces from the lattice and interacts with radiation fields from other electrons in the bunch. This interaction, spatially modulated with an X-ray wavelength period, leads to micro-bunches forming within the electron density. Once initiated, this effect is amplified as the radiation field gains strength downstream. This phenomenon, known as self-amplified stimulated emission (SASE), transforms an undulator into a free-electron laser (XFEL) designed to exploit the SASE principle [42]. Typical XFEL architecture is presented in Fig. 3.6.

For SASE to work effectively, the radiation field must be strong enough to induce micro-bunching, and the electron gas density plays a critical role. In third-generation storage rings, even with low emittance, the electron density often falls short because the bunch length, approximately  $100 \text{ ps} \times c$ , where  $c$  is the speed of light, is too large. One solution involves utilising a linear accelerator (LINAC) to generate small electron beams with diameters around  $100 \mu\text{m}$  (FWHM) and minimal angular divergence of approximately  $1 \mu\text{rad}$ . This approach boosts the electron density and enhances the SASE mechanism.

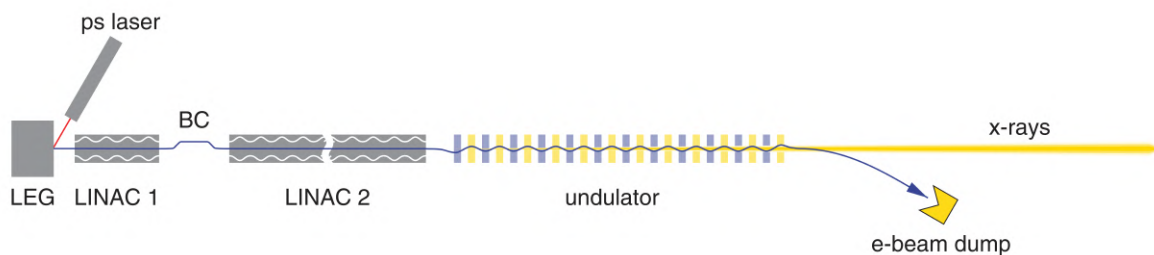


Figure 3.6: The figure is sourced from [40] and provides an overview of XFEL (X-ray Free-Electron Laser) facilities. It begins with a low-emittance gun (LEG) exposed to picosecond laser pulses, emitting electrons. These electrons accelerate in a short LINAC 1 and longitudinal compression via multiple bunch-compressor magnet chicanes (BC). Following this, they experience further acceleration in a much longer LINAC 2 before entering a lengthy undulator, often spanning hundreds of meters. The SASE (Self-Amplified Stimulated Emission) process within the undulator generates high-intensity X-ray pulses lasting approximately 50 fs. Subsequently, the electrons are deflected by a bending magnet and then dumped.

Understanding biomolecular processes and industrial chemical reactions at the atomic level is a crucial scientific pursuit. To capture chemical reactions with 1-angstrom precision, one needs incredibly fast flashes, approximately 100 fs in duration, since atoms move at speeds around 1000 m/s. Many dynamic processes, like bond formation and breaking, occur within femtoseconds to tens of femtoseconds. Scientists utilise lasers to study vibrationally and electronically excited states, but lasers have limitations; their wavelengths are much larger than typical interatomic distances. XFEL has revolutionised our ability to explore material dynamics, properties, and structure with unprecedented spatial precision, operating on a timescale thousands of times shorter than possible with synchrotron facilities.

Synchrotrons and XFELs produce pulsed X-radiation but differ significantly in terms of time structure and peak intensities (other properties can be seen in Table 3.1). Synchrotrons generate X-ray pulses at a rate of hundreds of millions per second, while XFELs vary from a few tens per second to over a million. Moreover, the energy in each synchrotron pulse contains tens of thousands of photons, whereas XFEL pulses can contain trillions of photons. XFEL pulses are also over a thousand times shorter than synchrotron pulses, resulting in peak power levels about 10 billion times higher than synchrotron. Certain experiments that can be conducted at synchrotrons are unfeasible at XFELs, mainly due to the exceedingly high X-ray peak powers, which would instantly damage the samples. However, high photon arrival rates are generally desirable in most X-ray experiments, even if they do not need XFEL-level intensities.

Table 3.1: Comparison of orders-of-magnitude synchrotron- and XFEL (Linac Coherent Light Source (LCLS)) properties

Property	Synchrotron	XFEL (LCLS)
<b>Pulse duration</b>	50-400 ps	5-50 fs
<b>Average flux***</b>	$2 \times 10^{14}$	$10^{14}$
<b>Peak flux****</b>	$6 \times 10^{15}$	$2 \times 10^{25}$
<b>Peak power</b>	1 W*	$10^{11}$ W**
<b>Average power</b>	25 mW*	600 mW to 140 W**

\* - after Si(111) monochromator

\*\* - Unmonochromatized, full SASE spectrum

\*\*\* - Photons/s/%/0.1% bandwidth

### 3.5 Detectors

Detectors for X-ray radiation play a crucial role in capturing and quantifying diffracted X-ray photons. Historically, photographic plates were one of the earliest methods for X-ray detection. They were paired with phosphorescent screens in contact with the emulsion. X-rays striking the phosphor screen emitted visible light, which exposed the film. However, photographic plates had limitations for quantitative analysis due to their non-linear response to signal intensity, poor dynamic range, spatial resolution, and lengthy read-out times.

Scintillation counters represent another detection method. In these detectors, absorbed X-rays undergo partial conversion into visible or near-visible light, which can be further amplified using a photo-multiplier tube (PMT). Typically, inorganic scintillator materials consist of salts or metal oxides doped with high-Z materials. When an X-ray photon is absorbed, the host material becomes electronically excited. This excited state efficiently transfers its energy to nearby states of the dopant ion. These states relax without photon emission to a slightly higher excited state, only a few  $eV$  above the final relaxed ground state. This final state emits a photon in the visible or soft ultra-violet range. Various organic and inorganic scintillator materials offer different temporal characteristics with dead times ranging from a few nanoseconds to several hundred nanoseconds.

The thickness of the scintillator material significantly impacts its X-ray stopping efficiency and signal strength. However, thicker scintillators can negatively affect spatial resolution in imaging applications, leading to increased point spread functions and photon reabsorption. Combining a scintillator with a photo-multiplier tube enables X-ray sensitivity but sacrifices spatial resolution, resulting in bulkier systems.

In some applications, silicon photodiodes are used as beamline diagnostic tools to fine-tune optics and align mirrors. These photodiodes allow for calculating photon flux based on generated current, provided that the diode's dimensions, thickness, and housing materials are well-defined.

Point detectors lack inherent spatial resolution capabilities. Any enhancement in spatial resolution beyond the physical size of their sensitive area relies on the use of slits. One-dimensional detectors like the MYTHEN Microstrip detector enable scans over extensive angular ranges simultaneously, drastically accelerating data acquisition by multiple orders of magnitude compared to zero-dimensional point detectors.

Two-dimensional detectors, also known as area detectors, have become prevalent in scattering and direct imaging experiments. They are increasingly adopted in dispersive spectroscopic setups as well. These detectors encompass CCD arrays, CMOS arrays, and hybrid pixel array detectors. Image plates, which utilised phosphor screens and scanning laser readouts in the past, have become virtually obsolete due to their exceedingly long readout times.

Two prevalent types of modern X-ray detectors are photon counting and integrating detectors. Photon counting detectors continuously update each pixel whenever they detect a single photon while integrating detectors collect the deposited charge throughout the exposure and read it out once the exposure concludes. The main difference between these two types of detectors is depicted in Fig. 3.7.

Photon counting devices operate based on the principle that photons with varying energies arrive randomly at the detector, as seen in diffraction experiments involving elastically scattered diffracted and lower-energy photons generated through photo absorption and subsequent fluorescence. When a photon reaches the detector or a pixel on an area detector, it produces a voltage spike in the detector electronics proportional to its energy. A



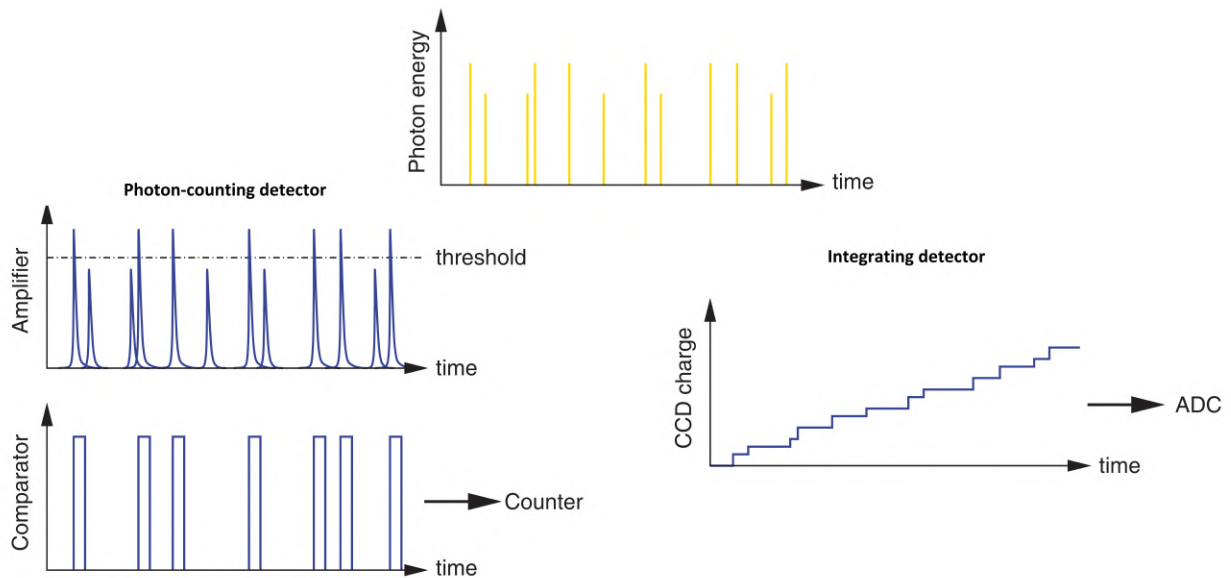


Figure 3.7: The figure, sourced from [40], depicts the interaction of X-rays with a detector. In integrating detectors, charge accumulation occurs, subsequently undergoing conversion through an analogue-to-digital converter (ADC). On the other hand, single-photon counting detectors amplify the charge generated by individual photons, transforming it into a voltage pulse whose amplitude is directly proportional to the photon's energy.

threshold voltage is set within the detector electronics to distinguish between fluorescence photons and elastically scattered ones. This threshold effectively filters out fluorescence, allowing only elastically scattered photons to be registered, incrementing a counter for each detected photon. Photon-counting detectors have a brief dead time after a photon arrives on a pixel, during which the electronics recover and prepare to record the next photon. This dead time, approximately 100 ns, implies that an arrival rate exceeding 10 MHz may result in lost counts due to pile-up.

In contrast, integrating detectors, such as charge-coupled devices (CCDs) and CMOS detectors, accumulate and store the charge generated each time they capture a photon, with the charge being proportional to the photon's energy and number of incident photons. This accumulated charge is converted into a voltage and then transformed into a digital signal only after the exposure.

CCDs consist of connected pixels arranged in rows, transferring charge along shift registers and reading out sequentially. They offer high sensitivity, uniformity, and efficiency but are limited by data transfer bottlenecks and noise introduced during readout. In contrast, CMOS detectors, known as active pixel sensors (APS), perform individual light-to-digital signal conversions on each pixel. They achieve faster readouts through digital multiplexing but sacrifice some light-capturing efficiency and response uniformity compared to CCDs.

Blooming affects both types of sensors. Blooming is spilling out of a bright signal in a certain sensor region to the neighbouring pixels because each pixel can only store a finite amount of charge before it spills over into neighbouring pixels. In the case of CCDs, this problem is compounded by the shift register readout mechanism, which smears the signal in the readout direction.

Photon counting detectors are ideal when photon arrival rates and pixel size are not limiting factors. However, in experiments like XFELs with extremely high photon rates, integrating detectors becomes essential - counting detectors are not fast enough to count photons that arrive within several fs pulse. Integrating detectors like

AGIPD, JungFrau and ePix10k, developed for XFEL experiments, employs capacitors to enhance dynamic range by switching between parallel circuits, providing a practical solution for high photon-rate scenarios.

Hybrid detectors combine two distinct semiconductor materials, widely used in synchrotron applications like macromolecular crystallography. The readout electronics are constructed from low-resistance doped Silicon, connected electrically through metallic bump bonds to a high-purity, high-resistance sensor material. This sensor material can be composed of entirely different materials, such as Silicon or compound semiconductors like GaAs or CdTe. Due to the physical bonding process, reducing the pixel size below approximately 20 microns in linear dimensions becomes challenging. The choice of higher-density materials like GaAs and CdTe for the sensor is primarily motivated by their enhanced absorption efficiency for high-energy photons.



---

# Data collection and analysis in protein crystallography

The first chapter elucidated the foundational principles governing the exploration of the atomic structure of the examined protein using X-ray crystallography. Subsequently, the second chapter delved into diverse X-ray sources, detailing their properties and the development of detector technologies tailored to physical constraints and specific requirements. This chapter overviews various data collection strategies and their corresponding analysis methods.

The primary section examines the main data collection techniques used in crystallography to obtain a complete dataset of the full diffraction intensities. Subsequent to this, the text delves into existing approaches for corresponding data analysis. Following that, an examination of the current data processing pipelines integrated across different beamlines is presented.

It is noteworthy that current automation methods for processing data at beamlines may fall short of meeting the demands of emerging experimental setups and collection strategies. Consequently, in developing data processing pipelines, it is imperative to prioritise adaptability without necessitating a complete restructuring of the existing framework. The imperative point of Chapter 5 emphasises the need for these pipelines to seamlessly adjust to diverse control systems installed at various beamlines.

## 4.1 Data collection techniques in X-ray crystallography

### 4.1.1 Laue crystallography

The Laue method, also known as crystallography using pink beam, helps obtain the full diffraction intensities using polychromatic radiation with wavelengths ranging between  $\lambda_{min}$  and  $\lambda_{max}$ . In this case, the volume encompassing the reflection conditions is limited by the two Ewald spheres with radii of  $1/\lambda_{min}$  and  $1/\lambda_{max}$  as shown in Fig. 4.1. Various cross-sections of the reciprocal lattice nodes are excited by X-rays with different wavelengths, resulting in diffraction. Therefore, the intensities of the reflections that lie completely inside the shell between the limiting Ewald spheres are fully integrated. We can collect a complete dataset of diffraction intensities for further structure determination if we measure the diffraction of a crystal in a sufficient number of different orientations.

Polychromatic diffraction data is much more difficult to interpret compared to monochromatic data. Thus, monochromatic techniques are much more widely used. Nevertheless, more and more works use polychromatic

radiation [43–46].

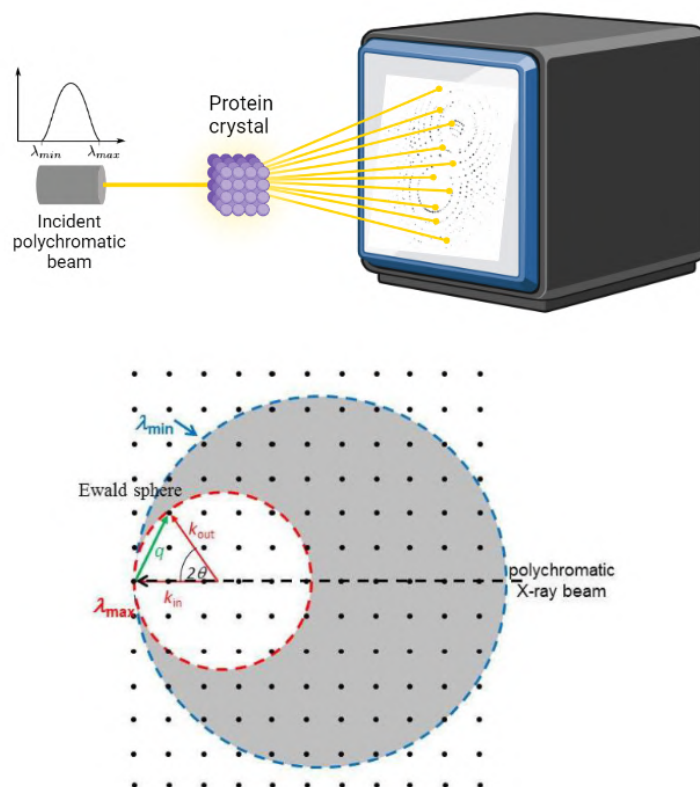


Figure 4.1: Laue method: a stationary crystal is put into the polychromatic X-ray beam with the wavelengths ranging from  $\lambda_{min}$  and  $\lambda_{max}$ . In this case, the Ewald sphere represents a shell within two limiting spheres with radii of  $1/\lambda_{min}$  and  $1/\lambda_{max}$ . Reflections that lie entirely within the shell will be integrated. The image presented is partially sourced from [47].

### 4.1.2 Single crystal rotation

A method such as conventional crystallography uses a series of rotations to collect data using monochromatic X-rays (Fig. 4.2). The crystal is rotated with respect to the incident beam to obtain the full reflection intensities. Typically, an increment of the angle  $\Delta\phi$  between 0.1 and 1 degree is used to record a diffraction pattern. In this case, each node of the reciprocal lattice will completely intersect the Ewald sphere, and its total diffraction intensity will be recorded either in one rotation pattern or in several consecutive patterns.

Due to the experiment's simplicity and relatively straightforward automatic data analysis, single crystal rotation is the most widely used crystallographic data collection method at both laboratory sources and synchrotron radiation facilities. With modern X-ray sources and recent developments in X-ray detectors, the collection of a complete dataset typically takes less than 2 minutes [48]. Thus, it became a well-established, reliable technique for macromolecular structure determination at many synchrotron beamlines worldwide.

### 4.1.3 Powder diffraction

Another possible data collection method is from polycrystalline material or powder. An ideal powder sample consists of many small, randomly oriented crystals. Then, each reciprocal lattice vector  $\mathbf{H}_{hkl}$  will be in all

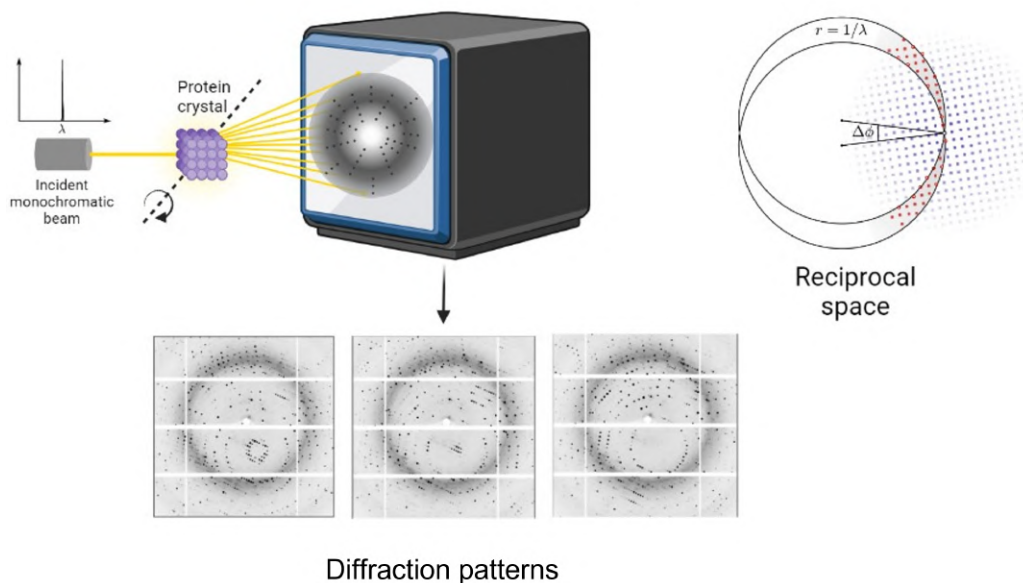


Figure 4.2: Single crystal rotation: the crystal is rotated along one axis in the beam. The rotational motion of the crystal in the beam can be represented as the rotation of the beam relative to a stationary reciprocal lattice. Reflection intensities that lie completely in the volume enclosed by the Ewald sphere during the rotation by the angle  $\Delta\phi$  will be completely integrated. Those reflections that intersect the Ewald sphere but do not lie completely inside the volume will be partially integrated. Nevertheless, their full intensity will be recorded over several consecutive patterns. The image presented is partially sourced from [47].

possible orientations with respect to the incident X-ray beam, forming a sphere of radius  $|\mathbf{H}_{hkl}|$  (Fig. 4.3). Thus, instead of a single point intersecting the Ewald sphere, each reciprocal lattice point will correspond to a circle. The powder diffraction pattern is recorded on the two-dimensional detector placed perpendicular to the incident beam and represents a series of concentric rings. The limitation of the powder method is that the rings begin to overlap at higher resolutions and become unresolvable, especially in macromolecular crystallography, where the unit cells are large. However, the method has been used successfully to determine the structure of proteins and remains a valuable complementary approach for single-crystal measurements [49].

#### 4.1.4 Serial crystallography

Serial crystallography (SX) is a method to investigate the structure of the biomolecule at an atomic level at room temperature [50] with minimal radiation damage, based on the principle of 'diffraction before destruction' [13, 51, 52]. Moreover, this technique can fully exploit the capabilities of modern X-ray sources such as X-ray free electron lasers (X-ray free-electron laser (XFEL)s) [9, 12] and 4<sup>th</sup> generation of synchrotrons [10, 11], and can also use modern detectors at full speed up to MHz. Nevertheless, the main drawbacks of such a method are the amount of data generated and the waste of many samples.

Many diffraction patterns are required for the 3D structure determination of the molecule because the X-rays are only exposed to each crystal at the time in the SX experiment. A proper sample delivery method, adjusted to the X-ray source's and the detector's properties, should be used to obtain a complete data set in the SX experiment. Various sample delivery methods, such as the use of injectors [53–57],[58] by acoustic droplet injection coupled with a conveyer-belt drive [59], etc. More detailed information on various sample delivery

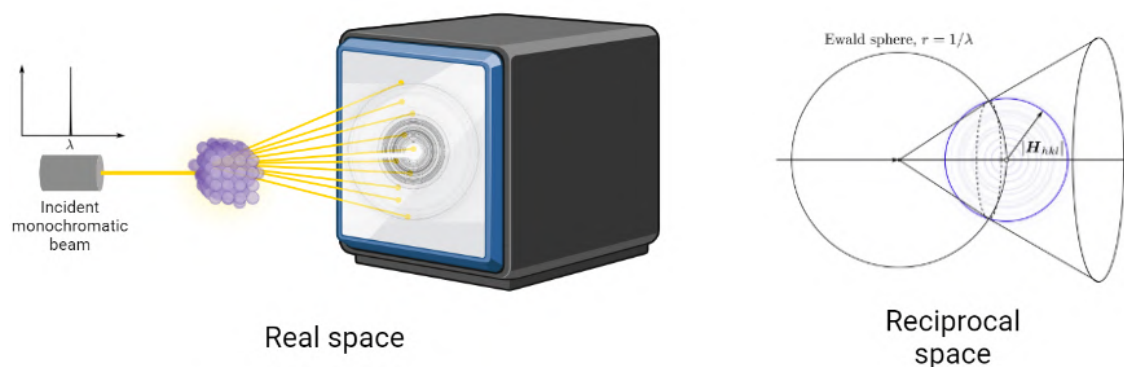


Figure 4.3: Powder Diffraction: the diffraction pattern of powder or micro-crystalline samples represents a series of concentric circles, or in other words, the reciprocal lattice becomes a series of concentric spheres corresponding to each reciprocal lattice vector, centred at the beginning of the reciprocal space. The image presented is partially sourced from [47].

methods can be found in Section 4.1.8. A typical experimental setup for SX is illustrated in Fig. 4.4. The main idea of such an experiment is to deliver samples to the X-ray interaction region in a serial way, during which the data are taken from many copies of similar samples.

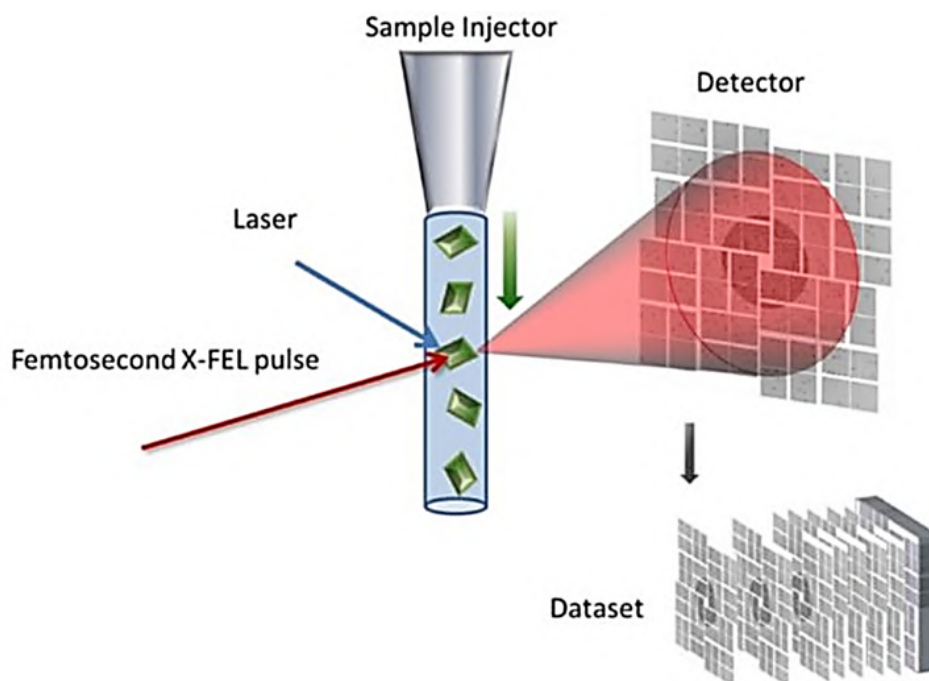


Figure 4.4: Scheme of a typical setup for serial crystallography experiment.

#### 4.1.5 Time-resolved crystallography

During the reaction, the molecule passes through a series of intermediate states (Fig. 4.6). In an enzymatic reaction, these steps lead from substrates to products. Chemical or physical trapping techniques can be used to determine the associated structures by increasing the lifetime of the molecules in the desired state and trapping a structurally homogeneous species. An obvious trapping mechanism is to initiate the reaction, wait for the reaction to proceed sufficiently, and then rapidly cool the crystal [60]. However, substrate analogues that block

further reaction development are often preferable. A disadvantage of this method is the loss of information on the rate at which the reaction occurs. The basic concept behind these methods is that

1. such macromolecular activity can be triggered within a crystal, and the structure of intermediate states can be characterised only if the macromolecule is active in the crystalline state;
2. efficient and synchronous activation of activity (at room temperature) can be achieved for all (or most) of the studied molecules;
3. structural information can be recorded on a time scale shorter than the lifetime of the intermediate state.

Unlike trapping a crystal in specific states, time-resolved crystallography follows the evolution of a spatially averaged structure in real-time [60–62].

X-ray free-electron lasers provide unique ultra-short and very bright pulses and pave the way for significant advances in time-resolved research in SX experiments (time-resolved serial femtosecond crystallography (TR-SFX)) to observe laser-induced protein dynamics at time scales down to sub-picoseconds [8, 63–69]. With TR-SFX, it is possible to directly visualise biomacromolecules in the intermediate state within the crystal and under their near-physiological conditions with high spatial and temporal resolutions [8, 67, 68]. The basic idea of the TR-SFX experiment is presented in Fig. 4.5.

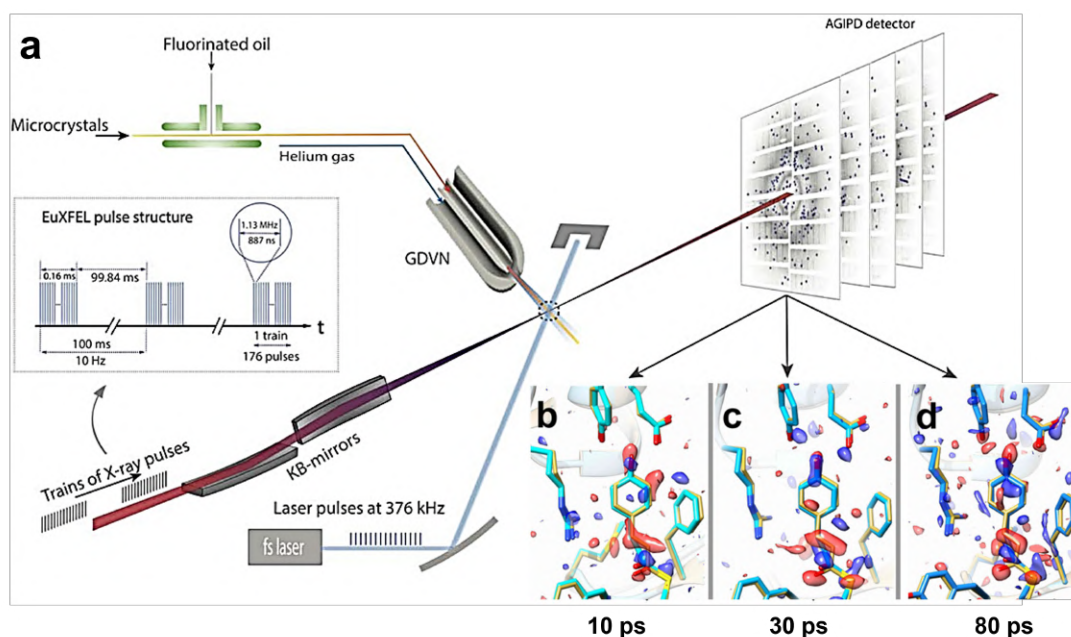


Figure 4.5: The picture is taken from [70] to demonstrate the idea of a pump-probe experiment at the European X-ray free electron laser (XFEL). (a) Setup of a MHz TR-SFX experiment at the European X-ray free electron laser (EuXFEL). (b-d) Structures and difference electron density at  $\pm 3\sigma$  counter level in the chromophore binding region of the photo-active yellow protein (photo-active yellow protein (PYP)) from 10 ps to 80 ps time delay.

Based on the reaction initiation method, we can identify two main categories in TR-SFX: pump-probe [71] and mix and inject TR-SFX [65]. The difference between them is that the first method uses light excitation to trigger the biochemical reaction of a bio-macromolecule. In contrast, the second one kicks out the reaction by a chemical factor. The biochemical reaction is initiated in the crystal and followed. However, to be possible, the crystal must not be disordered during the reaction. Because time is of the essence, data collection must be as fast



as possible. The advent of high-brilliance synchrotrons makes it possible to collect meaningful data in very short exposure times.

Time-resolved studies are also possible with Laue diffraction (discussed in Section 4.1.1). In this case, the structure can be determined directly from the diffraction pattern. However, the initial and final conditions can be determined much easier from static structures solved with monochromatic resolution. The time-dependent average structure in the crystal changes during the course of the reaction due to the time-dependent change in the concentrations of the intermediate states of the molecule. Thus, the diffraction varies with time. The data can then be analysed to identify the individual transient structures in the crystal. More details can be found in [72].

The use of small crystals in serial pump-probe crystallography experiments [73–75] provides advantages such as an increase in the volume fraction of optical laser-activated crystals compared to traditional time-resolved Laue experiments. Additionally, it offers larger diffusion volumes in mix-and-inject experiments, enabling the investigation of ligand-triggered biological reactions commencing at sub-millisecond times. Typically, such an experiment starts by initiating the biochemical reaction in the crystal with a burst of laser light (*pump*). The data (*probe*) would be collected after a well-defined time-lapse (pump-probe delay). To obtain a complete dataset, the pump-probe sequence is repeated. As a result, the user can determine the macromolecular structure-forming changes characteristic of the time delay after the start of the reaction. Generally, during the experiment, different pump-probe delays are tested; thus, each determined structure at each time delay is a snapshot of a molecular movie describing the path of conformational changes during the reaction.

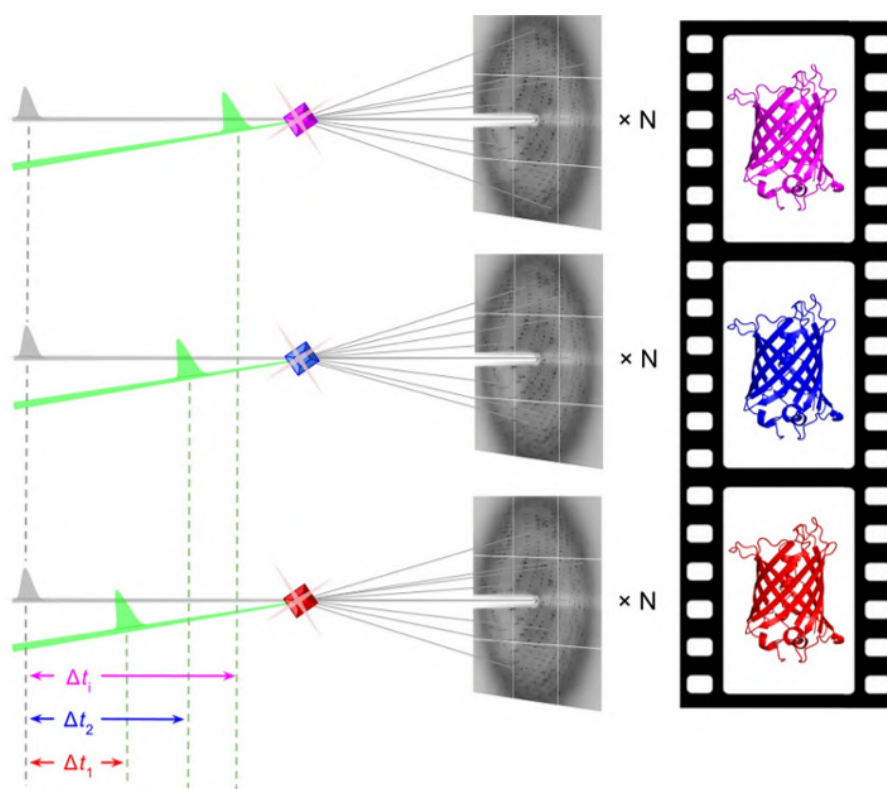


Figure 4.6: It is an original figure taken from [76] to demonstrated the idea of time-resolved experiment

For a pump-probe experiment to be successful, the reaction must be triggered efficiently and synchronously. Synchronously means that the pump has to initiate the reaction faster than the process of interest. It should be understood efficiently that a large fraction of the crystalline molecules has to be activated by the pump to observe structural changes by ensemble-averaged crystallography. We can use various pump methods depending on how

the reaction could be triggered: light-sensitive proteins or caged compounds in the complex with light-insensitive proteins can be activated by ultra-violet (UV)-visible light; in the case of enzyme catalysis, the start occurs by diffusion of substrates; a temperature jump could affect conformational equilibrium [77]; redox processes [78] or the destruction of strained intermediate states [79] could be triggered by X-ray irradiation; to study protein mechanics we can apply electric field pulses [80]. However, the widespread method is optical triggering due to robust technical implementation, the availability of crystals for many light-sensitive proteins with cyclic reactions to be studied, and the accessibility of ultra-fast time scales.

#### 4.1.6 Reflection partiality problem

The reflections of a perfect crystal have finite dimensions in the reciprocal space defined by the shape of the crystal. However, in real experiments, we are dealing with imperfect arrangements of atoms within a protein crystal. Protein crystals are usually represented as mosaic blocks, each perfectly ordered within itself and separated by lattice defects. The mosaicity of the crystal is a measurement of the misalignment of such individual domains. Since not all of the mosaic domains will satisfy the Bragg condition simultaneously, the crystal mosaicity leads to the deformation of the shape and the increase in the size of the reciprocal lattice peaks.

As mentioned in Section 2.9, in the experiment conducted using monochromatic radiation and a crystal in a certain orientation, diffraction occurs when the reciprocal lattice points intersect the Ewald sphere. Nevertheless, even with a monochromator, X-ray radiation has a finite bandwidth and is not completely collimated. Thus, the Ewald sphere has a finite thickness dependent on the scattering angle. This fact, in combination with the fact that reciprocal nodes have certain dimensions, results in recording only diffraction from a cross-section between reflections and the Ewald sphere (Fig. 4.7). The reflection intensity could be completely recorded only when the reciprocal peak lies fully within the Ewald sphere.

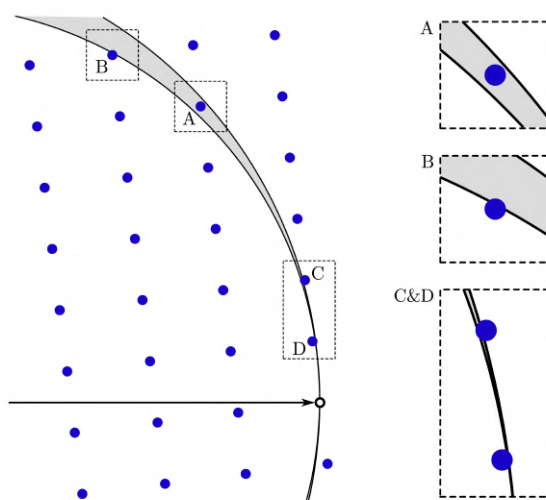


Figure 4.7: It is an original plot from [47] to show the partiality problem in crystallography: the intensity of reflection A is integrated fully while the intensities of reflections B, C and D are only partially integrated.

Partiality significantly affects the final quality of the data, specifically in SX, where full diffraction intensities are obtained solely from partially recorded reflections. To resolve a crystal structure, we have to precisely calculate partiality or obtain the diffracted intensities from the entire volume of the reciprocal lattice nodes. We should know accurately about beam bandwidth, divergence, unit cell parameters, the crystal's mosaicity, and its

orientation to estimate reflection partialities. There are several approaches to evaluate partialities [81–86]. One of the simplest models, used in the nXDS [81] and cctbx.prime [84], assumes that X-ray pulses are monochromatic and that the distance between the Ewald sphere and the reciprocal space node estimates the partiality of a reflection. Thus, the further away the reflection from the Ewald sphere is, the larger the scale factor will be applied to compensate for the partiality. Nevertheless, they depend on a vast amount of fully integrated reflection intensities. Therefore, crystallographic experimental methods are trying to measure reflections fully, for example, by increasing the X-ray spectral bandwidth [87].

#### 4.1.7 Crystals

Crystal formation from the macromolecules is the initial step in crystallography. Under certain circumstances, many molecular substances, including proteins, solidify to form crystals. From a solution to a crystalline state, individual molecules of a substance adopt one or a few identical orientations. The resulting crystal is an ordered three-dimensional array of molecules held together by non-covalent interactions.

Macromolecular crystallisation is an entropy-driven process. The local increase in order, achieved by organising the macromolecule on the crystal lattice, is balanced by the increased freedom of other particles in the solution. The crystallisation process depends on the concentration of various solutes and specific physicochemical parameters, such as pH, affecting macromolecules' surface charge. Crystallization tests are aimed at identifying favourable conditions for crystal growth. The main parameters varied during these tests are ionic conditions, sample concentration, temperature, pH and the concentration of so-called precipitants, which include certain salts (e.g. ammonium sulphate), polymers (e.g. polyethylene glycol, polyethylene glycol (PEG)) and organic solvents (e.g. ethanol). We could influence the crystallisation process by changing the volume of the sample (which affects the kinetics) or by adding small concentrations of additives such as certain metal ions, detergents, and urea. After crystallising a protein, obtaining crystals of chemically modified or ligand-bound forms is usually necessary. Several methods exist for growing protein crystals by placing the protein in an appropriate solution for nucleation and crystal growth, which are excessively discussed in [88, 89].

In vapour diffusion, a drop containing the protein is equilibrated against a larger reservoir of the mother liquor, and its size usually decreases during equilibration, increasing the constituents' concentrations. Due to the changing conditions inside the drop, the protein either crystallises or precipitates out [90, 91].

Crystals can also be obtained by dialysis of a protein solution against a crystallization solution [92]. With this approach, the protein concentration is kept approximately constant.

When protein crystals are too small for X-ray crystallography, seeding makes it possible to increase the crystal size. This involves adding a crystal to a new drop containing the protein. The crystal then acts as a nucleus from which a larger crystal can grow. Crystals of one protein also help initiate the growth of crystals of another, but similar protein [93].

By 2020, according to database <https://blanco.biomol.uci.edu/mpstruc/#Latest>, about 80% of membrane protein structures have been solved by X-ray crystallography. Such proteins are insoluble in conventional biochemistry buffers, which challenges the purification and crystallisation of these proteins. The usual approach is to extract the integral membrane protein in solubilised form with proper detergents and proceed with crystallization trials as with soluble proteins. Another approach has been to rely on lipid cubic phases to crystallise membrane proteins. Membrane proteins are thought to diffuse into regions of lower curvature, where they are incorporated into a lamellar organisation that binds to form highly ordered three-dimensional crystals. With more experience, more guides regarding crystallisation techniques for membrane proteins appeared to



overcome the mentioned problems and help users obtain a final 3D structure of membrane protein [94, 95].

#### 4.1.8 Sample delivery systems

Various sample delivery systems have been employed or investigated for delivering micro- to nano-scale crystalline samples into the X-ray beam for data acquisition, for example, injector methods with liquid or extrusion jets, fixed-target methods, and hybrid methods, discussed in the literature [50]. The principle schemes of developed sample delivery systems can be found in Fig. 4.8 and Fig. 4.9. The choice of the most suitable sample delivery method depends on the experimental goals, the required environment, and the characteristics of the crystals (such as size or quantity). If more than one sample delivery method can be used in an experiment, their advantages and disadvantages should be considered for each data acquisition step, availability at the beamline, and impact on the next data processing steps.

We can distinguish two methods to deliver micro- to nano-scale crystalline samples into the X-ray beam. One class of methods obtains a fine stream of crystals by ejecting a suspension of crystals through a small nozzle that flows orthogonal to the beam direction. At a high repetition rate, an X-ray beam interrogates the crystal stream and a diffraction pattern is produced each time an X-ray pulse hits a crystal. Variations include slowly flowing extrusions of crystals in a viscous medium or transporting the suspension as drops applied to a moving tape, as mentioned below:

1. Injectors with liquid jet (or high-flow injector technique like dynamic virtual nozzles, dynamic virtual nozzles (GDVN) [53], double-flow focusing nozzle, double-flow focusing nozzle (DFFN) [96]): the main advantage of such injectors is relatively low background, can be used in air and a vacuum, is capable of delivering the samples at a very high rate (up to MHz) and can be used for different time-resolved measurements (light-activated or mixing). The main disadvantage of this method is high sample consumption due to the high injection speed;
2. Injectors with extrusion jet (or low-flow injector technique as Lipidic cubic phase injector (lipidic cubic phase (LCP)-jet) [58, 97, 98]): in comparison with using liquid jet such sample-delivery method is slower and gives higher sample consumption efficiency, but the jet is thicker, so it produces higher background;
3. Transporting the suspension as drops (a droplet-based injection method, a droplet-based injection (ADE) [97, 99–105]) or thin stream applied to a moving tape [106–108]: low sample consumption and it is compatible with mix-and-diffuse methods for substrate or drug-design studies.
4. A hybrid electrokinetic technique called the microfluidic electrokinetic sample holder (the microfluidic electrokinetic sample holder (MESH)) method is a low-flow method [109, 110]. A crystal suspension flows through the capillary towards the X-ray beam under high voltage. A thin liquid stream is formed due to the deformation of the polarisable exited from the capillary solution surface by the electric potential and accelerates through the X-ray beam position towards a target electrode. This method is compatible with both liquid and moderately viscous carrier media, and larger capillary sizes may be used to alleviate clogging issues. Nevertheless, an electric potential may influence the observed protein structure. Additionally, since an open capillary is used to electrospin the crystal suspension in the X-ray beam path, the volume of liquid surrounding the crystals is larger than a thin liquid jet, creating a higher background; thus, it creates problems for further data analysis.

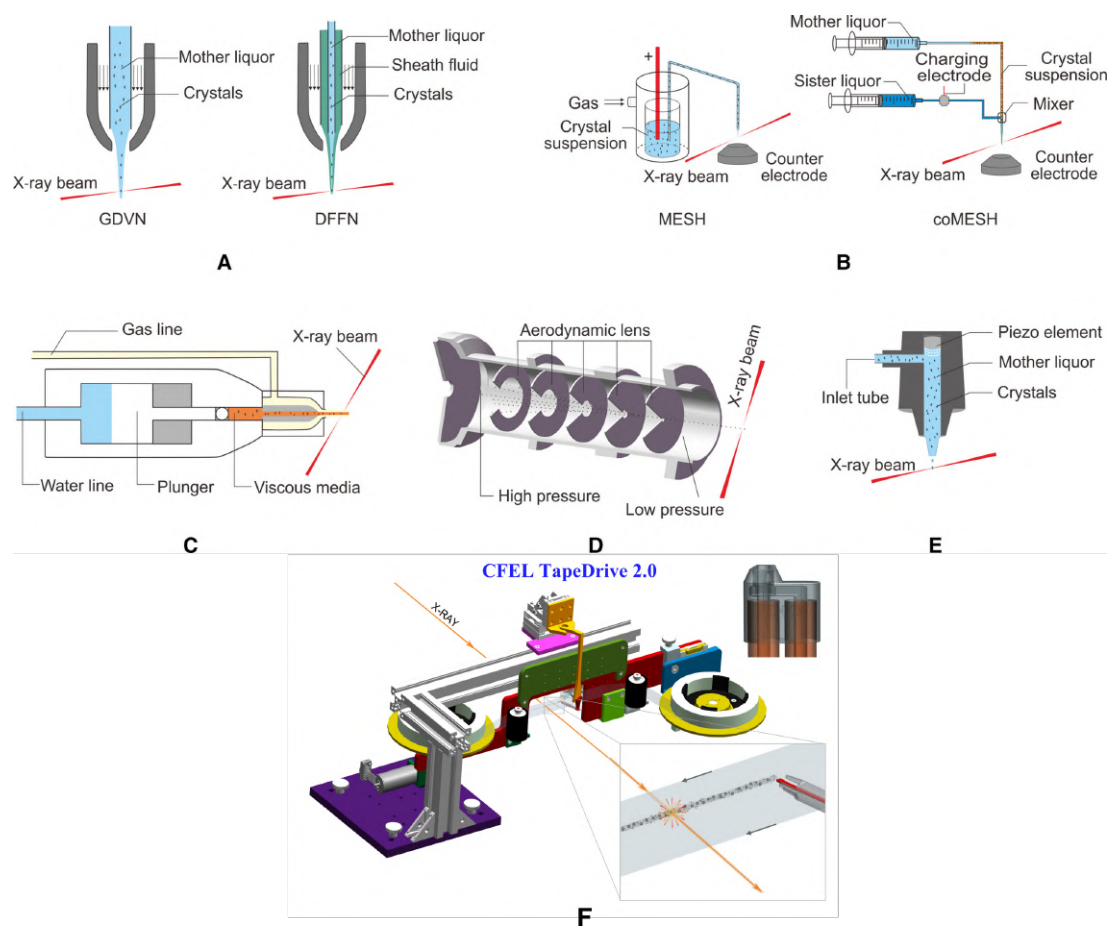


Figure 4.8: The original plot is taken from [20], which illustrates diagrams of various moving target systems. (A) The GDVN and DFFN system. (B) MESH and CoMESH system. (C) High-viscosity extrusion (high-viscosity extrusion (HVE)) injector. (D) The aerodynamic lens. (E) The nozzle of acoustic droplet ejection (ADE) technology. (F) CFEL TapeDrive 2.0: a conveyor belt-based sample-delivery system for multi-dimensional serial crystallography [108]. The picture of TapeDrive was made by Alessandra Henkel.

A conceptually different approach is to deposit crystals onto a solid supporting membrane, which is then raster scanned in the X-ray beam (Fig. 4.9). These "fixed-target" sample delivery systems encompass patterned silicon chips [43, 111–116] or plastic membranes [117–119]. This approach also facilitates on-the-chip crystallization [21], making it more suitable for fragile crystals that might otherwise suffer damage from injection-based methods. Some beamlines incorporate a robot to mount holders on the goniometer, eliminating the need to access the experimental hutch during the experiment. Data can be collected at room temperature with controlled humidity or cryogenic temperatures.

Recent years have witnessed the development of several distinct fixed-target designs for serial crystallography. A thin membrane chip, made of silicon or silicon nitride, is used in these designs and features periodic microscopic wells or pores [111–115, 120]. Appropriately sized pores trap the crystals when excess mother liquor is removed via blotting [115], resulting in minimal background scattering in the diffraction pattern [43, 114, 116]. The chip is commonly sealed between two membranes or maintained in a humid environment, such as a humidified gas stream, to prevent crystal dehydration. The former approach is suitable for vacuum measurements, while the latter offers the advantage of a lower background by avoiding introducing additional material into the X-ray beam.

An advantage of the fixed-target rastering approach is that it provides the possibility to measure every individual crystal, utilise the protein and reduce sample consumption efficiently. This is beneficial for protein samples that are expensive to produce. Furthermore, if the positions of the crystals are known before data collection, it is possible to measure only at those positions and efficiently use the X-ray beam. Also, for each crystal, small rotation series can be measured, which proves to be beneficial in mitigating the partiality problem [121]. Crystals can be placed into well-defined locations determined by the pore structure of the chip or found by inspection before the X-ray measurements, no matter where they are located [113]. The primary challenge associated with loading crystals into pores or wells is that they, particularly larger crystals, may not always be confined to those specific positions and can be distributed anywhere on the chip. Additionally, smaller crystals that are smaller than the pore size might be lost during the blotting process. The measurement of crystal locations can be achieved in different ways. For example, via UV tryptophan fluorescence imaging [122–124], UV-vis spectroscopy [113, 125], second-order nonlinear imaging of chiral crystals (second-order nonlinear imaging of chiral crystals (SONICC)) [126], or even manually selecting crystals using an in-line visible microscope [114, 127]. In the case of SONICC microscopy, it was found that micro-crystal positions could be determined with a spatial resolution of approximately  $2\ \mu\text{m}$  with fast image acquisition times in correlation with the crystal locations identified by raster scanning using an X-ray beam [126]. Using UV-vis spectroscopy to predetermine crystal locations in fixed-target room-temperature crystallography [113], an exceptional performance was demonstrated with a raw hit rate of nearly 100% and an effective indexing rate of approximately 50%. Unfortunately, all these methods have their limitations: SONICC has not been integrated at beamlines, and optical auto-search of crystal positions may fail since the crystals have very different shapes and sizes. At the same time, manual centering requires a lot of concentration and user intervention.

The hybrid sample delivery approach for SFX as a combined inject-and-transfer system (a combined inject-and-transfer system (BITS)) was introduced in [128] combined with a hybrid injection and fixed-target scanning method. Suspensions of crystals or crystals immersed in a viscous medium are applied to a UV/ozone (UVR) treated polyimide film with an injection needle and transferred to the point of interaction with X-rays in a programmed translation stage. In BITS, a crystal sample deposited on a film is scanned by translation in vertical and horizontal directions, depending on the desired length of the scanning interval. Thus, the sample consumption is sharply reduced. In [118], the development of universal, inexpensive, and robust polymeric microfluidic chips for routine and reliable serial measurements at room temperature, both in synchrotrons and in XFELs, was presented. The chip design included high X-ray translucency thin film layers tuned to minimise scatter background, adaptable sample flow layers tuned to die size, and a large sample area compatible with raster scanning and rotation-based serial acquisition data collection. The chip could be used for *in situ* crystallisation using micro-batch and vapour diffusion methods.

It is also possible to deliver crystals directly from mother liquor solutions at ambient temperature and pressure using the extractor crystal delivery method, successfully tested on the XFEL and synchrotron [130]. This method is easy to install and operate and compatible with various crystal sizes. A thin film of liquid on a crystalline carrier containing a random distribution of crystals is obtained by removing the mesh or thin film from the mother liquor solution using a solenoid driver. Then, the substrate is placed in a new support area in the X-ray beam between exposures. After data collection, the procedure is repeated from loading to exposing a fresh batch of crystals until a complete data set or a significant drop in the hit rate is obtained. In this way, unexposed crystals are returned to a solution with the possibility of exposure during the next cycle, resulting in a drastic reduction in sample waste. To solve the dehydration problem, the setup can be covered with a small plastic tube with an X-ray injection hole, and on the opposite side is a thin radiolucent film that allows diffracted X-rays to

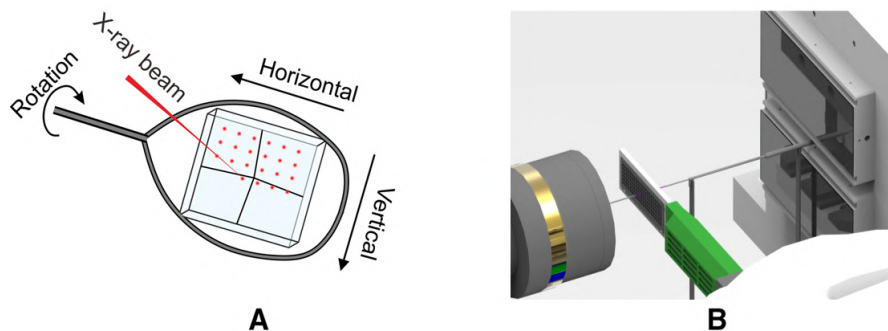


Figure 4.9: (A) Helical data collection of nylon loop [129]. This is an original plot taken from [20]. (B) Roadrunner II chip. This is an original plot taken from [47].

pass through.

Another recently developed approach for sample delivery is laser ablation of crystal-containing solutions [131], which is based on ultra-fast evaporation of liquids [132]. The idea is based on the excitation of water with a laser beam in the mid-infrared at the resonance of the  $H_2O$  stretch vibration to create sufficient force due to the evaporation to ablate the suspension with the crystals into a plume that may be exposed to X-rays [131].

## 4.2 Data analysis of conventional crystallography

In Section 4.1.2, we have already discussed the main principle of conventional crystallography. In such an experiment, by rotating the crystal at a fixed angular velocity in the beam, we could measure the structure factors from all relevant reciprocal lattice nodes to reconstruct the protein structure. Since a rotation in real space corresponds to a rotation in reciprocal space, diffraction occurs every time the points of the reciprocal lattice intersect the Ewald sphere. The total collected data from a crystal rotation is called a rotation series.

Now, we will talk about the main steps required for data processing. First, we must apply all corrections to the collected diffraction patterns because of the known detector and experiment artefacts. Generally, modern detectors have complex geometry. Nowadays, they contain a finite number of panels with gaps in between. Data are saved for each panel separately during data collection, but for further processing, we should work with the assembled images according to the detector geometry [133]. Also, we have to mask shadows of installations, bad regions, and 'misbehaving pixels' of the detector (like hot pixels) because they could cause problems at later stages of data processing. As mentioned in Section 4.1.8, different sample delivery methods affect the diffraction pattern by causing noise, which must also be corrected or masked. Some detectors require regular and proper re-calibration, such as AGIPD [134]. It is necessary to apply proper calibration constants to the raw data, such as the generated average noise signal and the so-called dark offset when the beam is off. These constants could significantly shift due to temperature changes and other effects; thus, it is necessary to measure them periodically. There are more corrections that constants may require, such as gains and optional switching thresholds, for which another calibration pipeline is needed [135].

The next step includes the identification of Bragg spots in the diffraction patterns by specialised algorithms named peak finders. The main idea of such algorithms is to assess the proper background and then determine peaks significantly above the noise level. Further, we have an indexing stage in data processing, which is the assignment of Miller indices to all locations where Bragg spots are supposed to be measured. To do this, we must identify the crystal orientation [136] for each diffraction pattern. Based on this determined orientation

matrix, the diffraction pattern is simulated to identify possible Bragg spots, and afterwards, the integration of intensities from the areas around predicted Bragg positions with the following scaling is performed.

Despite the ease of conducting such an experiment and the availability of well-developed data processing software (including graphical user interface (GUI)) [137–139], we face some limitations in traditional crystallography. One of the limitations is the use of large crystals for high-resolution measurements because of the tolerable dose. However, not all proteins, in general, could be crystallised (or hardly crystallised, such as membrane proteins) or obtained in the required size. The crystal should have minimum defects to carry out a successful rotational experiment.

Time-resolved studies with conventional crystallography do not fully meet the desired requirements. For example, a small crystal size is needed to initiate a chemical reaction. Small crystal size will enable an ensemble of molecules in the same state, but they will not survive while we measure them. Moreover, a time resolution that is longer than the time to rotate the crystal. Thus, conventional crystallography cannot employ time-resolved studies.

## 4.3 Data analysis in serial crystallography

The analysis of serial crystallographic data differs drastically from conventional crystallographic data and requires specifically developed software. First, due to the popularity of serial crystallography techniques for exploiting all capabilities of XFEL facilities [9] and modern synchrotron sources [10], serial crystallography results in the collection of large data sets in crystallography, requiring automated data processing pipelines and large-scale computing environments. Second, during data acquisition in serial crystallography, we collect a set of still diffraction patterns of randomly orientated crystals, for which conventional software cannot be applied in such cases for estimations of the structure factor moduli. We can solve the crystal structure using the standard crystallographic programs once the structure factor moduli have been determined. Here, we overview the main data processing steps in obtaining a final set of reflection intensities from a raw diffraction image dataset.

### 4.3.1 Pre-processing and hit-finding

Serial crystallography uses different sample delivery techniques based on a liquid jet or moving fixed-target support (more details can be found in Section 4.1.8) to put crystals into an X-ray beam. Moreover, serial crystallography (SX) experiments require the collection of an enormous amount of diffraction snapshots to get 3D structural information of the studied protein because carrying on such experiments means that there will be a probability of not hitting the crystal or partially hitting it, thus, at the end the full dataset will contain patterns with non-hits. Here we come across such characteristics as the hit rate, the fraction of measured frames with crystal diffraction compared to the total number of measured patterns. A typical hit rate value for such experiments is rather low - on the order of 0.1-10%. Nevertheless, using the fixed-target sample delivery can give a hit rate of up to 100%. Still, one possible issue is having multiple hits on diffraction patterns due to overlapping diffraction from several crystals, which affects data processing. But still, most experiments generate a large amount of data that contain a small part of a useful number of patterns for further data processing. Thus, the initial necessary step is to reduce the amount of data by saving only images with diffraction to save data storage and speed up the subsequent processing.

Cheetah is one of the world's most popular data processing and diffraction pre-processing software for SX experiments at FELs [133]. The main functions of Cheetah are detector corrections, searching for Bragg peaks,

sorting crystal diffraction patterns, and converting them into a facility-independent format for further analysis. Detector corrections are performed by masking misbehaving pixels (like bad and saturated pixels), applying dark corrections of each module and gaining corrections for each pixel. The next step is identifying Bragg peaks in the pattern using the so-called peakfinder8 algorithm. The main idea of peakfinder8 is to find all regions with a certain number  $N$  of connected pixels ( $n_{min} \leq N \leq n_{max}$ ) with values above a radially dependent threshold, determined from the radially averaged background intensity. The pattern is a hit if the number of found with a sufficiently high signal-to-noise ratio exceeds a certain minimum value  $n_{peaks}$ . Then, all 'useful' patterns are saved in HDF5 file format with such information as positions and intensities of the found peaks and various instrument and experiment setups.

### 4.3.2 Indexing

After performing all the steps mentioned in Section 4.3.1, the following analysis includes the determination of crystal orientation, integration of reflection intensities in each image, and then merging integrated intensities into a final data set. Several packages can be used to do these steps: cctbx.xfel [83], nXDS [81], DIALS [140] and CrystFEL [141]. However, compared to other programs, CrystFEL is more popular among the community due to being a free open-source software specifically developed for serial crystallography data processing, and now even has a user-friendly interface. Thus, the alternative packages will not be further mentioned.

After the peak finding, it is necessary to assign Miller indices to the found Bragg reflections or, in other words, to determine the crystal orientation from a diffraction pattern. This step is also known as indexing and can be performed within CrystFEL using a program such as `indexamajig`. `indexamajig` tries to project all Bragg peaks that were found during a pre-processing stage or within itself by using any available hit-finding algorithm (like `zaef` [142], `peakfinder8` [133] or `peakfinder9`), onto the Ewald sphere and determines reciprocal space coordinates of their corresponding reflections. Various indexing algorithms are applied to the positions of the peaks on the detector or to the computed reciprocal space coordinates to determine a three-dimensional periodic lattice that matches the observed reflections. There are several indexing methods implemented within `CrystFEL:MOSFLM` [143], `DirAx` [144], `XDS` [137], `asdf`, `TakeTwo` [145], `Felix` [146], `XGANDALF` [147] and `PinkIndexer` [45]. `asdf` is the indexing algorithm implemented internally in CrystFEL. `TakeTwo` [145], `Felix` [146] were developed to deal with the indexing of multiple crystals in a diffraction pattern, while `XGANDALF` [147] is an advanced indexing approach beneficial for indexing weak patterns with a few numbers of detected peaks and `PinkIndexer` [45] enables automatic indexing for emerging techniques such as serial electron crystallography [148, 149] and serial crystallography of pink beam [43].

When the indexing solution is obtained, the determined lattice parameters are then compared to the parameters of the expected unit cell. In cases when there is no prior information about unit cell parameters, it is still possible to index such datasets by the following indexers to get initial estimations: `asdf`, `DirAx`, `MOSFLM`, `XGANDALF`.

After that, the found orientation matrix is used to predict the positions of the Bragg peaks on the detector. The positions of found peaks are compared with the predicted ones, and indexing is considered successful if a certain fraction of the found Bragg spots coincides with the predicted ones. One of the metrics to evaluate the success of this procedure is the indexing fraction, which reflects the percentage of indexed hits.

Afterwards, the prediction refinement process takes place to refine the crystal lattice and the centre of the detector by minimising the distances between the found and predicted spot positions, considering an "Ewald offset" term in addition to the spot position terms [141]. Then, the reflection profile is calculated so that 98% of the spots that were assigned indices are predicted. Similarly, the individual diffraction resolution of each crystal

is estimated at the level of the 98<sup>th</sup> percentile of the scattering angles of the predicted peaks. Using this value, an individual resolution cut-off can be applied to each crystal during the integration and merging of intensities.

Successful indexing relies on having an accurate description of the detector geometry. Provided that the original geometry is accurate enough to index at least a few patterns, it can be refined by comparing the observed and calculated peak positions on the detector. Because indexing solutions based on information on peak positions from the entire detector, the predicted spot locations can be used as a reference and in the case of a mispositioned panel, we will have a systematic offset between the observed and calculated peak locations [150]. With `geoptimiser` [26], we could apply corrections to each panel and repeat the indexing procedure until we get the detector geometry with high precision.

These are some crystal symmetry classes when the crystal in several different orientations produces diffraction patterns with Bragg peaks in identical positions but with different intensities. In these cases, the indexing solution is ambiguous between two or more possibilities, while only one is correct. The *ambigator* program in `CrystFEL` includes a simplified version of the Brehm-Diedrichs algorithm [151], which can resolve such ambiguities using a clustering approach by calculating the correlations between the integrated reflection intensities.

### 4.3.3 Integration and merging of intensities

The result of the `indexamajig` program is a list of predicted reflections with their integrated intensities for each indexed crystal. The so-called 'three rings' integration approach is used to obtain the reflection intensity: three concentric rings located at the centre of the predicted reflection position define the peak, buffer, and background estimation regions. To obtain the integrated intensity, it is necessary to sum the pixel values inside the smaller circle and then subtract the background estimated from the ring between the middle and outer circles. Here, the importance of accurate peak prediction becomes apparent: the better the predicted reflection position matches the actual position of the Bragg peak, the smaller the radius of the inner ring can be used to integrate it, improving the signal-to-noise ratio.

After integration, integrated reflections are merged into a final set of *hkl* intensities, which are used further for the structure determination. The Monte Carlo approach is the simplest way to merge intensities: it averages the integrated intensities of each symmetrically unique reflection from different crystals. It was implemented in `CrystFEL` within the `process_hkl` program. Due to having many merged diffraction patterns to sample all possible crystal orientations, this approach is the same as angular integration [152, 153], performed in conventional crystallography by rotating the crystal during X-ray exposure.

`partialator` program implemented in `CrystFEL` includes other more advanced merging methods. Before the final merging of intensities, `partialator` performs the following procedures: scaling, partiality correction, and post-refinement [82]. The linear and Debye-Waller coefficients are two scaling factors that are determined for each crystal to bring the individual measurements of reflection intensities into as close an agreement as possible. The linear term accounts for changes in crystal size and beam intensity, while B-factor scaling compensates for variations in crystal quality. Partiality is calculated based on the geometric model described in [82], which evaluates partiality as the fraction of the reciprocal volume of the lattice node excited by X-rays. Post-refinement means an iterative refinement of the scaling and geometrical parameters for each crystal to achieve the best possible agreement of the corrected reflection intensities. After applying calculated scaling factors and partiality estimates to all crystals, the final intensity of each symmetrically unique reflection *hkl* is again determined as the average of the corrected intensities of all measurements. The errors in the merged intensities are estimated as a standard error of the mean.

$$\sigma_{hkl} = \frac{\sqrt{\sum (I_{hkl} - \langle I_{hkl} \rangle)^2}}{n_{hkl}} \quad (4.1)$$

where  $I_{hkl}$  - an individual measurement of reflection  $hkl$ ,  $\langle I_{hkl} \rangle$  is the final merged  $hkl$  intensity and  $n_{hkl}$  - is the total number of measurements of  $hkl$  reflection [141].

## 4.4 Initial phase estimate

### 4.4.1 Molecular replacement

If the protein's related or homologous (greater than 50% sequence identity) structure is already known, the molecular replacement (molecular replacement (MR)) method can be used. The main principle of MR is based on providing the highest correlation between the experimental diffraction measurements and the results calculated from the model by finding proper rotation and translation, which position the model structure in the unit cell. The basic ideas were described by Michael G. Rossmann and David M. Blow in 1962 [154, 155].

### 4.4.2 Direct methods

Molecular replacement is one of the powerful procedures for phasing observed structure factor amplitudes. However, it still has several limitations. Firstly, MR suffers from phase bias; the resultant structure may resemble the search model even though it is the wrong solution. Secondly, it cannot be used to determine the structure of a protein for which no suitable homologous structure is known to serve as a search model. We could use the *ab initio* phasing method in such situations. It is an essential method when the first protein structure is determined.

For very small molecules, it is possible to interpret the phases directly from the Patterson map. If the individual Patterson peaks are fully resolved, inter-atomic vectors can be accurately determined and sufficient to construct a model structure.

Some phase information is contained in the reflection intensities. A diffracting crystal is a real object consisting of atoms whose electron density is positive everywhere in the unit cell. Such assumptions as atomicity and positivity constrain the number of possible phases and, more importantly, create phase relationships between different reflections. For example, the phase of the reflection should correspond to the maxima of the real-space waves with high electron density positions overlapping in the crystal. Therefore, the phases of three waves that intersect at a high electron density location (for example, in an atom) have a phase relationship, ensuring the maxima overlap in a crystal. If we have indices satisfied such condition  $h + k + l = 0$ , we will have the following triplet relationship between their respective phases  $\phi$ :  $\phi_h - \phi_k - \phi_{h-k} \cong 0[2\pi]$ , where  $[2\pi]$  is the modulus.

Direct methods are based on statistical relationships between sets of structure factors. However, few protein structures have been solved *ab initio* in this way because direct methods require very good data quality ( $<1 \text{ \AA}$ ). The statistical relationships become weaker as the number of atoms increases. Therefore, this methodology is often used in conjunction with other phasing methods. With a small number of initial phases, the full dataset can be obtained iteratively.



### 4.4.3 Single and multiple isomorphous replacements (SIR, MIR)

The most common method for determining *ab initio* phases for proteins is to introduce heavy atoms into the protein structure [156–159]. Heavy atom derivatives are obtained by soaking native protein crystals in a buffer containing a heavy atom compound or by co-crystallization. The native and heavy-atom derivative crystals should be isomorphous, which means undistorted (crystal form and unit cell dimensions are unchanged). In this case, the structural factors and phases of the protein component of the crystal do not change. However, the newly observed intensities consider the presence of additional atoms. The influence of heavy atoms can then be separated from the rest of the molecule and with direct methods. Then, the position of the heavy atoms (i.e. the heavy atom substructure) can be determined. The structure factors of heavy-atom derivative ( $F_{ph}$ ) and native protein ( $F_p$ ) and heavy atom ( $F_h$ ) have the following relationship:

$$F_p = F_{ph} - F_h \quad (4.2)$$

Based on the Patterson function:

$$P(uvw) = \int_x \int_y \int_z \rho(xyz)\rho(x+u, y+v, z+w) dx dy dz \quad (4.3)$$

A peak at position  $\mathbf{r}$  in the Patterson map is proportional to the sum of the product of the electron densities of atoms separated by the vector  $\mathbf{r}$ . Since heavy atoms have greater density than the atoms usually presented in proteins, peaks at interatomic vectors between heavy atoms appear clearly in the map obtained by subtracting the Patterson diagram of the native protein dataset from that of the protein with heavy atoms.

Generally, there is a small number of heavy atoms per asymmetric unit. Thus, it is possible to estimate their coordinates from standard methods used in small-molecule crystallography and, therefore, phase the structure  $F_h$  corresponding to the heavy atoms alone. To choose the correct phase angle, it is preferable to use data from two or more different heavy atom crystals to produce a single solution for the phase of the protein crystal.

### 4.4.4 Single and multiple anomalous dispersion (SAD, MAD)

X-ray diffraction patterns of biological macromolecule crystals provide sufficient information to determine atomic structures, but atomic positions are complicated without having electron-density images. While diffraction measurements offer amplitudes, computing electron density necessitates obtaining phases for the diffracted waves. Anomalous scattering, a resonance phenomenon, provides a robust solution to this phase problem. By utilising scattering resonances from various elements, the techniques of multi-wavelength anomalous diffraction (multiple anomalous dispersion (MAD)) and single-wavelength anomalous diffraction (SAD) are now widely used to determine biological structures at the atomic level, mentioned in this Subsection.

All atoms have the property that their electrons can be excited from a lower to higher energy by incident photons with a given energy. A secondary photon is emitted as the electron returns to its shell. An anomalous signal refers to a phase change in scattering due to the absorption contribution, which results in a complex scattering amplitude  $f' + if''$  [160, 161]. It depends on the wavelength ( $\lambda$ ) of the incident X-ray and does not diminish with the diffraction angle ( $\theta$ ), because, in practice, the electron acts as a point scatterer.

In the case of normal scattering, Friedel's law states that the structure factors for  $(hkl)$  and  $(-h - k - l)$  (Bijvoet pairs) have equivalent amplitude but opposite phases. An anomalous signal would add extra vectors to the vector and result in the breaking symmetry [162].

An anomalous Patterson map is calculated from the intensity differences between Bijvoet pairs. The anomalous scattering is very weak, usually equivalent to the strength of a few electrons. Highly redundant

data enables the precise determination of each reflection's structure factor amplitude, reducing the noise of the Patterson map to a minimum. Vectors between anomalous scatters are nevertheless present and, when identified, provide phasing information.

The first choice is to choose the wavelength that produces the largest anomalous signal (i.e. where  $|f''|$  is largest). The wavelength is then varied to collect at maximal  $f'$ . The third and fourth data sets are usually collected at remote points of the absorption spectrum. As in the case of multiple isomorphous replacement (MIR), having as many anomalous derivatives as possible, each with a different anomalous scatterer is better.

The anomalous signal produced by heavy atoms soaked into the crystal is often combined with MIR. The anomalous Patterson map is correlated with the difference Patterson maps to position the anomalous scatter in the crystal with less ambiguity. Ligands with several anomalous scatters bound together are hence particularly valuable. Since the distances between each atom in the ligand are known, they can be readily identified in the Patterson maps and consequently help phase the rest of the structure.

## 4.5 Radiation damage

Elastic scattering, which gives rise to diffraction effects, preserves the photon's energy without any energy deposition in the sample. However, in macromolecular crystallography, X-ray energies of approximately 5-15 keV are typically used, resulting in an elastic scattering cross section orders of magnitude smaller than that of inelastic effects, primarily of photon absorption (Fig. 4.10). This means that for every coherently scattered photon contributing to diffraction, atoms absorb numerous photons, each ejecting a photo-electron. The photo-electron will produce a few hundred ionisation events before it thermalises, producing multiple secondary electrons. Ionised atoms experience a fall of electrons from a higher energy level into the vacancy created by the photo-electron, resulting in energy release in either a characteristic fluorescent photon or an outer shell electron ejected from the atom through Auger decay.

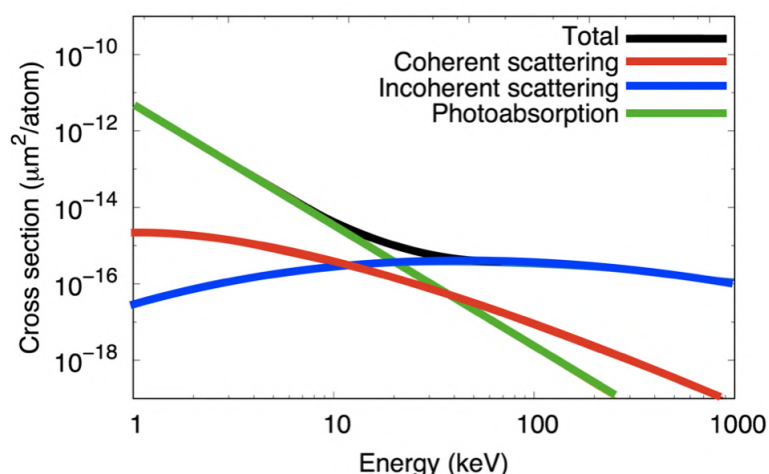


Figure 4.10: The image presented is sourced from [163], illustrating cross sections as a function of energy. It showcases the atomic cross-sections of neutral carbon, specifically focusing on photo absorption, coherent scattering, and incoherent (Compton) scattering. Notably, at 10 keV, photo absorption emerges as the predominant interaction.

Photoelectric absorption is a phenomenon that leads to energy loss, radical formation, deterioration of the crystal lattice, and temperature rise in the sample. This phenomenon is commonly referred to as radiation-induced

damage. The dose, defined as the sample's energy loss per unit mass, is quantified in the SI unit  $Gy = J/kg$ . In macromolecular crystallography, where the crystal thickness  $d$  is usually much smaller than the attenuation depth  $l$  of X-rays, the absorbed dose can be estimated as follows:

$$D = \frac{E_{abs}}{m} = \frac{N_{ph}h\nu(1 - \exp^{-d/l})}{\rho V} \cong \frac{N_{ph}h\nu l}{\rho S} = \frac{I_0}{l\rho}. \quad (4.4)$$

Here,  $\rho$  is the sample density,  $V = Sd$  is the irradiated volume, and  $I_0 = \frac{N_{ph}h\nu}{\rho}$  is the incident beam's fluence or energy per unit area. The dose can be estimated roughly using typical density and attenuation depth values. To calculate the dose accurately for an arbitrary wavelength and crystal content, the commonly used software RADDOSSE can be employed [164].

Blake and Phillips first investigated radiation damage in macromolecular crystallography in 1962 [165]. Radiation damage is broadly classified into two types: global damage and specific damage. Fading of diffraction intensity, particularly at high resolution, with an increase in absorbed dose, is the first sign of global damage. Metrics such as resolution degradation, increase in Wilson B-factor, unit cell dimensions, and mosaicity are often used to identify global damage [6]. Specific structural damage affects particular covalent bonds and is observed in electron density maps. Blake and Phillips [165] predicted such specific damage, as they observed changes in the structure factors of specific reflections with increasing radiation dose, indicating the occurrence of local structural changes in addition to global radiation damage. Their hypothesis has been validated, with the cleavage of disulfide bonds being the most notable example [166].

Henderson estimated a dose limit for macromolecular crystallography of three-dimensional crystals to be 20 MGy from observations of the dose  $D_{1/2}$  required for the biological two-dimensional crystals at 77 K to lose half of their diffraction intensity [167]. The Henderson limit was experimentally measured at 100 K to be 43 MGy [168]. However, the dose limit of 30 MGy, corresponding to 0.7 of the preserved diffraction intensity, is commonly used when planning diffraction experiments. Howells et al. [169] later gave the resolution-dependent dose limit as  $10d$  MGy, where  $d$  is the resolution in Å. At room temperature, protein crystals are much more radiation-sensitive, and  $D_{1/2}$  decreases by about two orders of magnitude when the temperature is increased from 100 K to 300 K [170], giving the dose limit of 300 kGy. This fact led to the development of cryo-temperature crystallography, which has remained the predominant technique for macromolecular structure determination since the early 1990s.

With the advent of X-ray free-electron lasers that produce ultrashort and highly intense X-ray pulses, it enables to outrun most of the radiation-damage processes occurring in the sample during exposure to XFEL radiation [52]. Although the total dose in XFEL experiments significantly exceeds the synchrotron dose limits [9, 171], the time scale of the interaction is so small that it does not negatively affect the quality of Bragg diffraction in high-resolution experiments [12], which was a proof of the concept "diffraction-before-destruction" [52]. However, some experimental and theoretical studies have indicated that structures from XFELs may not always be radiation-damage-free, and it is an ongoing concern for the success of time-resolved XFEL crystallography experiments. To understand this, we need to distinguish global and local damage effects [128, 172]:

#### 1. Global damage:

- The process includes the stochastic photo-ionisation and thermal motion affecting all atoms in the sample, irrespective of their location;

- Can be modelled with modified atomic form factors and a factor compensating for the non-equilibrium heating of the system [171, 173], which depends on pulse duration, pulse energy, and the wavelength [173].

## 2. Local damage

- The process occurs when there are significant differences in the rate of ionisation of different elements or when certain ions exhibit reproducible non-thermal motion;
- Depends on the spatial arrangement of the atoms, on the local molecular environment around each atom [174], it is element specific.

The studies of radiation damage at both synchrotrons and X-ray free-electron lasers are still ongoing [7, 163, 175, 176].

## 4.6 Existing auto-processing data pipelines at different X-ray facilities

According to <https://lightsources.org/lightsources-of-the-world/>, more than 50 light sources are built worldwide. Each facility has established its IT architecture and data policy and has different computational resources. Any beamline focused on conducting drug screening requires complete and reliable automation, from data collection without external intervention to final data processing with obtaining all the necessary figures of merit. The variety of X-ray sources leads to many possibilities for organising a common data management and processing scheme. However, to create a standard data processing pipeline for different types of experiments, such as conventional and serial crystallography, additional considerations are required due to their different nature. Multiple data acquisition software and GUIs, automated processing pipelines, and experimental information management systems based on beamlines have been developed, such as the Information System for Protein Crystallography Beamlines (*ISPyB*) [177, 178], automated data processing (*adp*) [179], Blu-Ice [180], BSS [181], CBASS [182], STARS [183], mxCUBE [184], JBluIce-EPICS [185] and GDA [186]. To have a beamline that can perform standard rotational and serial experiments within a single beamtime requires a good data handling workflow to accommodate different detectors and experimental setups. We will review some of the existing data processing pipelines at various facilities.

Currently, macromolecular crystallography (macromolecular crystallography (MX)) beamlines are designed to facilitate rapid data collection from a large number of samples while providing comprehensive data quality evaluations [187–193]. Robots are widely used for sample change on most MX beamlines. Using optical or diffraction-based techniques facilitates the automatic centring of a crystal in the X-ray beam with micrometre precision [194]. Moreover, collecting the final diffraction dataset from single or multiple crystals with further proper data analysis can be done automatically. The Macromolecular Xtallography Customised Beamline Environment (MXCuBE) is an interface for providing standard control software for a wide variety of MX beamlines [184]. MXCuBE performs data collections in pipeline mode, which means executing the same protocol for hundreds of samples [188], creating specially designed experimental protocols using a queuing function, and integrating experimental procedures using workflows [195]. At this moment, two main versions of MXCuBE are maintained on different beamlines (MXCuBE2 and MXCuBE3). The difference between these versions can be found in this work [196].

The High Throughput Crystallisation Laboratory (HTX Lab) is a large user facility at EMBL Grenoble that develops new methods for macromolecular crystallography, including sample evaluation and quality control

[197, 198]. CrystalDirect technology used in the HTX laboratory provides fully automated crystal mounting and processing [199–201]. At HTX, a web-based laboratory information system, the so-called Crystallographic Information Management System (CRIMS), was developed to provide automated communication with the ISPyB system [177, 178] that supports the management and processing of the collection of X-ray data. CRIMS also automatically extracts information about data collection results and the initial results of data processing carried out by synchrotron data processing systems and provides it to the user. The combination of a high-performance crystallisation platform, CrystalDirect technology, and CRIMS software paved the way for fully automated macromolecular crystallography pipelines that can be remotely controlled [202]. The BioMAX beamline is the first macromolecular crystallography (MX) beamline at a fourth-generation synchrotron source at MAX IV [203]. At this beamline, users can carry out different types of experiments, such as data collection with humidity control, at room temperature or cryo conditions, serial crystallography using fixed-target delivery systems or injectors [204], helical data collection, and rapid feedback mesh scans. Beamline control is provided through the Web-based MXCuBE3 [196] with the ISPyB database [177]. A similar concept of using the ISPyB database and the MXCuBE beamline control graphical user interface (GUI) to automatically process macromolecular crystallography X-ray diffraction data by running the XDS package [137] was implemented on MX beamlines at the European Synchrotron Radiation Facility (European Synchrotron Radiation Facility (ESRF)) [205].

The Berkeley Centre for Structural Biology (BCSB) has seven complementary macromolecular crystallography beamlines operating at the Advanced Light Source (ALS). Beamline control is carried out via the B4 Graphical User Interface on all BCSB beamlines [206]. The B4 GUI supports automatic data collection and on-the-fly data analysis through B4 Autocollect, which enables the characterisation of many samples in a high-throughput manner without user intervention. This is a great advantage for repetitive high-volume applications like drug screening experiments.

Two beamlines, CMCF-ID (08ID-1) and CMCF-BM (08B1-1), have recently been upgraded at the Canadian Macromolecular Crystallography Facility (CMCF) dedicated to macromolecular crystallography [190, 207]. At 08B1-1, the low-level beamline control system is based on the Experimental Physics and Industrial Control System (EPICS) [208]. MxDC, MxLIVE, in integration with the AutoProcess software package, provides data collection and experiment management [209]. AutoProcess is a Python code to run XDS [210], POINTLESS [211], BEST [212], XDSSTAT [213], and the CCP4 package [214]. To evaluate data quality, AutoProcess calls XTRIAGE from PHENIX [215].

At the Shanghai Synchrotron Radiation Facility (SSRF) biological macromolecular crystallography (MX) beamline [216], an automatic data processing and experiment information management system compatible with *BluIce* beamline-control system [180], the so-called Aquarium was designed. It contains three modules: one is for data processing, from data reduction to a model building if the anomalous signal presents. Another part is a daemon that submits processing jobs to a high-performance cluster triggered by the end of the data collection, and the last module is a website that is used to input the sample information and inspect the processed results of the collected data. The data processing module uses CCP4 [214], XDS [137], autoPROC [217], DIALS [139], xia2 [218], and SHELX packages [219].

Despite the advancements in MX, there are still challenges that need to be addressed. One such challenge is radiation damage, which can limit the quality of the collected data. Additionally, MX techniques may not always meet the requirements for proper time-resolved measurements, which are necessary for studying dynamic processes. While MX beamlines have made significant progress in automation, changing samples remains time-consuming and can hinder efficient data acquisition.

Moreover, the field of MX continues to face new challenges due to the introduction of new equipment and

data collection strategies. The rapid evolution of technology and experimental techniques requires continuous development and adaptation of MX methods to ensure optimal data quality and efficient data acquisition. Below, we will talk about how these challenges can be addressed.

In macromolecular crystallography, a small rotation series is a data collection strategy where the crystal is rotated by a small angle before each diffraction image is captured. The goal of the small rotation series is to minimise radiation damage by spreading the X-ray exposure across multiple small rotations instead of a single long exposure. The key advantages of the small rotation series are its ability to provide more complete and comprehensive data and that rotations are usually fast. Moreover, this method addresses the partiality problem often encountered in serial crystallography. By collecting data from multiple crystal orientations, the partiality bias is minimised, resulting in more reliable and representative measurements. The data from these crystals can then be combined into a complete set of reflection intensity data. KAMO [220] is an open-source data processing pipeline designed to automate the entire data processing procedure for multiple small wedge datasets utilising existing programs, including XDS [210], DIALS [139] and CCP4 [214]. The results are reported in an HTML file with visualisation using amCharts and D3.js.

Serial crystallography also presents unique challenges in data processing compared to traditional macromolecular crystallography. Specialised algorithms and data quality evaluation strategies have been developed to handle serial crystallography data, including programs such as Cheetah [133], OM [221], DIALS [139], and `CrystFEL` [141], etc. The processing software can be divided into two groups: online and offline. Online (like OM [221]) is used to analyse some diffraction patterns in real-time and provide some useful information (for example, hit rate) with minimal delay to help make strategic decisions about ongoing data collection. For example, it can be used for tuning the sample delivery (like a jet) position for optimal intersection with the X-ray beam. The second stage (offline) performs non-hit rejection and data conversion and runs `CrystFEL` for indexing and integration. All specially developed software provides tailored solutions for processing and analyzing the diffraction data obtained from serial crystallography experiments. Their development has greatly facilitated extracting of valuable structural information from serial crystallography datasets. A dedicated beamline exists that offers a unique experimental approach, combining elements from both macromolecular crystallography (MX) and serial crystallography (SX). One example of such a beamline is the BL32XU micro-beam beamline at SPring-8, which features an automated data collection system called ZOO [222]. ZOO facilitates the automation of various goniometer-based data collection protocols and performs faster data collection using the 'fast raster scan' system and the spot-finder by calling a `peakfinder8` [133] from the SHIKA program. This system optimises data collection by avoiding crystals with low diffracting power through low-dose raster scanning. ZOO offers several advantages, including considering radiation damage through KUMA, exploiting RADDPOSE [223], and selecting better datasets for merging by KAMO that use the `cctbx` library [218] and utilising XDS [137]. KUMA predicts the absorbed dose in the crystals and proposes suitable exposure conditions using user-defined parameters. As mentioned earlier, dedicated programs have been developed for processing serial crystallography (SX) data. SACLA is a notable facility that has utilized modified versions of Cheetah [133] and `CrystFEL` [141] to adapt them to the experimental and computational environments [224]. These tools offer real-time feedback and enable rapid structure solutions during beamtime. However, fine-tuning the analysis parameters for each experiment still presents a challenge and typically requires the presence of an expert in data analysis on-site. This highlights the need for expertise and optimisation in SX data processing workflows. The DA+ software, developed at the Paul Scherrer Institute and implemented at all three Swiss Light Source (SLS) MX beamlines (X06SA, X06DA and X10SA), is a data collection and analysis software consisting of distributed services and utilities, which communicate via messaging and streaming technologies [179]. The user interface,

acquisition engine, online processing and database represent the main components of DA+. DA+ also provides fast feedback on data quality assessment through distributed automated data analysis routines. A reliable network for DA+ is a key part of effective communication between beamline consoles, hardware components, data storage, compute clusters, and a database. Via DA+ GUI, the user defines an experiment. The request is sent to the DA+ server to execute the experiment. The data processing results are inspected in the web-based adp-tracker. The SSX suite is an extension of DA+ to perform high-throughput, efficient measurements on many crystals [225]. The following steps are executed for the SSX data collection and subsequent analysis. First, the DA+ GUI accomplishes sample mounting and centring (1) and identifies well-diffracting micro-crystals with a fast grid scan (2). Then, the CY+ GUI utility provides an efficient evaluation of the results of the grid scan analysis and the subsequent collection of multiple wedges of data from automatically selected positions in a serial and automated manner (3). The adp invokes instances of processes (JobWorkers) that process a single SSX mini set using fast\_xds (4). The automatic data merging utility (adm) performs online merging using the XSCALE program (5). Finally, the results of (5) and (6) are sent to the MX MongoDB database and displayed in the Web-based tracker (7). The highly Automated Macromolecular Crystallography (AMX) [226] and the frontier microfocus macromolecular crystallography (FMX) [227] are two macromolecular crystallography (MX) beamlines at the National Synchrotron Light Source II [228–230]. The data collection control software on FMX and AMX is the Life Science Data Collection graphical user interface (LSDC) based on MXCuBE2. This interface has three main parts: sample queue, data collection parameters, and sample viewing. Also, the user can select the appropriate data processing and structure determination pipeline. Depending on the type of experiment, there are multiple choices of pipelines. For conventional measured data, the user can select a data reduction pipeline fastDP [186]. fastDP exploits XDS [137], CCP4 [214], and CCTBX [218]. For novel structural determination, fastEP [https://github.com/DiamondLightSource/fast\\_ep](https://github.com/DiamondLightSource/fast_ep) can be executed, utilising SHELX [231]. Incorporated in the LSDC GUI, DIMPLE from the CCP4 package [214] is a molecular replacement and ligand visualisation pipeline for drug and ligand-screening experiments. Users provide structure models and enter other necessary information associated with each crystal into a sample information spreadsheet. The PyMDA multicrystal data processing pipeline [232, 233] is a data processing pipeline for SX that utilises DIALS [139] for processing single-crystal data sets and scales them using CCP4 programs POINTLESS and AIMLESS [214, 234]. Another possibility to process SX data is to use the WYpeline pipeline [235] developed for ultra-fast raster scans or to run CrystFEL [25]. ISPyB [177] is used to manage sample information. Collection and processing results are displayed through SynchWeb [178]. Moreover, other crystallographic software can be used during data processing, such as DOZOR [236] and DIALS hit finder [237] for spot finding in raster scans and serial crystallography data processing, DIALS [139] and HKL2000 [238] for data reduction, SHELX [231] and HKL2MAP [239] or novel structure determination, CCP4 [214] and PHENIX [215].

## 4.7 Modern problems in developing data processing pipelines for serial crystallography

Although macromolecular crystallography has long been a well-established technique for studying biomolecular structures, serial crystallography offers a potential solution to overcome the limitations discussed in Section 4.2. These limitations include the need for large crystals for high-resolution measurements, concerns about radiation damage, the presence of dynamic effects, and the challenges in meeting the requirements for time-resolved experiments. However, serial crystallography introduces its own demands and challenges regarding

data processing. Some of these challenges include:

1. The collection of a vast amount of datasets for various samples during the experiment. To obtain high-resolution datasets for samples with significantly different unit cell parameters, adjustments to the detector distance and centre in the geometry file are required each time such changes occur.
2. The tuning of parameters for hit finding and indexing algorithms significantly affects the overall statistics after the merging step. Mis-indexing can result in a substantial drop in correlation coefficients at high resolution.
3. Failure to apply proper masks can lead to many false-positive results, impacting the final model quality metrics.
4. The manual inspection of possible issues, such as shifting detector centre or peak saturation, becomes challenging when dealing with an enormous amount of data.

Another critical requirement for data processing in serial crystallography discussed further in Section 5.4, is the need to provide detector geometry with sub-pixel precision to retrieve maximum information from collected raw data. The list of potential issues extends beyond those mentioned above. Furthermore, achieving full automation for both conventional and serial crystallographic experiments can be problematic due to the emergence of new beamlines and X-ray sources.

Existing solutions for automating data processing pipelines at beamlines may not meet the requirements of these new experimental setups and collection strategies. Thus, the development of data processing pipelines should consider the possibility of adaptability without rebuilding the entire concept and the ability to adjust to different installed control systems at various beamlines, which is an imperative focus of Chapter 5.



---

# Improving data processing in protein crystallography

The ultimate goal of any protein crystallography experiment is to obtain a reliable and accurate structural model that represents the molecule under investigation. However, the quality of the model and the underlying data become crucial aspects to consider.

In this chapter, we delve into the intricacies of data processing and quality assessment in the context of structural biology. The process involves collecting X-ray diffraction data from protein crystals and applying mathematical algorithms to extract structural information. The initial step in structure determination is the measurement of the amplitudes of diffracted X-rays, which provide information about the electron density distribution within the crystal.

Recent advances in facilities, detectors, and sample delivery systems have led to advanced data processing pipelines. These pipelines, particularly beneficial in serial femtosecond crystallography (SFX), provide fast feedback and reproducible results. The final results of the analysis can be significantly influenced by various factors such as:

1. Sample dependent (imperfections in protein crystals, crystal size);
2. Defined by experimental setup (energy, detector geometry);
3. Parameters for data analysis (parameters for hit finding, indexing, integration, and merging).

The detailed information about each step required to process raw data is fully described in Section 4.3.

Phasing methods such as Molecular Replacement (MR) and Experimental Phasing, like Multiple Isomorphous Replacement (MIR) and Single-wavelength Anomalous Dispersion (SAD), are commonly employed in X-ray crystallography. In MR, a known protein structure is used as a starting model, and its orientation and position within the crystal lattice are determined by fitting the model to the experimental data. Experimental Phasing techniques, such as MIR and SAD, utilise heavy atom derivatives or anomalous scattering signals to determine the phases. These techniques enable the calculation of an electron density map, which is subsequently refined to obtain a high-resolution protein structure. These methods are described in detail in Section 4.4.

Data quality assessment becomes even more critical when dealing with data collected from diverse sources and over different periods. Throughout the data processing pipeline, various accepted data quality metrics are utilised to ensure the reliability of the results [240–242]. Key metrics include measured data quality and reproducibility (e.g.,  $I/\sigma(I)$ , Completeness,  $R_{split}$ ,  $CC^*$  or  $CC_{1/2}$ ), refinement quality of the model (R-values),

visual inspection of reconstructed electron density maps, and the ability to perform *de novo* phasing and/or structure refinement. When working with data manipulation, such as applying lossy compression (see Chapter 7) or testing modern processing pipelines (for more details, visit Chapter 6), it is crucial to have a robust quality check mechanism in place. This allows us to accurately assess the impact of any modifications on the final protein structure's quality. In Section 5.1, we comprehensively discussed established data quality metrics. Additionally, we highlighted the significance of considering additional characteristics of processed data to ensure the reliability and accuracy of each processing step (Section 5.1). The developed data processing pipeline, described in Section 4.3, includes mentioned metrics that are automatically generated at each step of data processing. Moreover, it was adjusted and incorporated into the data life cycle at the drug screening P09 beamline at PETRA III (for more details, visit Section 5.6).

In Section 5.1, we emphasise the importance of thorough data quality assessment and explore various factors that can impact analysis outcomes. We delve into novel approaches that advance data collection and processing in protein crystallography. Section 5.2 primarily focuses on the strategy of refining detector distance and detector centre. In Section 5.3, we discuss hit optimisation techniques for efficient data collection using chips. Furthermore, Section 5.5 introduces a tool for automatically generating ice rings and salt reflection masks per pattern. Subsequently, we describe the automatic data processing pipeline for serial crystallography (SX) data in Section 5.4. Lastly, Section 5.6 outlines the development of an automatic data processing pipeline specifically designed for an advanced High-Throughput Pharmaceutical X-ray screening facility located at the PETRA III.

## 5.1 Data quality metrics and some hints for data processing

As mentioned previously, the determination of macromolecular structures from X-ray diffraction patterns is a well-developed field with established data quality metrics. These metrics, applied at various stages of the data processing pipeline [240–242], play a crucial role in evaluating the quality of individual diffraction images and the overall dataset, ensuring the resulting structures are of high quality.

Some common data quality metrics used in serial protein crystallography are:

1. For measuring data self-consistency and reproducibility, we are using such data quality metrics as  $I/\sigma(I)$ ,  $CC_{1/2}$ ,  $CC^*$ ,  $R_{split}$  and completeness. They evaluate the quality and reproducibility of the merged data and are usually plotted as a function of the resolution [240, 241]. These metrics are also used for selecting an optimal high-resolution cut-off. Such metrics as  $R_{split}$ ,  $CC^*$  and  $CC_{1/2}$  are calculated so that the initial full dataset is split into two subsets of the same size. Each subset is then merged independently, and two resulting sets of the merged intensities ( $I_1$ ,  $I_2$ ) are compared using the formulas presented below. We will highlight the main concepts of applying these metrics to evaluate properly processed data.

- Pearson correlation coefficient  $CC_{1/2}$ : Higher  $CC_{1/2}$  values are generally better, indicating more reliable data:

$$CC_{1/2} = \frac{\sum_{hkl} (I_1 - \langle I_1 \rangle) (I_2 - \langle I_2 \rangle)}{\sqrt{\sum_{hkl} (I_1 - \langle I_1 \rangle)^2 \sum_{hkl} (I_2 - \langle I_2 \rangle)^2}} \quad (5.1)$$

- $CC^*$ : Originally,  $CC^*$  should estimate the correlation between the measured data and the hypothetical noise-free signal [241]. Nevertheless, it was shown that  $CC^*$  is a modified version of the correlation coefficient (CC)  $CC_{1/2}$  (see Eqn. 5.2) in order to estimate the correlation of the merged

dataset with the true intensities using the assumption that errors in the subsets are random [241, 242].  $CC^*$  ranges from 0 to 1, with values closer to 1 indicating a better fit between the observed and calculated structure factors:

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}} \quad (5.2)$$

- $R_{split}$ : In SX this metric is an analogue of  $R_{merge}$ , which is used in conventional crystallography. A lower  $R_{split}$  value indicates better agreement between two halves of datasets and, thus, higher precision. A  $R_{split}$  value of around 0.05 or lower is generally acceptable for high-quality crystallographic data:

$$R_{split} = \frac{1}{\sqrt{2}} \frac{\sum_{hkl} |I_1 - I_2|}{\frac{1}{2} \sum_{hkl} (I_1 + I_2)} \quad (5.3)$$

- $I/\sigma(I)$  ( $SNR$ ): It is a metric used to evaluate the quality of the diffraction data obtained from a protein crystal. High  $SNR$  values indicate that the diffraction pattern is of high quality and the reflections can be accurately measured. In contrast, low  $SNR$  values indicate that the reflections may be weak or difficult to measure accurately, often due to noise at the diffraction patterns.
- Completeness: This metric measures the percentage of all possible diffraction spots that were actually measured. Higher completeness is generally better, providing more complete information about the structure. Lower completeness often appears due to the preferred orientation of the measured crystals (in serial measurements) or because not the full rotation was measured, or due to the missing cone when the scattering is observed at high angles (in conventional crystallography). Also, low completeness is often observed in higher-resolution shells.

2. Redundancy  $\langle n_{hkl} \rangle$  represents an average number of individual measurements of each reflection
3. Wilson B-factor: The Wilson B-factor, also known as the Debye-Waller factor, indicates the degree of order within a crystal. It approximates the average atomic B factors, which can be refined later. When the B factor is high ( $> 50 \text{ \AA}^2$ ), it indicates significant disorder within the crystal. It is important to note that various factors, including radiation damage, the content of liquid or vitreous phases within the crystals, and systematic measurement errors, can influence the Wilson B-factor. Additionally, the relative motions of molecules within the crystal packing can contribute to scattering and affect the B-factor.
4. R-values: This metric measures how well the refined structure predicts the observed data [241, 243]. Lower R-values indicate better agreement between the observed and calculated data.  $R_{work}$  value assesses the residual differences between the measured (initial  $F_{obs}(hkl)$ ) and improved (refined  $F_{calc}(hkl)$ ) structure factors, while  $R_{free}$  assesses the same differences for a subset of reflections that have not been included in the refinement process. They can be expressed as follows:

$$R_{work}/R_{free} = \frac{\sum_{hkl} |F_{obs}(hkl) - F_{calc}(hkl)|}{\sum_{hkl} F_{obs}(hkl)} \quad (5.4)$$

5. The strength of the anomalous diffraction signal and the ability to perform *de novo* phasing and/or structure refinement. Experimental phasing techniques can be employed by harnessing the intensity differences between corresponding Friedel pairs, as explained in Subsection 4.4.4. However, the anomalous signal tends to be weak when measuring at the absorption edge for light elements like sulphur or if measuring far from an absorption edge. Consequently, it becomes even more susceptible to noise and any distortions in

the data, which can inadvertently arise during data processing. For anomalous signal metrics, such as the anomalous correlation coefficient  $CC_{ano}$ , are often used. It measures the accuracy of each anomalous difference.  $CC_{ano}$  is the resolution-dependent metric, which typically becomes lower as high-resolution data are included. However, it is possible to encounter when  $CC_{ano}$  is low, but SAD phasing was successful. Thus, it can be concluded that this metric may not be a good quality indicator for SAD phasing with SFX data due to large fluctuations [244–246].

6. Resolution: In X-ray crystallography, resolution is the smallest distance between crystal lattice planes resolved in the diffraction pattern. Higher resolution data is generally better, as it allows more accurate placement of atoms in the structure. However, it is worth noting that the resolution value can be somewhat subjective and is often determined based on a subjective decision after considering other statistical indicators. The resolution cut-off can vary depending on the metric used for evaluation. For instance, it can be determined based on criteria such as a  $CC^*$  value of 0.5 or an  $R_{split}$  value exceeding 100
7. Visual analysis of the reconstructed electron density to detect non-physical chemical phenomena

Data quality metrics are commonly employed in conjunction to evaluate the quality of data obtained from serial protein crystallography experiments. It is crucial to use these metrics appropriately. For example, determining whether valuable information has been lost due to applying data compression requires us to define a set of metrics to measure potential information loss. We need to quantify not only whether the quality of the final protein structure is affected but also whether the ability to perform any of the many intermediate analysis steps is compromised due to loss of data quality.

Such metrics as  $I/\sigma(I)$ , *Completeness*,  $R_{split}$ ,  $CC^*$  or  $CC_{1/2}$  represent the quality of the merged data and its reproducibility and are usually plotted as a function of the resolution [240, 241]. Thus, these metrics are generally used for selecting an optimal high-resolution cut-off. Nevertheless, they cannot guarantee that the reconstructed electron density is correct [247]. Additionally, metrics such as  $R_{split}$ ,  $CC^*$ , or  $CC_{1/2}$  are often plotted together to visually assess the degradation of data quality caused by artefacts at different resolutions. For example, suppose the diffraction patterns contain ice rings. In that case, these metrics will exhibit a significant drop in values at different resolutions (for example, around 3 Å), as depicted in Figure 5.1. This serves as an indication of the impact of ice rings on the quality of the data.

Plots of  $R_{split}$  and  $CC^*$  as a function of  $1/d$ , where  $d$  represents the resolution length, can be plotted together. This plotting strategy facilitates the identification of an intersection point, which can serve as an approximate and reliable estimation of the potential resolution for the final dataset. This approach is easier to use for automatic determination of the resolution cutoff due to the fact that conditions like  $CC^* > 0.5$  can be misleading (often in the case of low statistics or, for example, ice rings). We use this for the automation of binning compression (which is discussed in Section 7.4.2.4), as the resolution at which these metrics intersect can be used for the calculation of the *max-res* parameter needed for the peak finding step in the recent version of Cheetah.

Ultimately, the interpretation of the quality of the reconstructed map is based on expert judgment. In some cases, relying solely on refinement quality assessed R-values can lead to incorrect conclusions about the model quality. For instance, expert inspection of the structure may reveal nonphysical chemistry phenomena, such as atom or bond overlaps, or the presence or absence of additional electron density, such as a structural solvent, which should be incorporated into the structural model to enhance agreement with the experimental data. Therefore, the method, which involves visually inspecting the quality of the reconstructed electron density, proves to be valuable. Nevertheless, before submitting the obtained crystallographic structure model, the user has

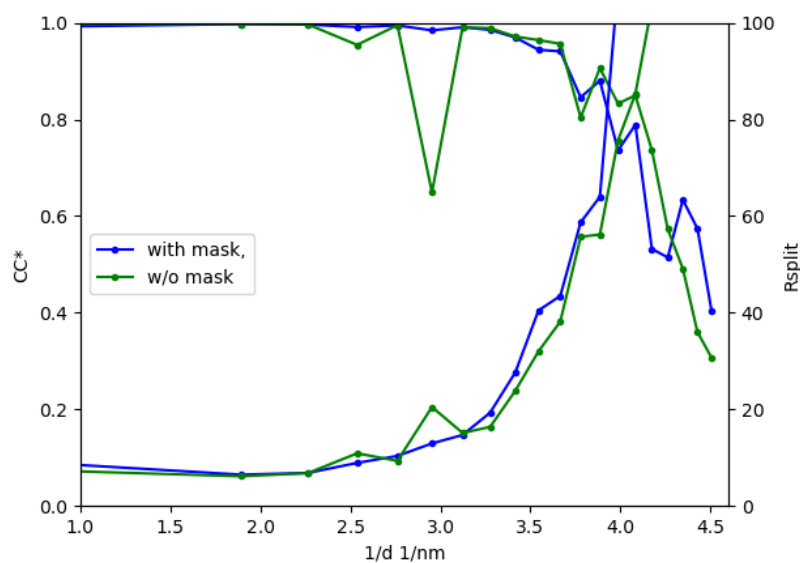


Figure 5.1: The data quality metrics  $CC^*$  and  $R_{split}$  were compared for the dataset with (blue) and without (green) a per-pattern ring mask.

to validate it against geometrical criteria and the diffraction data. Such an automated procedure as PDB-REDO helps the crystallographers improve and carefully validate their model [248]. However, including this step into a fully automated data processing pipeline remains challenging and requires an individual's expertise to determine whether the electron density appears reasonable.

Furthermore, it is worth noting that all the aforementioned statistical values are calculated within specific resolution ranges. These ranges are carefully selected and analysed to determine the appropriate resolution cut-off point for the model. By utilising high-resolution range statistics, informed decisions can be made regarding the resolution cut-off, ensuring that the model accurately represents the underlying data while maintaining optimal data quality.

As discussed briefly, in some specific cases, for example, we must establish a better approach to assess data quality degradation after applying new lossy data algorithms. Thus, one of the most appropriate methods could be to check the strength of the anomalous signal and the *ab initio* reconstructability of the structure from such data. The anomalous signals in single-/multi- wavelength anomalous diffraction (SAD/MAD) datasets are usually weak, and the method can work only if the error in the determination of the structure factors is lower than the Bijvoet differences. This is why this method, especially for SX data, usually requires good data quality to work. To enhance the sensitivity of this approach to data quality, an effective strategy is to use a subset of data that is precisely sufficient for the pipeline to operate successfully [249]. By employing this strategy, any degradation in data quality becomes crucial as it can lead to failure in reconstructing the structure. This targeted subset selection ensures that even subtle variations in data quality can significantly impact the overall reconstruction process.

While the previously mentioned metrics provide valuable insights into the quality of the final merged data and refined model structure, they do not directly address the specific issues that can arise during data processing. Factors such as incorrect detector distance or origin, the presence of saturated pixels, or unmasked unreliable detector regions can significantly impact the overall statistics of the dataset and the accuracy of the reconstructed biomolecular structure. To tackle these challenges, additional tools have been developed to assess the status of each stage in the data processing pipeline. These tools include `peakogram-stream`,

detector-shift, ave-resolution and others. The scripts for these tools can be found in the GitHub repository at <https://gitlab.desy.de/thomas.white/crystfel.git>. For example, we actively use scripts like peakogram-stream to generate peakograms (the histogram of the intensities of found peaks) and the script detector-shift, which determines the shift of the incident beam at the detector, for evaluating and monitoring the data processing stages, ensuring the reliability and accuracy of the final results. The peakogram-stream script helps users determine the maximum intensity to consider for a reflection. The peakogram represents the distribution of intensities of the detected peaks and can also highlight issues such as saturated reflections (see Fig. 5.2) or the presence of ice rings, which need to be excluded during data merging. To address this, users can set the parameter flag\_morethan in the geometry file to mark pixels as "bad" if their values exceed a given threshold. Another useful application of this script is determining the proper detector placement (or x-ray energy) for appropriate data collection. From the peakogram, we can estimate the achievable resolution for the current dataset. And if the peakogram is cut at the resolution corresponding to the detector limit, we could decrease detector distance and/or increase energy to measure at a higher resolution.

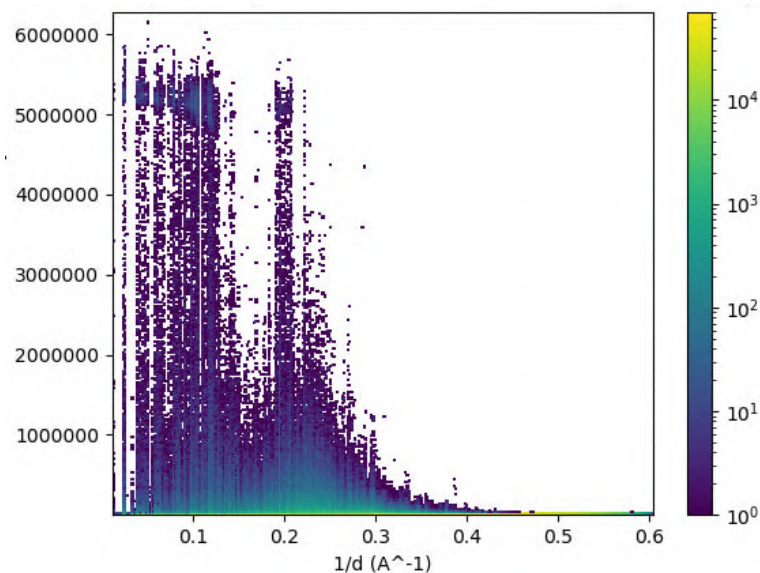


Figure 5.2: The peakogram for the dataset collected at ID29, ESRF, demonstrates the saturation problem (cloud at the level of 5000000 counts). The vertical axis represents the highest pixel value in each reflection, and the horizontal axis represents  $1/d$ , where  $d$  is the resolution length. The colour scale represents the density of points.

Similarly, the detector-shift program provides valuable information about the position of the detector centre. It can also identify if there were some issues with the experimental geometry (see Fig. 5.3). The Fig. 5.3 illustrates the splitting of the detector centre during one of the experiments carried out at the P11 beamline, PETRA III. Also, this problem is quite common for XFEL when the beam's position drifts. Thus, we have to refine the detector centre per each run. These two examples showcase the necessity of introducing auto-refining of the detector centre into the data processing pipeline, which was done and described in Section 5.4

Suppose data were collected using fixed-target sample delivery systems like chips or tape drives. In that case, it is important to analyse each indexed crystal's unit cell (unit cell (UC)) vector distribution. One way to do this is to create a three-dimensional scatter plot of the UC vectors. This plot allows one to visualise the position and density of the vectors, which can provide insights into the crystal's preferred orientation (see Fig. 5.4): If certain UC vectors are significantly more frequent or densely packed in specific regions of reciprocal space, it indicates

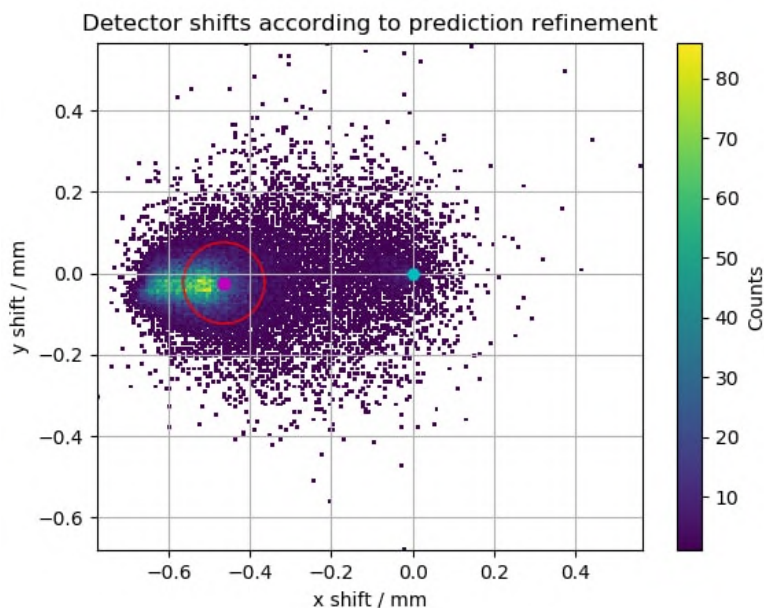


Figure 5.3: The plot generated by the `detector-shift` script illustrates the issue related to the misalignment of the detector position with respect to the X-ray beam direction.

a preferred orientation of the crystals. The truly random distribution of the crystals' orientations should look like uniformly filled ellipsoids. The modified script `orientation-v2.py` can be found in the GitHub repository at [https://github.com/galchenm/data\\_processing\\_pipeline.git](https://github.com/galchenm/data_processing_pipeline.git). This script, originally written by Yaroslav Gevorkov, is designed to visualize the distribution of the unit cell vectors, as depicted in Fig. 5.4: each subplot shows the distribution of the corresponding reciprocal lattice vector. In the case of the absence of preferred orientation, the ellipsoids should be uniformly filled. Otherwise, the preferred orientation is observed (as demonstrated in Fig. 5.4).

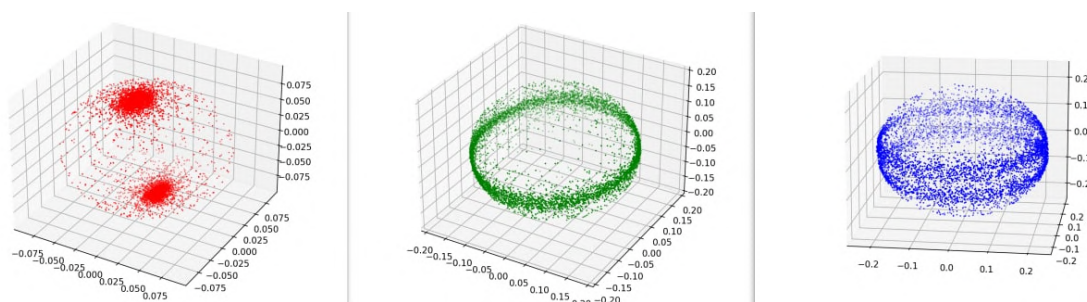


Figure 5.4: An example illustrating the distribution of unit cell vectors in reciprocal space, highlighting the presence of preferred orientation. The left plot corresponds to the distribution of  $a^*$ , the middle -  $b^*$ , and  $c^*$  is depicted on the last one.

For the experiment planning, choosing the experimental geometry that would allow measuring data to the desired resolution is important. By knowing the detector's geometry, including the pixel size, number of pixels in each direction, detector distance, and wavelength, we can calculate the resolution measurable at each diffraction pattern (see Fig. 5.5). The formula used for this calculation is shown in Equation Eqn. 5.5, where  $\Delta x$  represents the resolution,  $\lambda$  is the wavelength in  $\text{\AA}$  (calculated as  $\frac{12.4}{E}$ ,  $E$  is the energy of the x-rays in keV),  $d$  is the distance from the detector centre to the edges/corners, and  $l$  is the detector distance. One can calculate  $d = \text{pixel\_size} \times N_{\text{pixels}}$ , where  $N_{\text{pixels}}$  is usually measured from the incident beam position at the detector



(detector centre).

$$\Delta x = \frac{\lambda}{2 \sin \left( \frac{1}{2} \arctan \left( \frac{d}{l} \right) \right)} \quad (5.5)$$

The equation Eqn. 5.5 not only allows us to estimate the achievable resolution but also provides a way to determine the proper detector distance  $l$  if we know the unit cell parameter ( $\Delta x$  in this case) and the desired distance between the Bragg peaks ( $d$ ) at the measured diffraction patterns. This is needed to optimise the experimental setup and achieve the desired resolution and Bragg peaks separation. Such calculations are crucial for experiments using 2D detectors with a few pixels (1 mega-pixel (Mpix)) while measuring samples with relatively big unit cells. If a user fails to put the detector properly in such a case, the whole measurement might not be processable, and the experiment would fail.

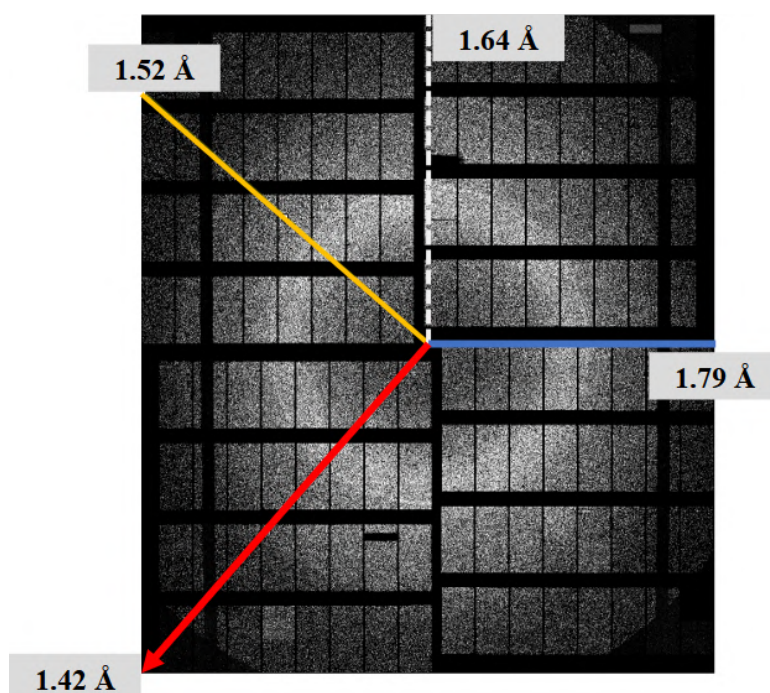


Figure 5.5: The illustration demonstrates the achievable resolution on a diffraction pattern collected with an AGIPD detector. The pixel size equals 200 microns. The numbers of the pixels to the corner and edges are 840, 552 and 633, respectively. The wavelength for the experiment was  $\lambda = 1.3 \text{ \AA}$ , and the detector distance  $l = 0.1196 \text{ m}$ .

## 5.2 The calibration of detector distance and origin using conventional crystallography

The proper data analysis requires knowledge of the detector centre and distance with sub-pixel precision. It is especially crucial for 3D merging, the idea of which is to map the diffracted intensities into three-dimensional reciprocal space instead of integrating each image in two dimensions as in the classical approach [250]. Moreover, such information can be used in other techniques, such as powder plots calculation or any analysis of the Bragg peaks position where the exact distance from the origin is required.

In 2020, as part of the project focusing on continuous scattering usage, experiments were conducted at the P11 and P14 beamlines of PETRA III [41]. During those experiments, rotational series were collected from



individual crystals under cryogenic conditions. The collected data were indexed using XDS [137] for subsequent 3D merging. XDS performs refinements on various experimental parameters, such as the beam direction and detector position. However, to effectively utilise this data, it is necessary to transform the XDS coordinate system into the detector coordinates. To accomplish this, the following procedure is proposed: the basic workflow involves the following steps: first, the coordinates of the actual detector centre are determined using the three vectors of the laboratory XDS system and the  $k$ -vectors obtained from the GXPARM file (see Figure 5.6). The required corrections to the detector centre are then calculated accordingly:

$$\Delta x = \frac{k_x \times d}{p}, \Delta y = \frac{k_y \times d}{p} \quad (5.6)$$

where  $d$  is a detector distance and  $p$  is a pixel size. Then, XDS is executed on the Maxwell cluster to validate the results by using an input file (XDS.INP) containing the already refined parameters. This step ensures that the detector centre is accurately determined and fixed, which is crucial for the subsequent merging process. By performing this verification, we can confidently utilise the determined detector centre for the subsequent steps in the data analysis pipeline.

The described strategy can be implemented as a standard procedure for any beamline. It involves measuring a protein crystal, such as lysozyme in a sleeve, at room temperature and performing a scan through XDS with the post-analysis steps mentioned above for calibration. This strategy offers a significant improvement in precision compared to the traditional method using ice rings, where the precision is typically limited to several pixels. We can achieve a precision of less than 0.1 pixel with the proposed strategy.

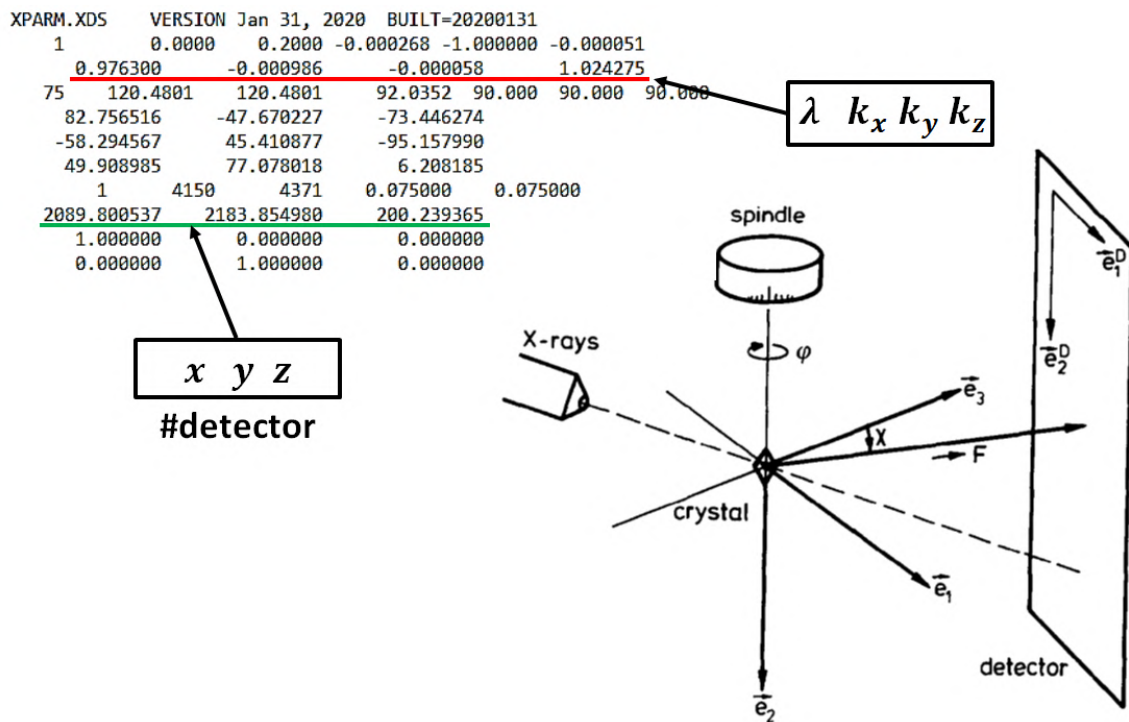


Figure 5.6: The picture is partially taken from [251]. It shows which parameters the DDR script takes to transform the XDS coordinate system into the detector coordinates.

For this particular task, the DDR program was implemented ([https://github.com/galchenm/detector\\_distance\\_res](https://github.com/galchenm/detector_distance_res)) and can be executed as follows:

```
./DDR-v3.py
```

```
[-path, The path of folder/s that contain/s GXPARM file]
[-f, File with paths to folders of interest]
[-tr, Threshold for resolution]
```

The output could look like this:

```
The number of processed files is [N]
Detector centre is ([ORGX], [ORGY]),
standard deviation for centre is ([sigma_ORGX], [sigma_ORGY])
Detector distance (mean) is [d_mean],
Detector distance (median) is [d_median],
Detector distance deviation is [sigma_d]
```

DDR also generates a log file with the resolution for each processed dataset.

## 5.3 Efficient data collection using chips

Many experimental strategies for serial crystallography are in use, depending on the type and sizes of the crystals or other needs of the experiment. Such strategies should ideally minimise the wastage of samples or beamtime without compromising experimental goals. Section 4.1.8 provides a thorough exploration of diverse sample delivery methods, delving into their unique characteristics and limitations. In this part of the chapter, the imperative focus would be on optimising data collection using solid support such as chips.

To improve the data collection with fixed-target sample delivery, we propose a two-stage scanning protocol for the chip: first, a fast fly scan (measuring during the movement) is made to find the positions of the chip with diffracting crystals, followed by a mini-rotation series (scanning over 1-5 degrees) only at each of those found positions with crystals (Fig. 5.9). In practice, it is implemented in the following way: the fly-scan is performed with a low fluence X-ray beam and at the maximum scanning speed to limit the exposure and prevent damage to the crystals. The collected data is then analyzed to determine the chip positions at which the crystal diffraction was observed. A data analysis process is initiated in parallel after each row of the raster scan to provide results shortly after the full scan has finished. Then, at each scan position where crystal diffraction was detected (according to some criterion, such as the presence of a certain number of Bragg peaks), a rotation series is collected over a small range of angles while not exceeding the total tolerable dose. Such an approach speeds up the data collection and reduces the total data volume collected. Due to the fact that pre-scanning is performed in the same configuration as the actual data acquisition, the chances of missing some crystals or scanning non-crystalline samples are rather low.

Here, we demonstrate a proof of principle of this “smart” X-ray chip scanning by introducing the intermediate step of crystal localisation into the CrystalControl software developed at the P11 beamline of the PETRA III synchrotron radiation facility in Hamburg, DESY.

### 5.3.1 Experimental setup and data collection

A micro-patterned silicon chip (Suna precision) with a  $4 \times 10$  mm size was used as a fixed-target sample holder [112] (see Fig. 5.7). The silicon chip was perforated with  $25 \mu\text{m}$  holes through which excess liquid can be sucked. The silicon chip holder has a cavity that serves as a mother liquor reservoir and provides an equilibrated

humid environment for the sample. Batch crystals with a size range of 25-40  $\mu\text{m}$  were deposited on the chip, and the excess reservoir solution was sucked through the chip holes with a tissue. A thin mylar foil sleeve was used as humidity protection against drying out.

The chip was manually mounted on the goniometer using the standard magnetic mount. The alignment of the chip is performed using the in-line microscope to ensure that the centre of rotation stays at the chip for any position within the scan. The diffraction measurements were carried out at a photon energy of 18 keV using an Eiger 2X 16M detector placed 155 mm behind the sample, and the beam was focused to a spot of  $9 \times 5 \mu\text{m}^2$ . The flux of the unattenuated beam was  $5 \times 10^{12}$  ph/sec.

The P11 goniometer and data acquisition process is controlled through a custom Python-based graphical user interface (GUI) called CrystalControl (CC) [252]. In addition to the conventional data acquisition modes for macromolecular crystallography (MX), this GUI offers various features specifically designed for micro-crystallography, such as a grid scanning capability. The grid can cover the entire chip or a specific area of interest (one region of interest per data collection). The user has the capability to draw a grid directly onto the image from the in-line microscope. Two modes for grid scans are implemented in CC: fly scan, where measurements are taken during the horizontal movement of the chip, and step scan, where the chip is first shifted to a position and then the measurement is performed. In our protocol, we implemented a two-step process for data collection. Firstly, we conducted a low-dose finder scan using the fly scan method. The obtained results from this finder scan were processed using the method described below to identify the positions on the grid where crystal diffraction was detected. Subsequently, we modified the step scan to visit the determined positions for data collection sequentially.

The maximum speed of the fly scan, used for hit finding, is ultimately limited by the detector frame rate of 133 Hz (7.5 ms exposure). In practice, the speed is often limited by the speed of the goniometer movement. Thus, large steps usually require a longer acquisition time, which is then coupled with a reduced beamline transmission. For example, with an exposure time of 40 ms, a step size of 50  $\mu\text{m}$ , and 1% beamline transmission, this hit-finding scan delivered a dose of 1.3 kGy to each lysozyme crystal. In the fly scan mode, the measurements are performed at a fixed orientation of the chip.

The following protocol for sample crystallisation of studied protein was used [21]: hen egg white lysozyme was purchased from Sigma-Aldrich and dissolved in 50 mM sodium acetate pH 3.5 (140 mg/mL) and filtered through a 0.2  $\mu\text{m}$  filter. A cold solution (4 °C) of 60 mg/mL lysozyme was mixed 1:1.5 with a pre-chilled (4 °C) precipitation solution (50 mM sodium acetate pH 3.5, 0.75 M sodium chloride, 30% ethylene glycol, 11.25% polyethylene glycol 400), adapted from [21]. The mixture was incubated at 4 °C for 16 hours, mixed at intervals using an Eppendorf Thermomixer C (1600 rpm for 30 seconds, every 5 minutes). Crystals ranging in size from 25  $\mu\text{m}$  to 40  $\mu\text{m}$  were obtained, with a mean size of 30  $\mu\text{m}$ .

### 5.3.2 Data analysis

We used and compared two hit-finding algorithms, `peakfinder8` [133] or `Dozor` [236]. `Peakfinder8` finds frames with Bragg peaks by identifying regions in the diffraction pattern consisting of a specific number  $n$  of connected pixels ( $n_{min} \leq n \leq n_{max}$ ) with intensity values above a threshold determined from the radially averaged background intensity. A pattern is considered a hit when the number of regions found, each exhibiting a sufficiently high signal-to-noise ratio, exceeds a predetermined minimum value of  $N_{peaks}$ . Each horizontal line of the grid during the fly scan is saved as a separate HDF5 file, so as soon as the file is saved, the hit-finding analysis is started. To speed up the calculations, the processing of each scanned line was submitted as a job to

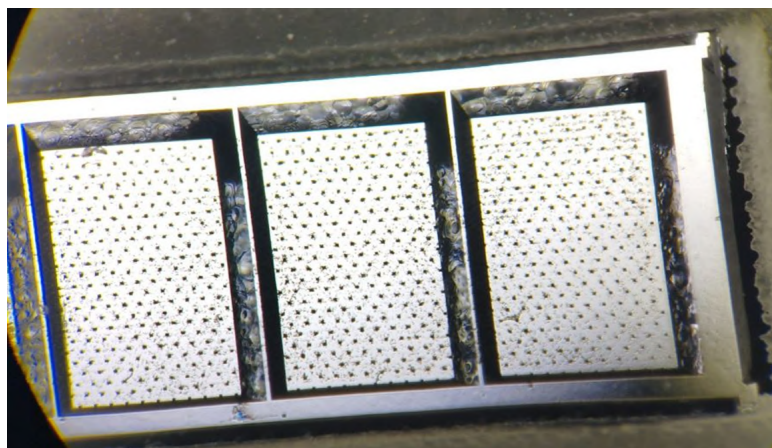


Figure 5.7: The photo of used  $4 \times 10 \text{ mm}^2$  silicon Chips, manufactured by Suna precision <https://www.suna-precision.com/products/serial-crystallography/silicon-chips>

the DESY Maxwell HPC Cluster (<https://confluence.desy.de/display/IS/Maxwell>). Since the Maxwell cluster contains many powerful nodes, this strategy performs hit-finding almost in real-time.

The Dozor program [236] was also executed for each line separately on the dedicated P11 cluster. The algorithm for finding the Bragg peaks used in Dozor is quite similar to the one used in `peakfinder8` with some differences in the statistics calculation and the implementation [236]. The hit-finding programs provide the positions of the substrate where the beam intersects crystals. A comparison of the crystal positions determined by the two programs is given in Fig. 5.8. These coordinates were then saved, but the diffraction frames recorded during the fly scan could be ultimately discarded.

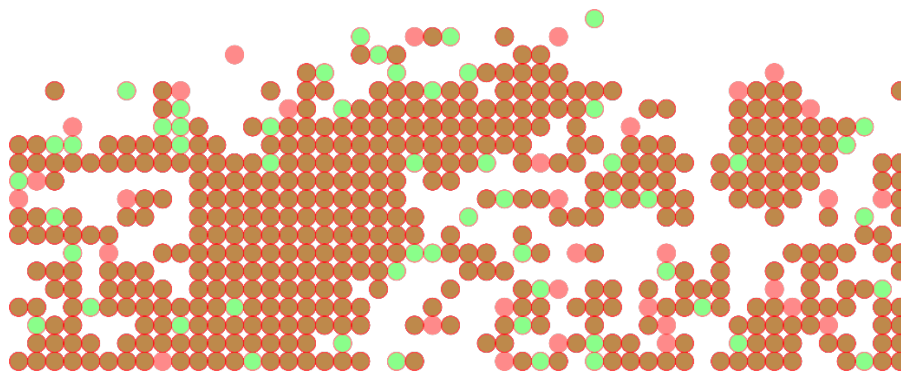


Figure 5.8: The positions of detected protein crystals by `peakfinder8` (pink) [133] or Dozor (green) [236]

After the hit-finding, the modified grid (excluding positions where no crystal diffraction was recorded) was loaded into the step scan. In a typical step scan, the chip was shifted to each occupied grid position where a rotation mini-series was performed, consisting of 11 frames, 0.36 degrees/frame. The beamline transmission was set to 10%, and the exposure per frame was 100 ms (an exposure 25 times larger than the initial scan). This measurement deposited a dose of 326 kGy, which can be considered tolerable [253].

The data collected using small rotation series at the positions of the found crystals was further processed using `CrystFEL 0.10.1`. The program `fdip_tweaker` was used to fine-tune the parameters for data processing. The `peakfinder8` algorithm was used for identifying the Bragg peaks with parameters: `--min-snr=4 --threshold=5 --min-pix-count=2 --max-pix-count=20`. We want to note here that the parameter `--max-pix-count` helps to discard the Bragg peaks produced by the silicon chip. Detected ‘hits’

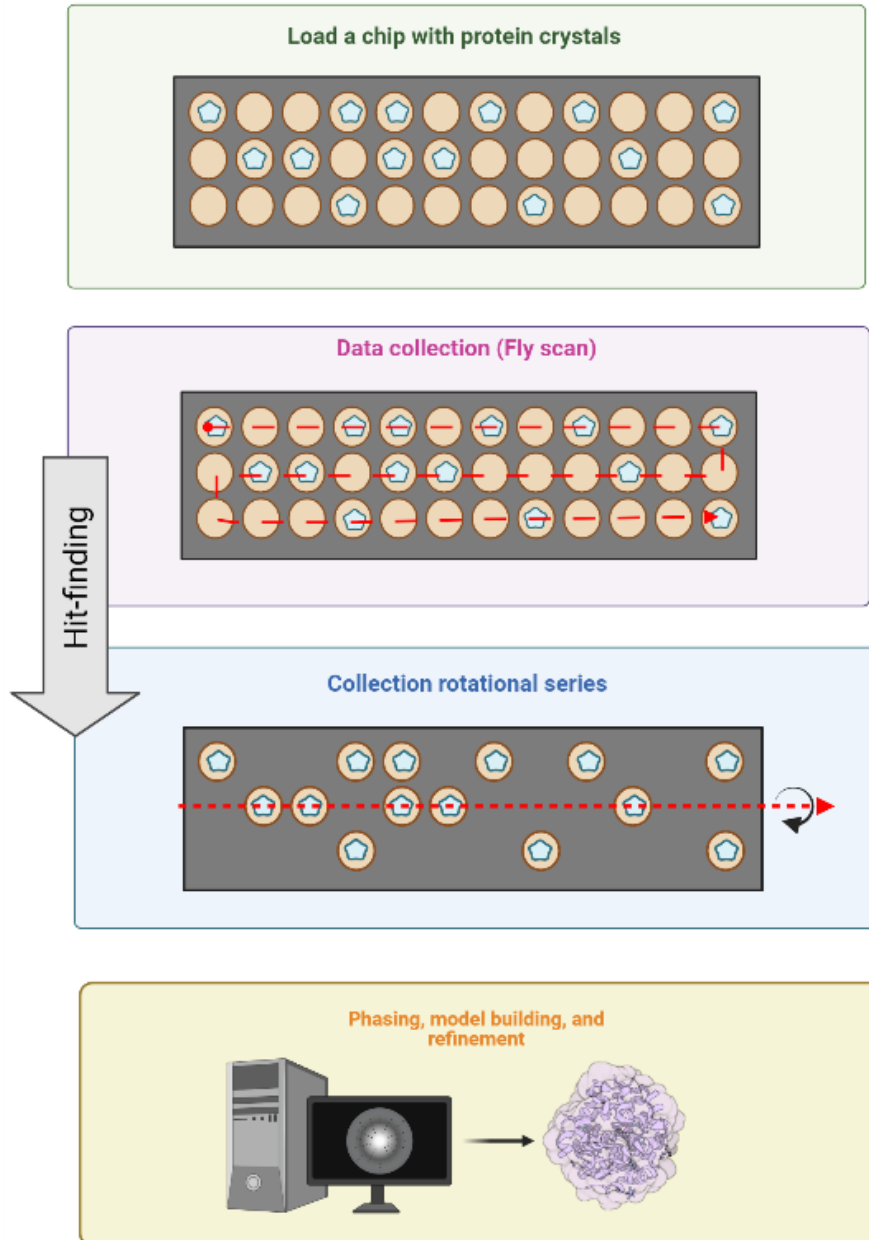


Figure 5.9: The principle of the smart chip-scanning approach. First, an on-the-fly scan is performed to locate well-diffracting protein crystals, and then mini-rotation series were measured at each detected crystal position.

were indexed using XGANDALF [147] and using `--no-cell-combinations --no-check-peaks --multi` options and integrated with `--int-radius=2, 4, 6`. Small rotational series of lysozyme from two experiments were scaled and merged into group  $4/mmm$  using `xsphere` as the partiality model by executing the `partialator` in `CrystFEL` using three iterations and `--push-res=1.0`. Figures of merit ( $SNR$ ,  $Completeness$ ,  $R_{split}$  and  $CC^*$ ) were calculated using `compare_hkl` and `check_hkl`, all part of the `CrystFEL` package, with `--highres=2.0 --nshells=20` options. MTZ files for crystallographic data processing were generated from `CrystFEL` merged reflection data files using `F2MTZ` of the CCP4 program suite.

The structure refinement of processed data was performed with `phenix.refine` [215] (Phenix/1.20) with



Table 5.1: Several scans of different loads of lysozyme samples on a chip. Every grid contained 900 positions (50x18).

Dataset	Number of indexed patterns while indexing all frames	Number of indexed patterns, while indexing only frames determined as hits	Compression rate
Lyso1_grid1	27	27	33.3
Lyso1_grid2	94	94	9.6
Lyso2_grid1	232	232	3.8
Lyso3_grid1	545	545	1.7
Lyso4_grid1	511	511	1.8
Lyso5_grid1	155	155	5.8

such parameters as `xray_data.high_resolution=2.0` and `xray_data.low_resolution=20` using 6FTR as the search model. The results are presented in Table 5.2.

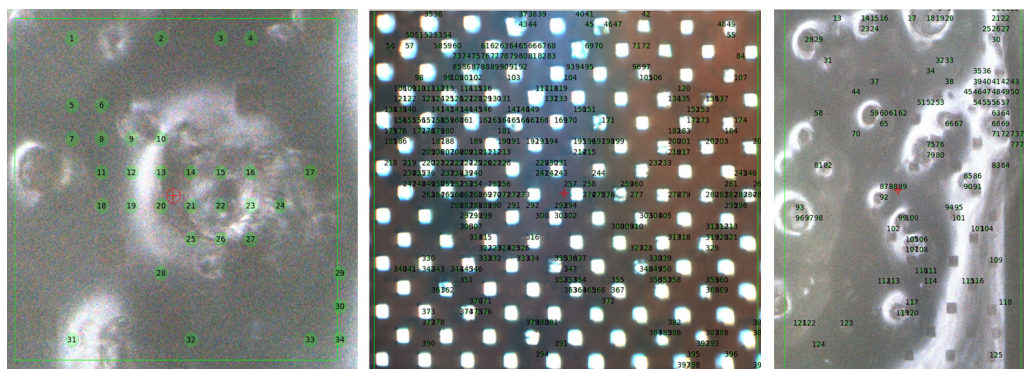


Figure 5.10: Three examples of the grids after the hit-finding. The squares are the holes in the chips, and the numbered green circles show the positions at which the crystal diffraction was determined using the peakfinder8 algorithm.

### 5.3.3 Results and discussion

To demonstrate the results of the proposed approach, three examples of scanned chips are shown in Fig. 5.10. The images were recorded using the in-line microscope at the beamline, for which the magnification and position are calibrated relative to the scanning stage. In each case, the green rectangle was set to define the range of the finder scan, and the fly-scan parameters determine the number of rows and columns in the grid. The green circles show the locations of crystal diffraction hits. Notably, these locations do not correlate with the positions of the square-shaped pores in the chips nor with visible features that might be mistaken as crystals.

The speedup achieved through the proposed approach heavily relies on the concentration of crystals deposited on the chip. Comparing the fly scan to the step scan, the former proves to be significantly faster. For instance, a fly scan covering a  $50 \times 18$  grid (900 positions) with a 40 ms exposure time and a  $50 \mu\text{m}$  step completes in approximately 100 seconds. Out of this time, around 36 seconds are devoted to the actual data acquisition, while the remaining 1-minute accounts for scan preparation, motor movement, and positioning. On the other hand, a step scan necessitates several seconds for positioning at each point, resulting in a scan with 900 points taking over 30 minutes. Scanning 545 pre-selected points with 11 frames per position took approximately 28

minutes. Step scans with 27 positions required only a few minutes to complete. This evident contrast highlights the immense advantage of investing some time in conducting a fast fly scan to determine the positions of the crystals.

To check if the positions of crystals on the chip were determined correctly, we performed the following test: the whole dataset measured during the low-dose fly scan was processed using `CrystFEL` [141], and the indexing results were compared to the result of processing a subset containing only hits found by the `peakfinder8`. In all cases, indexing the whole measured dataset led to the identical number of indexed crystals as the indexing of just the identified patterns. This suggests that measuring the positions where the crystals were not detected during the hit-finding step gives no additional information. The number of indexed crystals on several measured grids is presented in Table 5.1. Each grid consisted of 900 positions.

The last column of Table 5.1 lists the ratio of the total number of recorded frames to the number of indexable frames. Our approach offers significant advantages in terms of time saved during data recording and the amount of data saved. As described above, after determining the positions of the crystals inside the grid, the mini-rotation series were collected at each position determined as a hit. Such a dataset can be treated as serial data using `CrystFEL` – in this case, the fact that each rotation series was measured from one crystal was not considered. Alternatively, one can process each series independently using, for example, `XDS` and merge the integrated reflections obtained for different positions. While the first method is simpler for users, the second method might give better results since it solves the partiality problem within each rotation series.

We have processed data collected during two experiments to demonstrate that the measured data can be used for structure determination. The datasets were treated with consistent parameters during raw data processing and structure refinement (see the Methods section for details). The resulting statistics are summarised in Table 5.2.

Table 5.2: Overall statistics for two datasets collected during different chips after pre-determination of exact crystal positions with a low dose.

	<b>Experiment 11013662</b>	<b>Experiment 11013278</b>
<b>Number of patterns</b>	9230	15023
<b>Indexed patterns/ crystals</b>	1899/ 2088	1189/ 1321
<b>Resolution, Å</b>	79.00 - 2.00 (2.03 - 2.0)	79.00 - 2.00 (2.03 - 2.0)
<b>SNR</b>	4.81 (1.36)	2.96 (0.33)
<b>CC*</b>	0.975 (0.603)	0.962 (0.284)
<b>CC<sub>1/2</sub></b>	0.907 (0.222)	0.860 (0.042)
<b>R<sub>split</sub>, %</b>	26.82 (106.18)	35.97 (203.09)

Table 5.2: Overall statistics for two datasets collected during different chips after pre-determination of exact crystal positions with a low dose.

	Experiment 11013662	Experiment 11013278
<b>Completeness, %</b>	100.0 (100.0)	95.75 (75.12)
<b>Multiplicity</b>	59.08 (41.6)	16.25 (4.9)
<b>Unique reflections</b>	8589 (414)	8223 (311)
<b>Wilson B-factor</b>	28.03	28.18
<b>R<sub>free</sub>/</b>	0.29/	0.28/
<b>R<sub>work</sub></b>	0.27	0.27

Parts of the reconstructed structures are presented at Fig. 5.11. Visual inspection of obtained electron density maps from both datasets did not indicate radiation damage.

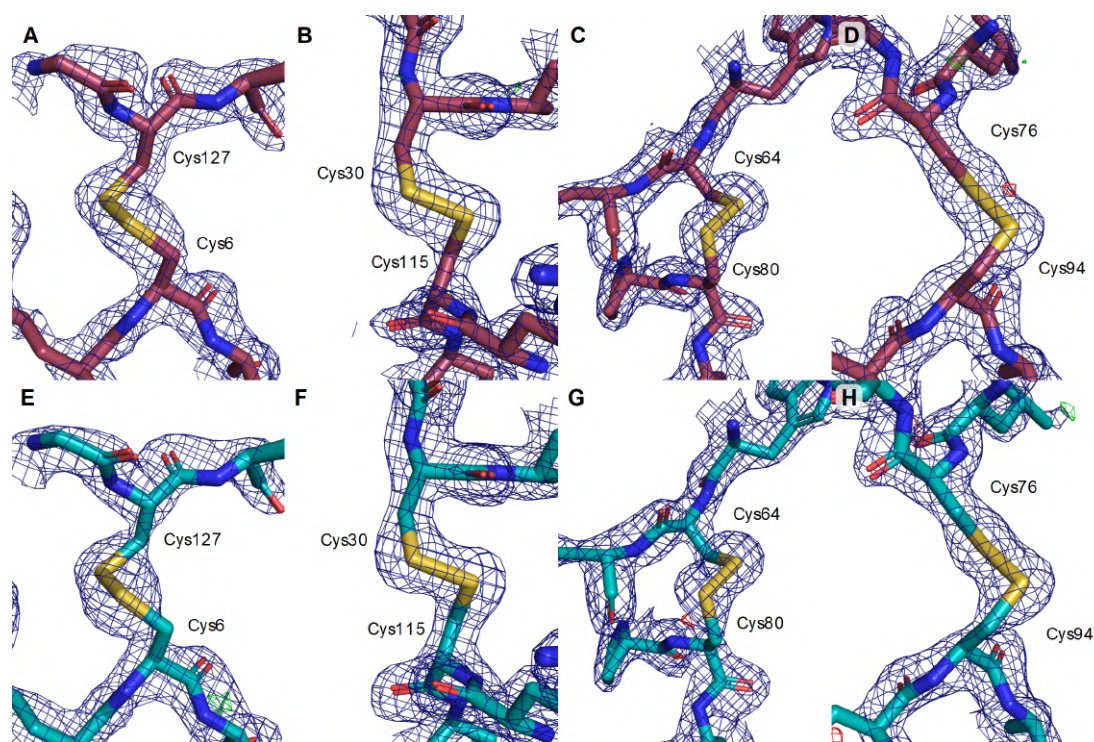


Figure 5.11: Electron-densities around the four disulfide bridges in the structures solved from two experiments: A-D – 11013278 and E-H – 11013662. Blue mesh,  $2F_o - F_c$ ,  $1.5 \sigma$ ; green/red mesh,  $F_o - F_c$ ,  $\pm 3 \sigma$ ; both carved at  $1.6 \text{ \AA}$  around the depicted atoms.

A potential improvement of the methods involves analysing the fly scan data to accurately determine the centres of the detected crystals, especially if they are larger than the step size of the scan. By measuring each crystal at its centre, more representative and comprehensive data can be obtained. Furthermore, an optimal strategy would be to adjust the size of the beam based on the detected crystal's size, possibly by modifying the



aperture. However, it is essential to consider that such improvements might require additional adjustments to the beamline hardware. Additionally, this approach might encounter difficulties when dealing with a high density of deposited crystals, where distinguishing individual crystal centres could become problematic.

The data processing of the mini-rotation using `CrystFEL` may not be optimal, as it disregards the additional information obtained from the rotation of the measured sample. Incorporating `XDS` along with `XSCALE` for processing the measured data can enhance the data quality. Such modification aims to address the partiality problem and leverage the benefits of additional constraints provided by the rotation.

The method proposed in this study is versatile and applicable both at synchrotrons and laboratory sources. In principle, the method can be utilised at FELs to measure rapid dynamical processes, such as light-activated phenomena, in protein crystals. However, even an attenuated beam would cause serious modifications to the structure. Moreover, to address the partiality problem, the rotating increment has to be small (within the divergence and the bandwidth of the beam) due to the pulsed nature of the FELs.

Optimising the scanning process offers multiple benefits, including faster data collection and reduced data volume. By avoiding the collection of empty frames from crystal-lacking positions, unnecessary data is prevented from being stored, resulting in significant resource savings. In the shown tests, depending on the concentration of well-diffracted protein crystals on the chip, we achieved storage savings ranging from 1.7 up to 33 times. The method exhibits its most significant improvement when the chip is loaded with only a few crystals, which is often the case for certain proteins that are particularly challenging to crystallise. In such instances, the proposed method becomes instrumental in maximising the utilisation of all available crystals, greatly enhancing the likelihood of successfully obtaining the structure of the measured protein.

The developed “smart” chip-scanning approach was implemented in a separate branch of the `CrystalControl` software at the P11 beamline. This approach can be easily integrated into the controlling software at other beamlines.

## 5.4 Offline data processing pipeline for serial crystallography

In recent years, advancements in data collection methods and modern X-ray sources have revolutionised serial X-ray crystallography. Researchers now have the ability to collect vast amounts of datasets from various samples, fully utilizing the capabilities of beamlines. Tools like `OM` (Online Monitor) [221] offer real-time monitoring of X-ray imaging experiments, providing users with rapid feedback on hit rates and helping them make timely decisions regarding data collection strategies for the current sample.

Some research groups are even exploring the concept of online indexing, aiming to eliminate the need for storing intermediate files and addressing the data storage challenge. However, the current status of these developments does not allow for complete reliance on the obtained results for SX experiments. As we discussed in Section 5.1, precise knowledge of the detector geometry with sub-pixel accuracy and optimized data processing parameters is crucial to extracting the maximum information from the collected data. Achieving these two aspects requires careful evaluation and proper corrections, which are only possible during offline data processing.

Moreover, the duration of experiments, often spanning several days, adds pressure and increases the likelihood of errors during data processing. Manually executing scripts and inspecting results individually become impractical under such circumstances. The situation becomes even more challenging when dealing with beam drift or the need to reposition the detector due to the large unit cell parameters of the sample. Additionally, computational resource limitations hinder the fast feedback from the data processing team.

All mentioned in Section 4.7 challenges highlight the necessity of developing a robust data processing pipeline capable of overcoming these difficulties and enhancing the quality of the obtained results. Such a pipeline would automate the data analysis process, allowing for efficient and reliable processing by generating figures of merits at each data processing step, even in complex experimental scenarios.

Here, we will present an offline data processing pipeline that fully analyses SFX data from raw images to merged *hkl* intensities with the corresponding calculated data quality metrics collected in the table format required for publishing articles. Section 4.3 provides a comprehensive discussion of the essential processing steps involved in this pipeline. The essence of this pipeline is to have a certain folder structure of processed data in order to simplify and speed up the processing and to have a reliable data quality evaluation. The main advantage of this pipeline is the ease of use and reliability of the results at each data processing stage. In addition, each script has optional arguments that make this pipeline more flexible, allowing more advanced features such as checking the detector centre followed by automatic generation of the detector geometry using calculated detector centre offsets. This data processing pipeline is universal for various file formats like cbf, HDF5, and Nexus. It is also possible to process data divided into blocks according to the same biophysical parameters, such as pH, temperature, ligand, time delay, etc. As a result, it will greatly simplify combining results in the late stages of data processing. The pipeline consists of separate blocks that are called sequentially. Such a structure is beneficial because these blocks could be used as standalone programs if the user is interested in a specific processing step. Moreover, they could be invoked externally, making the stage of integration into other programs more transparent. Below, the main workflow of the data processing pipeline is presented. The original code can be found here [https://github.com/galchenm/data\\_processing\\_pipeline.git](https://github.com/galchenm/data_processing_pipeline.git). In supplementary materials Section C.1, a detailed description of usage developed data processing pipeline is presented.

After the final step, merging intensities, we need to evaluate the data reliability. We mentioned various data quality metrics in Section 5.1. To decide where is the resolution cut-off point, it is necessary to plot metrics such as  $CC^*$  and  $R_{split}$  versus resolution length  $d$  (or  $1/d$ ). The plot of the behaviour of metrics such as  $CC^*$  and  $R_{split}$  versus resolution length  $d$  (or  $1/d$ ) is needed for making a cutoff decision. The cut-off point determined at this step will significantly affect the quality of the final reconstructed protein structure. Since there is usually a huge amount of collected raw data to process, manually evaluating each block and visualising the metrics of interest for each of them seem not rational and difficult to handle. To ease this step, the `many_plots-upt-v2.py` program was implemented, and the original source code can be found here [https://github.com/galchenm/plot\\_func](https://github.com/galchenm/plot_func). An example of usage can be found in supplementary materials Section C.1. In Fig. 5.12, we can observe the possible outcomes of the script `many_plots-upt-v2.py`.

As mentioned in Section 4.3.1 and Section 4.2, handling a bad pixel mask is an important part of SFX data processing. X-ray detectors may have broken pixels or areas that should be excluded from data processing due to the experiment setup (for example, shadows). `CsPadMaskMaker` is a graphical interface for making a static pixel mask (<https://github.com/kbeyerlein/CsPadMaskMaker>). It was developed before the virtual datasets (vds) appeared in HDF5 file format. This program was updated with features such as radial background subtraction and applying polarisation correction. Also, support for new file formats (3D HDF5 files and files with VDS data) was added (<https://github.com/galchenm/vdsCsPadMaskMaker>). The background subtraction itself is done by evaluating  $I(r)$  at the respective radius of every pixel and subtracting it from the intensity of the pixel. The details regarding the installation and usage of the updated version of `vdsCsPadMaskMaker` are presented in Section C.1. In Section 5.5, we will talk in detail about generating

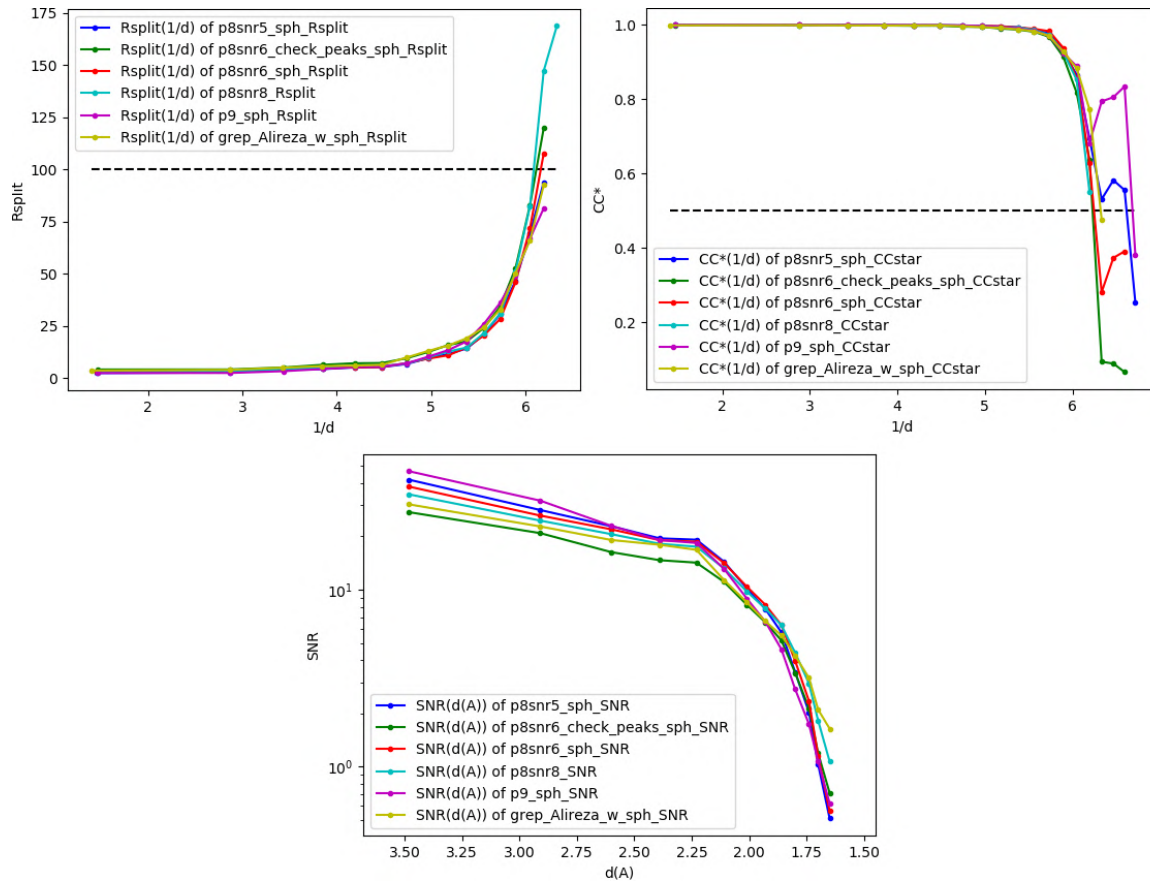


Figure 5.12: The  $CC^*$ ,  $R_{split}$ , and  $SNR$  metrics were visualised to compare the results for a specific dataset collected at EuXFEL with the AGIPD detector. The datasets were processed using different processing parameters.

static ring masks for each pattern.

In Fig. 5.13, the block scheme of each data processing step is depicted. In Section 4.6, existing data processing pipelines at different facilities are described, and then Section 4.7 discusses the remaining problems that have not been solved yet. The developed pipeline overcomes the mentioned issues in Section 4.7. The key idea behind this pipeline is to employ the plug-and-play principle, which means the whole pipeline can be adjusted without any problem to new experimental setups and include processing steps for newly appeared data collection strategies. As seen from Fig. 5.13, the main concept is built around maintaining the same folder structure for processed data as for raw folders. Having such consistency significantly eases the process of full automatization of data processing. Following simple steps described in C.1, inexperienced users can process serial data collected at any modern facility. The modularity of this pipeline facilitates the possibility of integrating data processing steps independently into the software control system, which is the main focus of Section 5.6. This processing approach includes the steps of auto-refinement detector centre and accumulation intermediate data quality characteristics such as peakograms, preferred orientation plots and so on, excessively discussed in Section 5.1.

This pipeline was used to process FEL and synchrotron-generated data. Nevertheless, some parameters defined in the detector geometry file must be refined before running this pipeline. Optimisation of the detector geometry is crucial for accurate peak prediction and integration. In Section 4.3.2, we mentioned that sometimes refinement of the relative positions and rotations of individual detector segments with `geoptimiser` is needed. Still, it is not required for each experiment. Generally, previously determined detector geometry could be

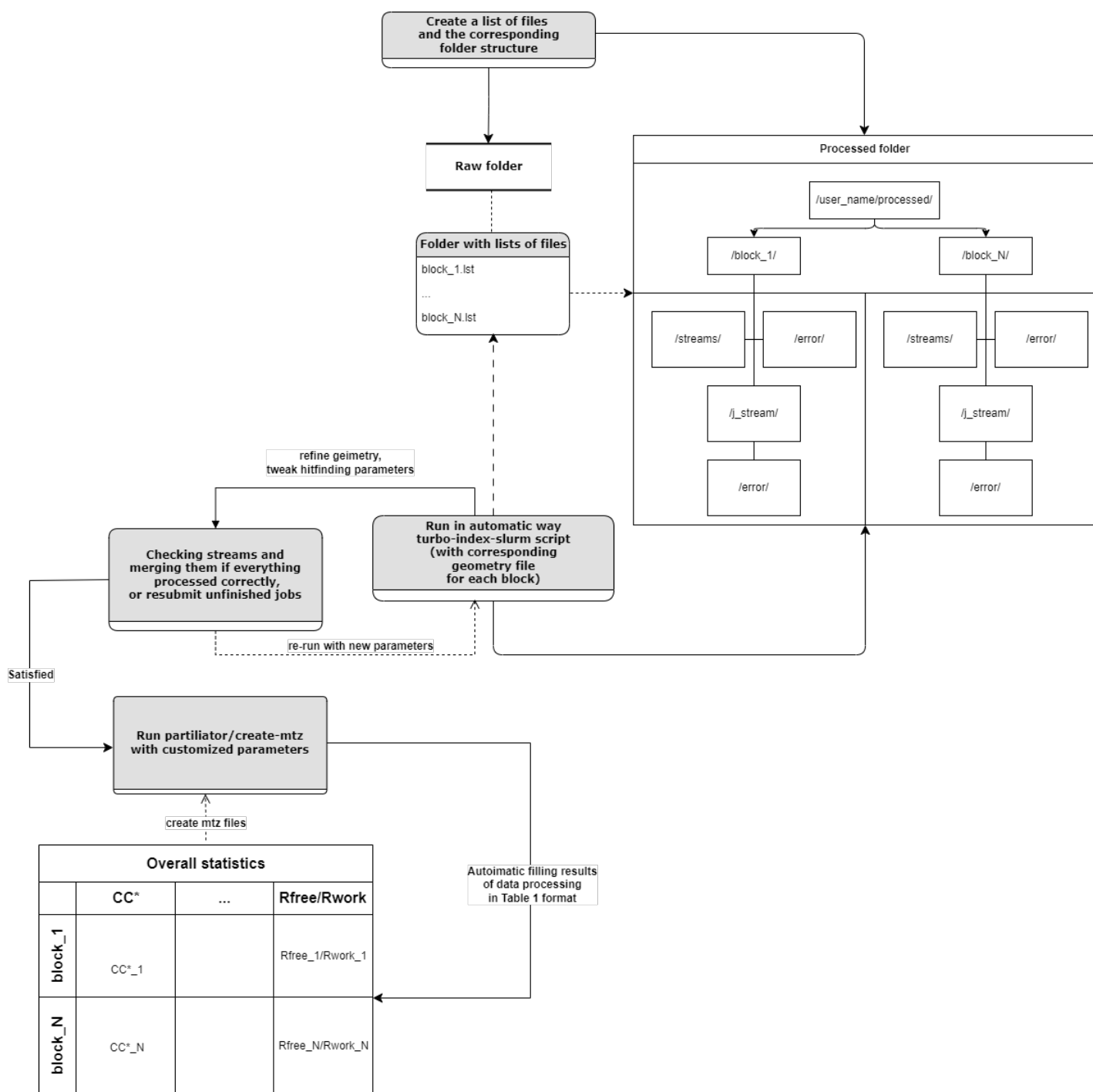


Figure 5.13: Main steps of the automatic data processing pipeline for serial crystallography

sufficient. However, parameters such as the beam centre and sample-to-detector distance must be optimised for each experiment. In the first iteration, in order to evaluate the centre of the detector, it is necessary to obtain a virtual powder pattern from all detected peaks from a large dataset and then manually align the detector to the centre of the obtained powder rings. The indexing fraction, the agreement between the obtained and expected unit cell parameters and the shape of the distributions of the obtained unit cell parameters are the best parameters to assess the accuracy of the sample-to-detector distance and knowing the relationship between these parameters and the detector distance, we can easily recalculate the real value for this parameter.

## 5.5 Tool for generating per pattern mask for salt or ice reflections

Artefacts such as ice rings and salt reflections can significantly affect the final result of data processing. Conventional crystallography usually excludes these artefacts by omitting the whole resolution ranges where the rings are observed, leading to poor data quality. However, because of the nature of serial crystallography, the advantage is that such artefacts may not be present in every recorded diffraction image. So, the resolution ranges containing rings must be excluded only for the patterns where the rings are detected. Thus, we can get more information from the collected data, including the problematic resolution regions. Nevertheless, because these artefacts may vary from frame to frame, a static mask, used to mask shadows from the experimental setup and misbehaving pixels, is unsuitable for solving this problem - a proper mask must be generated for every diffraction pattern. A Python script was developed to automatically create a salt and/or ice-ring mask per pattern in serial crystallography.

A radial curve is first calculated to identify the ring artefact. Then, the median filter is applied to smooth this radial curve. After that, the difference between the original and the smoothed curve is calculated, and all regions where the difference is greater than the specified threshold are masked. In this way, the resolution ranges containing sharp rings (typical for ice or salt diffraction) are excluded.

Another artefact the ring-masking algorithm cannot handle is bright and wide Bragg spots associated with salt or silicon (often used for the sample delivery) diffraction. We modified the `peakfinder8` outlier mask to identify these blobs. `peakfinder8`, for each resolution, identifies outliers, masks them, and then examines these masked areas regarding the number of associated pixels. If the number of connected pixels exceeds 20, this region is added to the output mask generated for this diffraction pattern.

In conventional crystallography, such a mask will lead to poor data quality due to the absence of the whole range of resolution where these artefacts were observed. In serial crystallography, due to the collecting of snapshots from a vast amount of randomly oriented crystals, the presence of ice and salt features in each diffraction pattern will be approximately low. Thus, masking ice rings and salt reflections will not lead to omitting the whole resolution bin.

A two-stage mask creation was implemented for each pattern in the script, which can be found at [https://github.com/galchenm/ring\\_mask\\_auto](https://github.com/galchenm/ring_mask_auto). An example of usage is provided here:

```
./main.py
[-p, The folder with files]
[-f, File contains all files that it is necessary to copy to the folder ]
[-o, Output folder]
[-e, Extension of files (cbf, h5, nxs or cxi)]
[-h5, The path to data in files]
[-m, Static mask]
[-mh5, The path to data in static mask]
[-g, Geometry filename]
```

Later, this script was upgraded by adding a non-hits rejection step ([https://github.com/galchenm/auto\\_ring\\_salt\\_masking\\_with\\_non\\_hit\\_rejection](https://github.com/galchenm/auto_ring_salt_masking_with_non_hit_rejection)). It helps to avoid mask creation for blank patterns or patterns with insufficient information.

The developed program was tested on datasets acquired from the experiment carried out at drug-screening HiPhaX beamline PETRA III in July 2023. Kapton and silicon chips were used as the sample delivery systems.

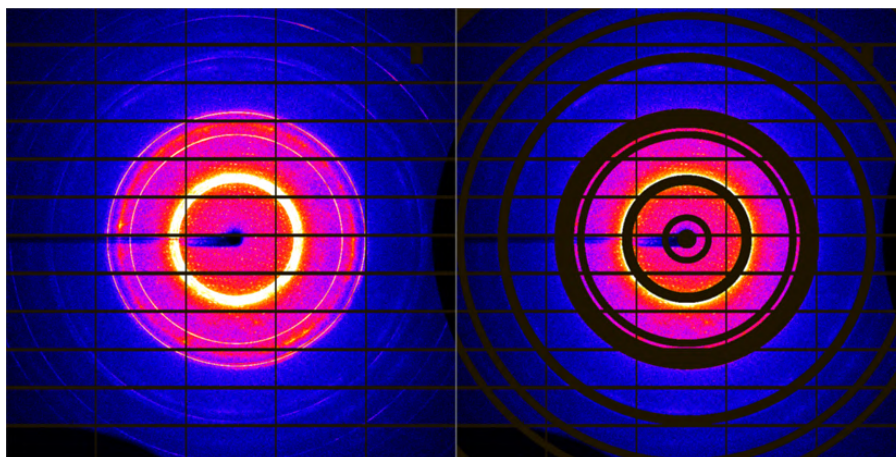


Figure 5.14: Example of the pattern of Sample 1 on the left side is without RingFinder mask and on the right side with RingFinder mask.

More than 440 datasets were collected in total. *Klebsiella Pneumoniae* fosfomycin resistance protein (FAKP) was studied with 97 different ligands. Particularly, the evaluation of the impact of different masking approaches was performed as a part of summer student work (DESY, 2023):

- Using parameter `flag_morethan` in CrystFEL, which masks the pixels with intensities higher than the set value. We used values of 2500, 5000, 7500, 10000 and 15000;
- Static mask (edges, shadows and the beamstop);
- Dynamic mask, which includes the static mask and generated per pattern ice/salt mask.

Additionally, we performed the generation of dynamic mask combined with non-hits rejection. Such an approach lets us not occupy storage with blank frames and generate for them a dynamic mask. The details of an experimental setup can be found in Table C.2.

Here, we present the processed data of a kapton chipD, which was divided into 12 windows. For each window, FAKP with different ligands were allocated (see Table C.1). Salt reflections were detected for this chip only in `window_5`. The results are represented in Fig. 5.15 and Fig. 5.16. From Fig. 5.15, it is seen the drop in values for the reference data at  $1/d = 6$ , which indicates the presence of salt reflections. It is generally difficult to claim that one specific masking option is better across all metrics. However, overall, we can state that the dynamic mask provides an optimum solution.

The chip is constantly flushed in a chamber under humidified conditions to avoid crystal degradation. However, there is a possibility that crystals could experience a gradient of humidity if the system remains open. It is known that dehydration of protein crystals leads to water loss. Thus, it results in the shrinkage of the crystal. To visualise the dehydration process, the unit cell volume was mapped back to the position of the crystals on the chip. An example of the spatial distribution of the unit cell volume for the `chipD` is shown in Fig. 5.17. As can be seen, the unit cell volume is in the range of  $280 \text{ nm}^3$  and  $305 \text{ nm}^3$ . The vertical white lines indicate that these window positions in the chip were skipped during the measurement.

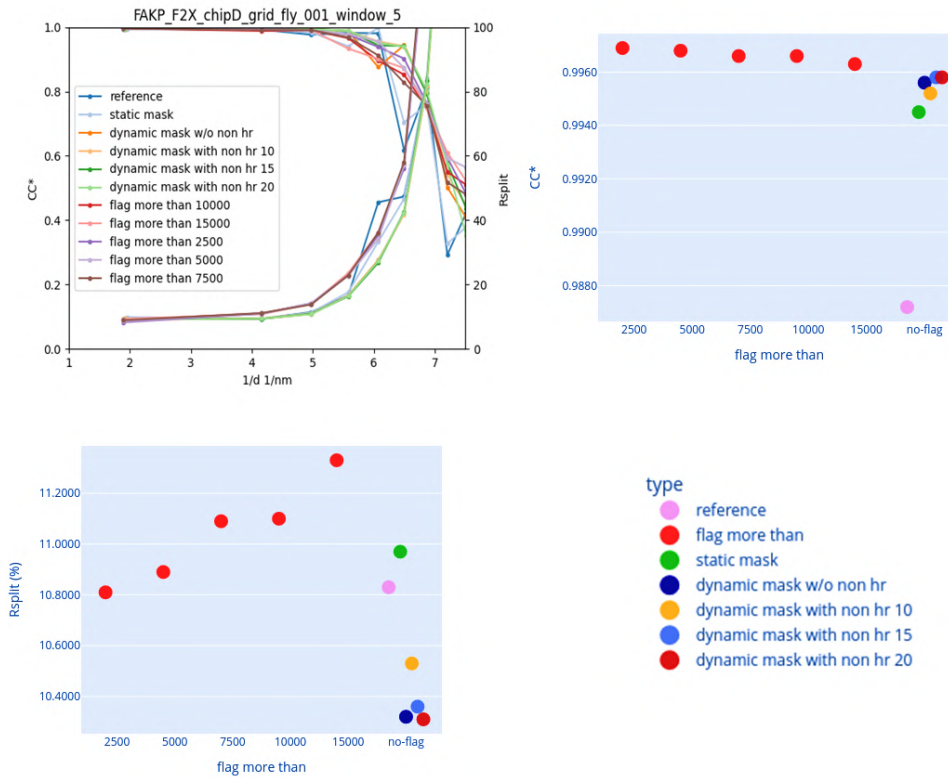


Figure 5.15: The top left graph shows the dependence of the  $CC^*$  and  $R_{split}$  metrics on over resolution. The top left and bottom right plots depict such overall statistics as  $CC^*$  and  $R_{split}$  for performed tests. If we look at the overall  $CC^*$  graph, it may seem that masking reflections using the `flag_morethan` is more optimal than other tests. However, looking at the neighbouring overall  $R_{split}$  graph, we see that the dynamic masks have better overall  $R_{split}$  values. We can also notice that increasing the value of `flag_morethan` negatively affects the data quality.

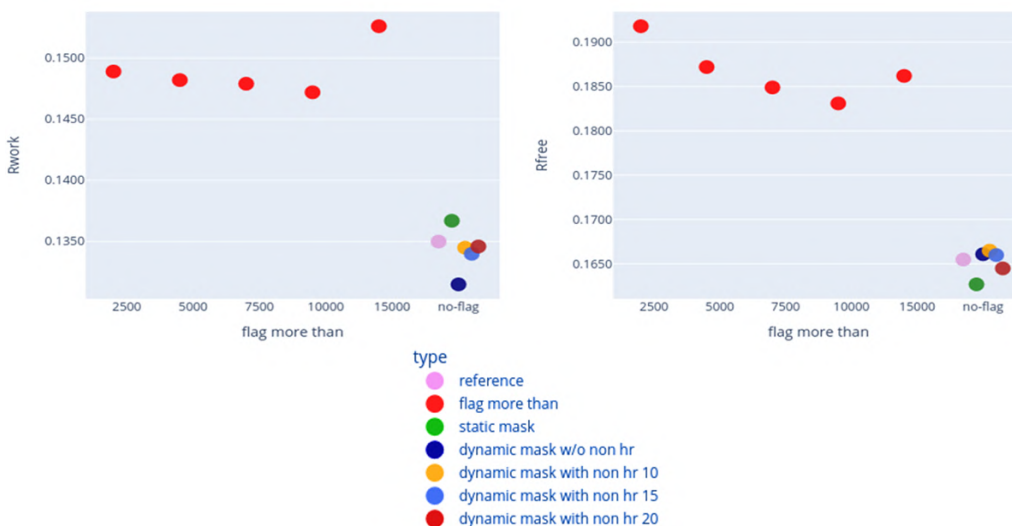


Figure 5.16: Dependence of  $R_{work}$  and  $R_{free}$  on the masking approach is depicted. The values for the dynamic masks are also better here than for the test with using `flag_morethan`.





Figure 5.17: Spatial distribution of unit cell volume of FAKP crystals on the chip with  $6 \times 2$  compartments, each  $4.22 \times 4.8 \text{ mm}^2$  in size.

## 5.6 Auto-processing pipeline for HiPhaX - a drug screening beamline P09, Petra III

In this part of the chapter, we will discuss the development of a pipeline for automatic data processing in serial and conventional crystallography, specifically for HiPhaX - the drug screening beamline P09 at Petra III, described in Section 3.3.1. This pipeline is based on the one described data processing approach in Section 5.4 and enables both online and offline data handling for conventional and serial data collection strategies automatically. The pipeline supports all detectors employed at HiPhaX: Lambda (1.5M), Pilatus CdTe (2M), Pilatus 6M and Eiger 4M. Customised online monitoring of data processing results is performed by extracting them into Google Sheets. Due to the modularity of the established software, initial robust processing can be invoked directly from the established at P09 software control system named Janus. A huge advantage of the developed data processing package is its independence from the experimental setup and data collection method. This makes the pipeline data universal and easily changeable to meet new needs arising on other beamlines.

In this part of the chapter, a detailed description of the developed data processing workflow, from data collection to obtaining quality assessment results, is offered. The Outlook section focuses on the current status of the P09 beamline and outlines future goals for further advancement.

### 5.6.1 Specification of configuration file for current experimental setup

To streamline data processing, an automatic pipeline has been developed, with the main workflow depicted in Figure 5.18. The associated scripts can be accessed at the following GitHub repository: <https://github.com/galchenm/P09>. The workflow requires a filled configuration file with such information as experimental setup (the offset of detector distance, detector centre, detector geometry), the path to raw data, and the output folder and templates of files (XDS.INP file for rotational experiment and `turbo-index-p09` script for serial crystallography) needed for further data processing. Depending on the type of experiment, the user can specify the software that the workflow will run for processing raw data. For example, in the case of conventional crystallography, the user can run the `xds_app` [137] or `autoPROC` program [217, 254]. The `turbo-index-p09` script can be configured for SFX with the hit-finding parameters and indexing method.



Still, these options can be adapted later for offline data processing, which can be done using the pipeline described in 5.4. An example of the content of the configuration file can be seen below.

```
crystallography:
  XDS_INP_template: "XDS_template.INP"
  ORGX: 2560
  ORGY: 1234
  DISTANCE_OFFSET: 20
  command_for_processing_rotational: "xds_app"
  command_for_processing_serial: "turbo-index-p09"
  raw_directory: "/asap3/petra3/gpfs/p09/2022/data/11016565/raw"
  converted_directory: "./convert"
  processed_directory: "./processed"
  geometry_for_conversion: "geometry_for_conversion.geom"
  geometry_for_processing: "geometry_for_processing_template.geom"
  data_h5path: "/data/data"
  cell_file: "lyzo.pdb"
```

However, before starting the experiment, it is necessary to perform some preliminary steps to establish sufficiently accurate values of the distance to the detector and its origin. There are several ways to obtain sufficiently accurate values of these parameters based on experimental data. The detector centre could be obtained by constructing a virtual powder pattern from all detected peaks in a large set of collected diffraction patterns and manually aligning the detector to the centre of the resulting powder rings. The centre is determined more accurately during the prediction refinement procedure in `indexamajig`. As for the sample-to-detector distance, people usually adjust it by maximising the indexing fraction or checking the agreement between the obtained and expected unit cell parameters. Unfortunately, both methods might fail: a fraction of indexed patterns might get into a local maximum, especially if the detector distance is far from the real one. The determined unit cell parameter might be wrong due to the incorrect energy of X-rays or some changes in the reference crystal (for example, due to different humidity). Therefore, for a more accurate estimate, the shape of the distributions of the obtained unit cell parameters can be used - at a true distance, the distribution of the UC lengths and angles is usually symmetric (except when using XGANDALF indexing algorithm with hexagonal UC). Another approach to determining the detector distance is based on measurements of a standard sample at two different distances. It could be serial data or powder diffraction. Based on the measured features corresponding to the specific distance, we can recalculate the actual distance using the triangulation formula:

$$L = \frac{a_2 L_1 (L_2 - L_1)}{a_1 L_2 - a_2 L_1} \quad (5.7)$$

where  $a_1$  is a cell parameter determined at the distance  $L_1$  and  $a_2$  correspond to the distance  $L_2$ .

The detector distance, its centre and the beam direction could also be estimated by measuring the rotational series of the well-characterised sample at two different positions. These collected diffraction images will be processed with `xds_app` [137]. The results of XDS can be used further to execute the developed script `calibration_script-v2.py` that can be found at GitHub repository [https://github.com/galchenm/detector\\_distance\\_res.git](https://github.com/galchenm/detector_distance_res.git) to obtain the real detector centre, its origin and beam direction.

## 5.6.2 Data collection and auto-processing strategy

When the configuration file is filled with the necessary information, data collection is started. The main advantage of an established pipeline is the ability to call it from an external program, such as a GUI installed at the beamline for each block of collected diffraction images in real-time or to execute it after the experiment for automatic offline data re-processing.

Lambda 1.5M detector is not supported by XDS, so, for now, the data generated by this detector is converted into assembled cbfs, and then these converted files are further processed. The current version of the pipeline automatically determines whether a conversion step is needed or should be skipped and moved on to the next step in the workflow.

The raw data folder for the measured sample contains an `info.txt` file with important information about each collected dataset: the collection method, energy, wavelength, starting angle, and rotation increment parameter (if this is a rotational experiment), preliminary information about the distance to the detector, preliminary detector distance information, the total number of collected frames, etc. The pipeline parses this `info.txt` file to fill template files such as the `XDS.INP` or `geometry.geom` file for further processing with XDS (or its analogues) or `turbo-index-p09`. As mentioned in the configuration file, the user specifies the software that should be used for processing. Because of the implementation of fully automatic data processing without any additional manipulations or intervention, the configuration file has parameters to customise processing for both conventional experiments and serial crystallography.

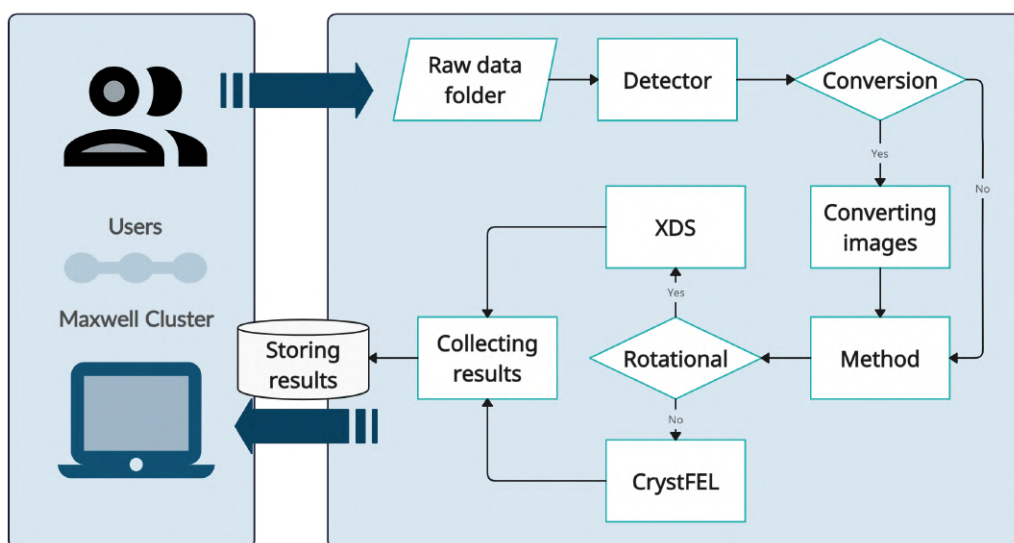


Figure 5.18: The workflow of data processing for a drug screening beamline P09, Petra III

## 5.6.3 Google Sheets as an optimal database for monitoring results and saving metadata

In recent years, there has been a heightened awareness within the high data rate macromolecular crystallography community regarding the critical significance of having thorough and uniform metadata [255–259]. This recognition underscores the necessity for metadata completeness and consistency, enabling seamless data processing at any location. This capability extends to data collected at various times, across different facilities, or spanning multiple facilities, even over extended periods, ranging from months to years in the past. An enormous amount of databases is developed dedicated to specific needs [260–262], for example, AMARCORD originally

developed for compound screening pipelines. It has been used to screen the Main Protease of the SARS-COV-2 virus against thousands of available compounds and offers rich introspection into the process [263]. However, all available databases have common issues, such as accessibility for users outside of the institution maintaining the database and their dedication to specific environments. Thus, Google Sheets should be considered an alternative way of saving metadata and saving the results of initial data processing, such as total frames, hit rate, indexing rate, number of indexed patterns and indexed crystals, resolution, and user-requested features that they consider important. In supplementary material in Section C.3.1, the fundamental steps to utilise the online interaction feature with Google Sheets are presented, highlighting that Google Sheets can log information on the actual state of data processing during the experiment, but with some changes also will be able to save metadata regarding experimental setup like energy, distance, sample name and etc. Moreover, it is easy to implement and integrate long into Google Sheets into the main pipeline due to its flexibility. Google Sheets can be reached at any time, even in offline mode and do not require any special setting from external users. The developed simple Python script, named `upt-cheetah-to-logbook-V2`, was adapted to the P09 beamline needs and introduced in the main data processing workflow. As mentioned above, this pipeline can also be used for offline data processing, in combination with the workflow described in Section 5.4 for SX.

#### 5.6.4 Outlook

We conducted a comprehensive review of existing data management pipelines employed at various macromolecular beamlines. These pipelines typically consist of three main components: control software for experimental setup and data collection, automatic data processing pipelines executed after data acquisition, and a database for storing sample information and processing results. Each component presents its unique bottlenecks and limitations.

During data collection, hardware and data transfer from the detector to memory storage often impose limitations. At this stage, raw data may undergo pre-processing, including necessary corrections like dark and gain corrections and data reduction. After data collection, the processed data is passed to the data processing pipeline, including appropriate crystallographic tools, depending on the experiment type.

Conventional crystallography, a well-established technique, has seen significant automation, from data collection to providing users with results and data quality assessments. On the other hand, serial crystallography, a more advanced method, leverages the capabilities of new-generation synchrotrons, free-electron lasers (FELs), and modern detectors but still presents challenges in full automation without expert intervention. Serial crystallography data differs from conventional crystallography as each diffraction pattern corresponds to a random orientation, leading to the unavoidable partiality problem. Moreover, new types of equipment appear that challenge even automatised MX experiments. Thus, this chapter was dedicated to developing data processing pipelines that can be easily adjusted to any software and hardware environment.

We have developed a pipeline for automatic data processing that supports serial and conventional crystallography experiments at the P09 drug screening beamline in Petra III. The hardware at P09 accommodates various sample delivery methods, such as conventional MX loops, chips, jets (GDVN or LCP), and tape drives. The experiments at P09 utilized four different detectors: Lambda (1.5M), Pilatus CdTe (2M), Eiger 4M, and Pilatus 6M. Our pipeline adapts accordingly to determine the appropriate detector.

Furthermore, our data processing pipeline offers both online and offline data processing capabilities. We have also integrated a step to automatically deposit data quality metrics for each experiment into Google Sheets, enabling users to receive quick feedback. The pipeline was partly integrated with software control on the

P09 beamline and successfully tested during an experiment conducted in March 2023 and integrated into the software control system, the so-called Janus. The developed data processing pipeline has been used for several experiments already. Current results are under publication process.

---

# Enhancing data quality through modern data processing pipelines

Not every experiment yielded success in high-resolution structure determination; in fact, some fell short of attaining the desired protein structure. The amassed data, stored on tapes or disks, not only occupies significant storage space but also necessitates substantial financial investment in expanding storage capacities. This raises crucial questions: Can we reprocess data to augment the quality of the final biomolecule structure? An intriguing query follows: Can we extract additional insights from the existing data and reconstruct structures that initially failed? Lastly, should we retain all previously collected data, or is there a means to reduce data volume without compromising scientific outcomes? This chapter endeavours to address these inquiries by examining three exemplary experiments conducted at distinct facilities.

These experiments assess the impact of reprocessing data using an established pipeline, as detailed in Section 5.4, on final results. The reliability and versatility of the pipeline are showcased across diverse datasets, with the generation of key data quality metrics outlined in Section 5.1 for meticulous data evaluation at each stage. The parameters for the hit-finding algorithm in data processing were meticulously fine-tuned using the `fdip_tweaker` software. Special attention was given to detector geometry, and a proper mask was crafted following guidelines elucidated in Chapter 5.

The initial dataset featuring lysozyme, with an anomalous signal from *Br* and collected at a synchrotron P11 beamline, illustrates how the omission of *ab initio* phasing can be rectified through judicious reprocessing with the developed pipeline. The subsequent section delves into the comprehensive reprocessing of an entire experiment conducted at LCLS in 2011, discussing not only the enhancement in resolution achieved but also contemplating the optimal storage approach for previously collected data from a data reduction perspective. The final experiment elucidates further enhancements achieved through data reprocessing and encompasses comparing structures obtained using X-ray pulses of different durations in terms of radiation damage.

## 6.1 Re-processing previously collected data

The field of X-ray crystallography is witnessing the emergence of new X-ray facilities with higher repetition rates, along with advanced area detectors that offer a greater number of pixels. However, the growth rate in storage capacity has not kept pace with these technological advances. Consequently, there is an impending challenge where raw measured data may no longer be feasible to store for extended periods before final processing. This requires the reduction of data to address storage limitations.

Concurrent with these developments, programs and algorithms for data processing are continuously improving. As a result, more valuable information can be extracted from the same old measurements, even when processed with modern software. We examined data from one of the initial high-resolution Serial Femtosecond Crystallography (SFX) experiments to evaluate the potential improvements achievable by processing old data with modern software. Our primary objective is to quantify the extent to which results can be enhanced and identify the key data processing stages that contribute significantly to these improvements.

This chapter focuses on assessing new data treatment's influence on the final results for several datasets and samples collected at various facilities with different sample delivery systems and detectors. Specifically, we analyse data of lysozyme, photosystem I, photosynthetic centre and cathepsin B collected using liquid jet and CSPAD at LCLS in 2011; hemoglobin datasets with different pulse duration collected using liquid jet and CSPAD at LCLS; and data with an anomalous signal of lysozyme with Bromine acquired with TapeDrive and Pilatus 6M at Petra III in 2015. The data collected at LCLS have been previously processed and published [12, 264–266]. However, due to reconstruction difficulties, the anomalous dataset of lysozyme with Bromine obtained at the P11 beamline at Petra III was never published. Furthermore, during the experiment described in [266], hemoglobin samples were measured using pulses of different duration (3 femtoseconds and 10 femtoseconds), and only the data obtained using a pulse of 10 femtoseconds was successfully reconstructed. These cases underscore the necessity of storing old data in a reduced format to enable reprocessing with advanced data processing pipelines, leading to improved results. Furthermore, the anomalous lysozyme dataset from P11, 2015 highlights the potential for reprocessing to facilitate the determination of the final structure by using the *ab initio* phasing method.

### 6.1.1 Anomalous dataset

A dataset containing an anomalous signal is sensitive to inaccurate data processing procedures. Not well-optimised parameters, unrefined detector geometry and unmasking unreliable regions can lead to the inability to reconstruct the final structure using *ab initio* phasing. The physics behind the anomalous signal and *ab initio* phasing can be found in Section 2.10 and Section 4.4.4, respectively. Here, we will demonstrate how proper data analysis will succeed in obtaining the final protein structure.

#### 6.1.1.1 Reprocessing previously collected SAD data at P11 beamline, PETRA III

In Section 2.10, we discuss the physical principles of an anomalous signal. In this part of the chapter, we will demonstrate the influence of re-processing previously collected data containing anomalous signals from heavy metals with a new data processing pipeline on the ability to reconstruct the structure.

In 2015, an anomalous dataset of lysozyme incorporating bromine (Br) as the anomalous scatterer was acquired at a wavelength of 1.0915 Å using the Pilatus 6M detector [267] situated at the P11 beamline, Petra III. The TapeDrive sample delivery system was employed during the experiment to precisely position the crystals within the X-ray beam for data acquisition, as depicted in Figure 6.1. Notably, this experiment marked the first successful attempt at collecting anomalous datasets using such an advanced sample delivery system. Data were saved in separate CBF files, and the geometry was refined by `geoptimiser` [26].

However, the data processing pipeline available at that time did not yield successful results in determining the protein structure due to unidentified issues. To address this, we retrieved the data from the tape archive and subjected them to reprocessing using a more refined detector geometry. For this purpose, we adopted a

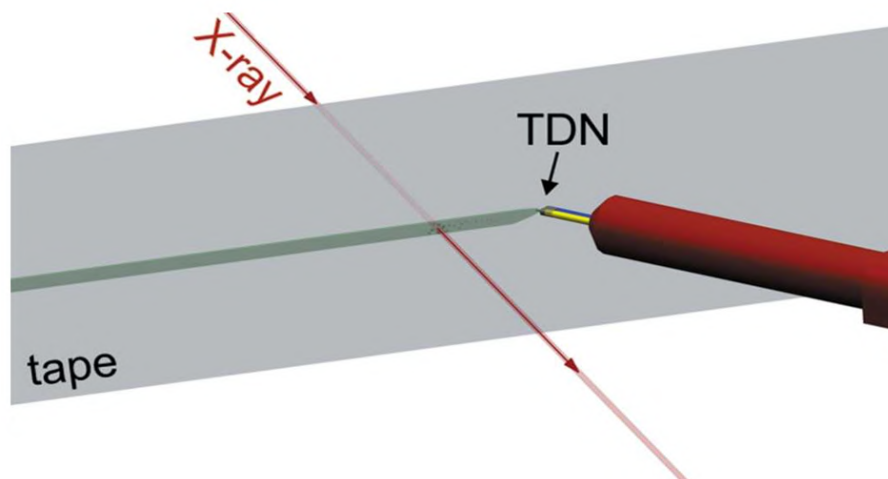


Figure 6.1: The image depicted here is sourced from [107]. It provides an overview of the sample environment setup used in the experiment. The key components highlighted in the drawing include the TapeDrive nozzle (the TapeDrive nozzle (TDN)), tape, sample line, and X-ray beam. The TapeDrive nozzle delivers the sample to the interaction region, while the tape serves as a platform for holding and presenting the samples. The sample line connects the tape to the experimental setup, enabling the controlled movement and positioning of the samples. The X-ray beam, emitted from the X-ray source, interacts with the samples at the designated region, allowing valuable data collection. This illustration offers a clear visualisation of the experimental setup, facilitating a better understanding of the sample delivery and X-ray interaction process.

new, well-established data processing pipeline that integrated the recent version of `CrystFEL` (v0.9.1) and `Phenix/1.20`, in conjunction with the `CCP4 crank2` suite [268], for the final structure refinement.

The reprocessing efforts successfully determined the protein structure through *ab initio* phasing, and the conclusive outcome is presented in Table 6.1. This case proves the necessity of saving and reprocessing old data with newly developed software to get much better results.

Table 6.1: Overall statistics of a reprocessed anomalous dataset of lysozyme with bromine (Br) obtained in 2015 at P11, PETRA III with Pilatus 6M detector.

	<b>lyso+Br, 2015</b>
<b>Num. patterns/hits</b>	474812/306574
<b>Indexed patterns/crystals</b>	159891/193162
<b>Resolution, Å</b>	31.44 - 2.0
<b>R<sub>split</sub> (%)</b>	6.32
<b>CC<sub>1/2</sub></b>	0.998
<b>CC<sub>ano</sub></b>	0.68
<b>CC*</b>	0.9995
<b>SNR</b>	21.191
<b>Completeness (%)</b>	98.944
<b>Multiplicity</b>	857.831

Table 6.1: Overall statistics of a reprocessed anomalous dataset of lysozyme with bromine (Br) obtained in 2015 at P11, PETRA III with Pilatus 6M detector.

	<b>lyso+Br, 2015</b>
<b>Total Measurements</b>	13262074
<b>Unique Reflections</b>	15460
$R_{\text{free}}/R_{\text{work}}$	0.1729/0.2331
<b>Wilson B-factor</b>	13.19

### 6.1.2 Reprocessing previously collected data at LCLS in 2011

The raw data for lysozyme, cathepsin B, the reaction centre of Photosystem II, and Photosystem I were obtained in 2011 at the Coherent X-ray Imaging (CXI) instrument located at the Linac Coherent Light Source (LCLS). Measurements were made using X-rays with an energy of 9.4 keV (wavelength of 1.32 Å). The experiment was performed at a 120 Hz repetition rate. The data, which had been stored on tape, was successfully recovered, and a comprehensive analysis pipeline was employed using state-of-the-art software and recent algorithms.

To facilitate the data collection process, a liquid micro-jet technique [53] was used to inject the crystals into the Free Electron Laser (FEL) beam while they were in their storage solution. The Cornell-SLAC hybrid Pixel Array Detector (CSPAD) 2.3M detector [269] was utilized to capture diffraction patterns at the full speed of LCLS. The CSPADs consisted of 64 tiles, each with dimensions of 192 pixels by 185 pixels. The CSPAD detector achieved the 120-Hz readout rate required to measure each x-ray pulse from LCLS [270].

Furthermore, for comparative analysis, the "old" reduced dataset deposited in the Protein Data Bank (PDB) was retrieved under the entries 4ET8, 4ET9, 4CAS and 4HWY. This dataset consisted only of crystal diffraction frames, as determined in 2011. The processing results of newly converted and previously deposited data were compared with the findings published after the original experiment [12, 264, 265]. Refer to Table 6.2 for comparing these results.

Table 6.2: Reprocessing previously published data in [12, 264, 265] with a new data processing pipeline

<b>Sample name</b>	<b>Resolution Å published/reprocessed</b>	<b><math>R_{\text{free}}/R_{\text{work}}</math> published/ reprocessed</b>
<b>Lyzo (40fs), pdb id: 4ET8</b>	1.9/1.5	(0.229/ 0.196)/ (0.195/0.172)
<b>catB, pdb id: 4HWY</b>	2.1/1.66	(0.213/0.182)/ (0.188/0.178)
<b>RC*, pdb id: 4CAS</b>	3.5/2.7	(0.329/0.295)/ (0.295/0.228)

Data conversion and reduction steps were performed using the Cheetah program [133]. Specific hit-finding parameters were fine-tuned for individual runs to optimise the data analysis and get the maximum from the



data. Defective pixels were carefully masked to ensure the accurate use of Bragg peaks as small as one pixel. Individual masks were created for each run to exclude artefacts such as ice scattering, ASIC edges, bad pixels, and shadows. The `CrystFEL` software [141] versions 0.9.1 and 0.10.1 were used to process the datasets using different parameters, and the results were subsequently compared. For detailed comparisons, see Tables D.1-D.17.

The `XGANDALF` indexing algorithm [147] demonstrated superior performance for SX data and exhibited robustness in indexing multiple lattices within a single diffraction pattern (utilizing the `--multi` option in `CrystFEL`). During the indexing process, the detector geometry was refined using the "detector-shift" script, and the `geoptimiser` program [26], provided by `CrystFEL`, was utilized. At the LCLS facility, the metrology of each of the four quadrants of the CSPAD was meticulously measured using an optical microscope. Following the successful indexing of the diffraction patterns, the intensities of the Bragg peaks were integrated for further analysis.

The `partialator` program, part of the `CrystFEL` suite, was used to achieve the final merging of intensities. Various partiality models were applied to examine their impact on different datasets. The results obtained by employing different modes of data merging can be found in Tables D.1-D.17.

The structure was determined using `phenix.refine` from Phenix/1.13 software [215]. All refinement procedures were performed consistently to ensure unbiased results by executing molecular replacement from the command line with the same set of parameters defined for each sample. The reprocessed data exhibited improved data quality, such as indexing rate, reconstructed electron densities, and overall statistical measures. A visual comparison of electron maps between the old and reprocessed data, as shown in Figures 6.6a to 6.6d, clearly demonstrates the remarkable enhancements achieved by recent pipelines. The electron densities appear notably smoother and more detailed in the reprocessed data.

Furthermore, the reprocessed data led to a significant discovery of the original goal of the 2011 experiment. The experiment aimed to address the question of whether extremely short pulses, such as 5-10 fs, are necessary to observe the "diffraction-before-destruction" effect or if slightly longer pulses of 40-60 fs duration, which are more intense, would suffice. The initial processing of the data did not provide a conclusive answer, as the quality of reconstruction for both datasets (one with a pulse duration of 5-10 femtoseconds and the other with 40 femtoseconds) was found to be quite similar [12]. However, reprocessing using a modern pipeline has highlighted the distinction between the measured structures.

Interestingly, the 40-femtosecond dataset exhibited slightly higher resolution than the 5- to 10-femtosecond dataset, likely due to the higher number of photons and indexable patterns. However, the reconstructed structure obtained from the 5-10 fs dataset showed better overall quality. This observation suggests that the structure measured using 5-10 femtosecond pulses experienced less damage than those measured with 40 femtosecond pulses.

In addition, we conducted a comparison between the reconstructed electron densities obtained from the lysozyme data collected at LCLS in 2011 (with a total of 60,000 patterns) [12] and the EuXFEL dataset acquired in 2018 (with a total of 1 million patterns) [271]. The results of this comparison are presented in Figure 6.6d. Interestingly, based on the findings shown in Figure 6.6d, it can be concluded that the quality of the data obtained from the reprocessed 2011 dataset exceeds that of the 2018 dataset.

Chapter 7 extensively discusses data evaluation after applying lossy compressions. However, it is important to briefly address the results obtained from the 40 fs lysozyme dataset after using non-hits rejection. In serial crystallography, it is common for not all detector frames to exhibit crystal diffraction. This is primarily due to the random intersection of the sample with the X-ray beam as it is passed across. Whether an individual

detector frame contains diffraction is determined by statistical factors. The hit rate, the fraction of measured frames containing useful crystal diffraction, is a critical parameter in serial crystallography experiments. It is influenced by factors such as sample concentration, flow rate, the relative size of the X-ray beam, and other factors. In practice, the hit rate during experiments is often relatively low, typically ranging between 0.1% and 10%. Consequently, it becomes evident that the volume of data can be significantly reduced by selectively retaining patterns with crystal diffraction while discarding data frames that do not exhibit any crystal diffraction.

The central question revolves around whether it is adequate to preserve only the frames that exhibit crystal diffraction or whether retaining the entire dataset would allow for the utilisation of more sophisticated data processing algorithms, leading to enhanced results in the future. It is important to note that serial crystallography (SX) data processing has significantly improved since its initial trials in 2010 [9]. These advancements have resulted in improved methodologies and techniques that offer the potential for more refined analyses and better insight from the preserved data. We have selected a lysozyme diffraction dataset measured with the X-ray pulse length 40 fs and processed it in several different ways:

1. just the results published in the original paper (<https://www.cxidb.org/id-17.html>);
2. the same patterns that were used originally but re-processed from the “raw” data;
3. modern hit-finding applied to all “raw” data further processed with the new algorithms.

In Table 6.3, one can see some results compared to the ones obtained in the original article.

Table 6.3: Different processing of the lysozyme dataset from 2011, The structure refinement of processed data was performed with `phenix.refine` (Phenix/1.13) with such parameters as `xray_data.high_resolution=1.6` and `xray_data.low_resolution=20` using 6FTR as the search model.

Name	Number of hits/ indexed patterns/ crystals	Resolution	$R_{\text{split}}$ / Completeness	$R_{\text{free}}$ / $R_{\text{work}}$
<b>Originally published results (4ET8)</b>	66k / 122k / 122k	1.9 Å	0.158 (n.a.)/ 98.3% (96.6%)	0.229/ 0.196
<b>Reprocessed the same events as original</b>	66k / 59k / 124k	1.51 Å	0.029 (0.14)/ 99.9% (97.87%)	0.195/ 0.172
<b>Fully reprocessed all frames from raw data</b>	109k / 71.5k / 137k	1.49 Å	0.029 (0.15)/ 100% (100%)	0.189/ 0.168

According to Table 6.3, the data processing performed just after the experiment [12] resulted in much lower quality than after any reprocessing performed now. It should be noted here that the resolution is 1.49 Å, which corresponds to the corner of the detector (see Fig. 6.2), so the actual gain in using the modern pipeline can be even higher, as seen from Table 6.2 (for one of the measured samples, the achievable resolution improved from 3.5 Å to 2.5 Å). However, determining the resolution cut-off point is primarily a subjective process. Nevertheless, all the quality metrics mentioned above are considered to reach a reasonable compromise. The reconstructed electron density and the structural model between the original and the re-processed data are shown in the Fig. 6.3

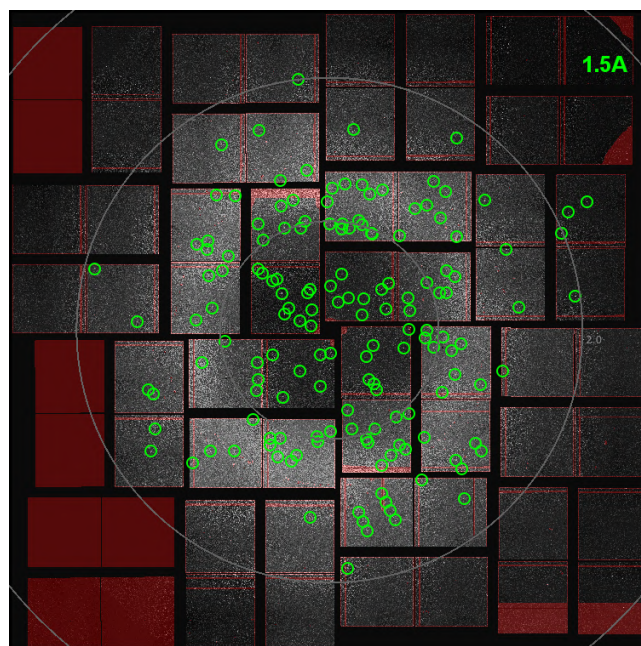


Figure 6.2: Diffraction pattern of lysozyme measured with the CSPAD detector. Red regions were masked and not considered during the data processing. Green circles indicate the found peaks. The resolution rings demonstrate the fact that there is almost no data measured below 1.5 Å resolution.

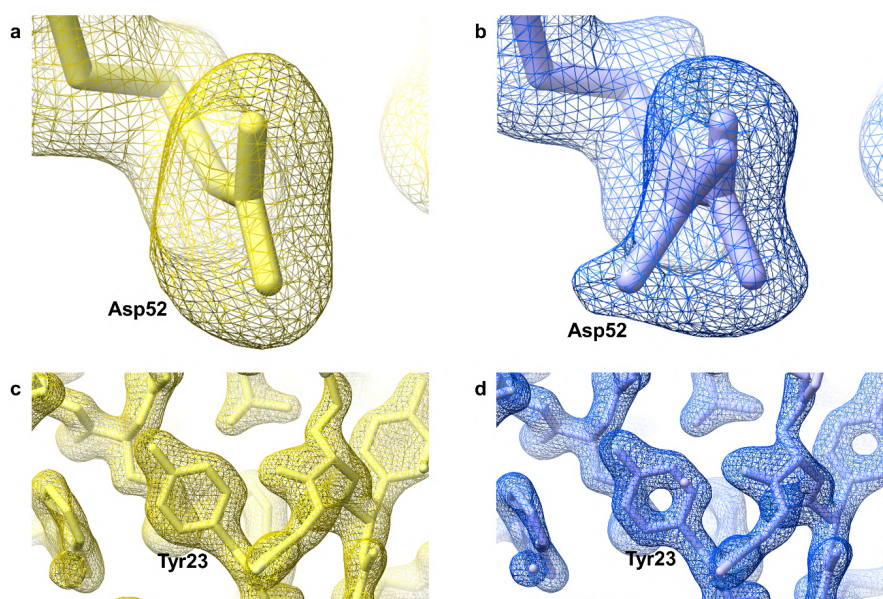


Figure 6.3: (a, b) Re-processed data resolves alternative conformer of the active site residue Asp52. Electron density maps (contour level  $\sigma = 0.8$ ) with models of residue Asp52 based on (a) originally published results (yellow) and (b) reprocessed data (blue). (c, d) The overall quality of the electron density is improved. Electron density maps (contour level  $\sigma = 1.5$ ) with models based on (c) originally published results (yellow) and (d) reprocessed data (blue).

and in Fig. 6.4. Fig. 6.3 also demonstrates the improvement in the reconstructed electron density between the original and reprocessed data. The example shown, Asp52, is an active site residue essential for the enzyme mechanism of lysozyme [272]. The re-processing data results in electron density maps that allow one to resolve an alternative conformation of Asp52, which is an active site residue essential for the enzyme mechanism of lysozyme [272], which allows a more accurate interpretation of biological function.

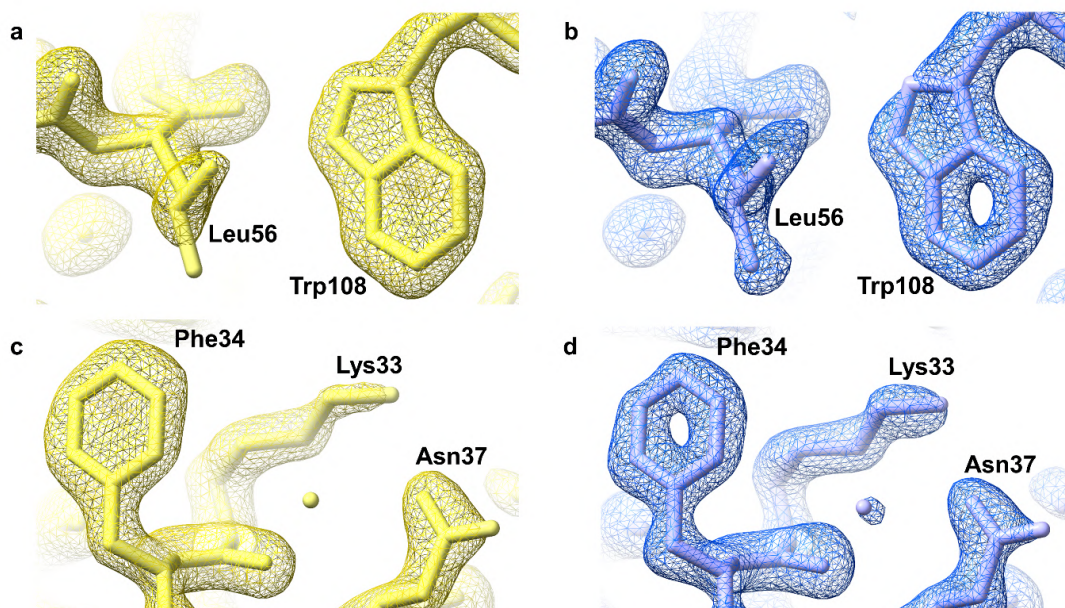


Figure 6.4: Additional examples of improved electron density comparing original (yellow, a/c) and re-processed data (blue, b/d). All maps are contoured at  $\sigma = 1.5$ . (a/b) Residues Leu56 is located in the core of the protein, and Trp108 is located within the active site cleft. (c/d) Residues Phe33, Lys34 and Asn37 are located at the protein's surface.

It is important to highlight that performing structure refinement using the `phenix.refine` suite from Phenix/1.13, with a resolution cut-off at 1.9 Å for the re-processed data, yielded  $R_{\text{free}}/R_{\text{work}}$  values of 0.206/0.173, which were inferior to the values obtained when reconstructing the data to a resolution of 1.49 Å ( $R_{\text{free}}/R_{\text{work}} = 0.189/0.168$ ). We conducted molecular replacement using the original mtz file deposited in 2012 to discern the primary factor behind the observed improvement. The outcome exhibited only a marginal enhancement compared to the published statistics ( $R_{\text{free}}/R_{\text{work}}=0.229/0.196$ ) compared to Phenix/1.20 ( $R_{\text{free}}/R_{\text{work}}=0.2109/0.1730$ ). Consequently, we deduced that raw image processing played a pivotal role in the improvement rather than advancements in the phasing software.

To fortify the conclusion, we conducted a comparative test. We refined the same dataset using two versions of Phenix (1.20 vs. 1.13) with default parameters without manual interventions. The refinements were executed with both 1.9 Å and 1.5 Å resolution cutoffs. The results in Table 6.4 demonstrate that the discrepancies between the same constraints are relatively similar. However, the resolution cutoff significantly impacted the outcomes.

Table 6.4: Comparison Phenix/1.20 versus Phenix/1.13 refinement results

Resolution range, Å	$R_{\text{free}}/R_{\text{work}}$ , Phenix/1.20	$R_{\text{free}}/R_{\text{work}}$ , Phenix/1.13
20 - 1.9	0.205/0.171	0.201/0.165
20 - 1.5	0.216/0.198	0.204/0.188

Upon closer examination of the aforementioned results, it becomes evident that the primary factors contributing to the observed improvement are the new algorithms for indexing (including indexing multiple crystals per pattern) and integration implemented in `CrystFEL` [25], better knowledge of the detector geometry [26], and a different strategy for the background subtraction. By effectively incorporating reflections from higher-resolution



data, the overall quality and accuracy of the structure have been notably enhanced. These advances collectively contribute to the improved quality and accuracy of the reprocessed data compared to the original publication.

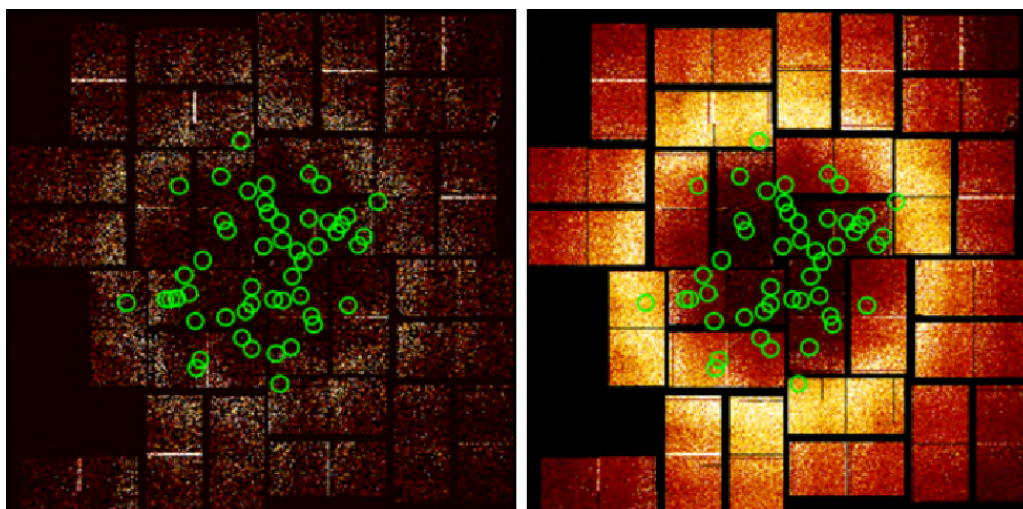


Figure 6.5: Comparison of X-ray diffraction patterns: original pipeline, deposited at CXI-DB (Left) vs modern pipeline (Right).

But from the point of view of data reduction, discussed in Section 7.4.2.1, the critical finding from such analysis is that improving the identification of frames containing crystal diffraction through the hit-finding step did not significantly enhance data quality. This is primarily due to the nature of the hit-finding process, which relies on a simple metric based on the presence of Bragg peaks. When the hit finding step was repeated on all raw data, most of the additional patterns detected were associated with weak diffraction signals, representing small crystals or crystals hit by the tail of the X-ray beam.

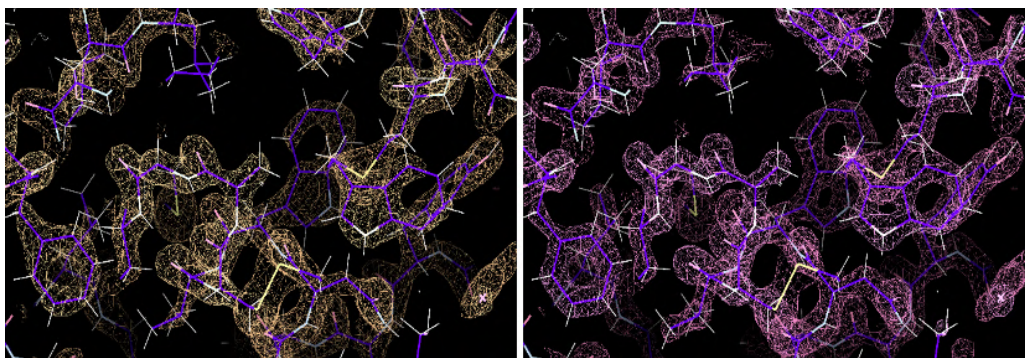
It is crucial to note that weak diffraction patterns provide less informative data compared to the strong patterns identified in the initial analysis, especially at higher resolutions. As a result, these weak patterns minimally improve the refined structure's quality.

Based on our analysis, we can conclude that retaining only frames exhibiting clear diffraction peaks for crystals with strong diffraction is a reasonable compromise. However, it is advisable to save these data in the "raw" format, enabling the application of future improvements in detector calibration if desired. This approach ensures that the data can benefit from advancements in calibration techniques, which may enhance the quality of the refined structure. However, dealing with weakly diffracting crystals poses a more complex challenge. There is always the possibility that weak diffraction frames may go undetected during the initial hit-finding process [273]. Moreover, additional information may be identified in the diffraction patterns after the experiment, such as diffuse scattering outside of Bragg peaks [274]. Whether to retain all weakly diffracting data depends on each facility's judgement, considering the potential benefits of such data for further analysis.

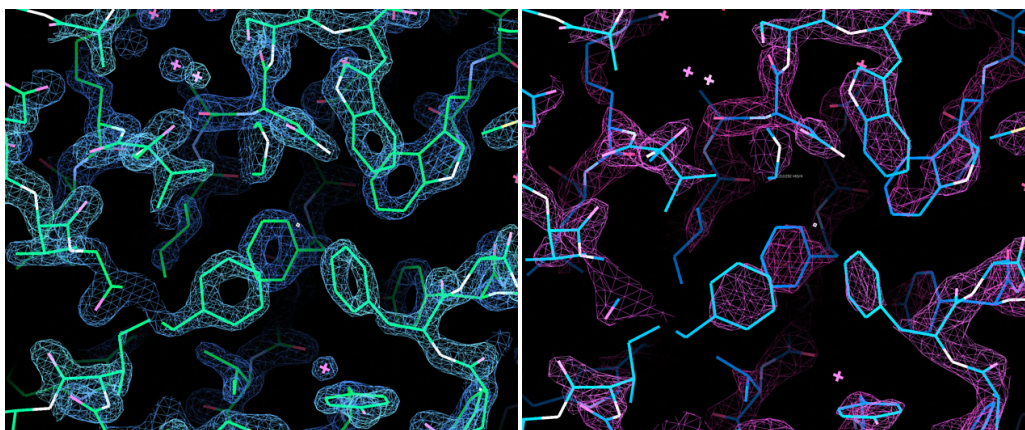
### 6.1.3 X-ray Diffraction Analysis of Hemoglobin Samples at LCLS MFX

Experimental data for hemoglobin samples from different batches were collected at the LCLS MFX experimental station at the SLAC National Accelerator Laboratory in Menlo Park, CA, USA, during LR17 beamtime [266] (example of diffraction pattern can be seen in Fig. 6.7). The experiments were conducted using a photon energy of 7.15 keV and pulse lengths of 10 femtoseconds and 3 femtoseconds, with a repetition rate of 120 Hz. Sample delivery was achieved using the DFFN nozzle on the RoadRunner system, specifically adapted for liquid jets

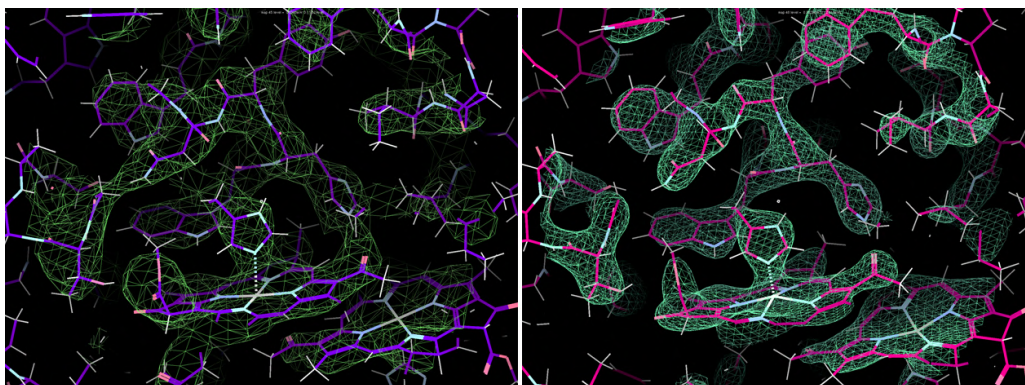
Figure 6.6: Reconstructed structures and corresponding electron densities



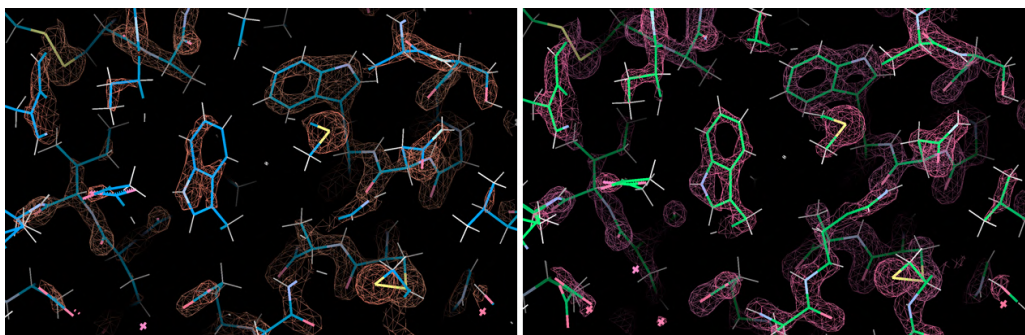
(a) Lysozyme 40 fs pulse duration from the dataset published in 2012 (Table 6.3, - left) and 2023 (Table 6.3, - right).



(b) Cathepsin B: left - after re-processing with modern pipeline; right - published results in [265]



(c) A photosynthetic reaction centre: left - published results in [264]; right - after re-processing with modern pipeline



(d) Lysozyme: left - published results of the experiment at EuXFEL in 2018 [271]; right - after re-processing datasets collected at LCLS in 2011 with modern pipeline



[116, 266]. To optimise the experimental conditions, the capillary beamstop and a helium enclosure were employed [43].

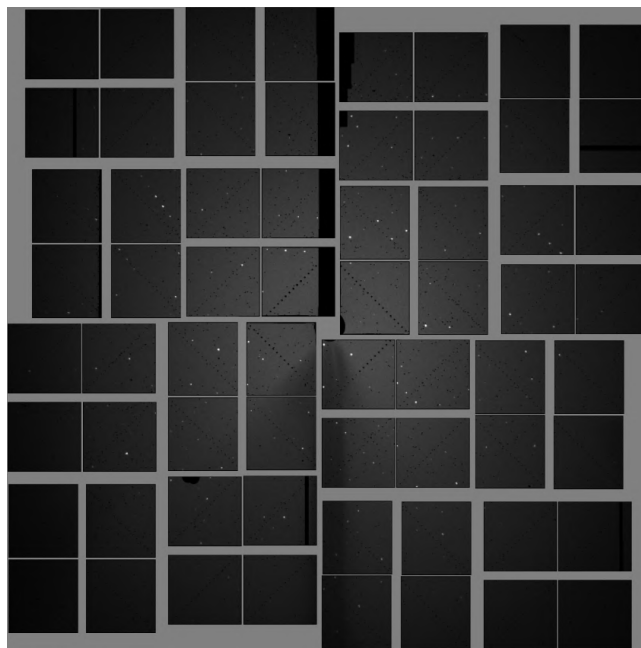


Figure 6.7: The image presented here shows a diffraction pattern obtained from the LR17 beamtime.

The data were collected using the Cornell-SLAC pixel-array detector (CS-PAD) [269]. The detector geometry was refined using `geoptimiser` to enhance the quality of data analysis [26]. Real-time monitoring was performed using the OnDA tool to assess the hit fraction and data quality on-the-fly [221]. The collected diffraction patterns were identified as individual "hits" and converted to HDF5 format using Cheetah software [133].

The primary objective of reprocessing this experiment was to evaluate the influence of pulse duration on the structural model of hemoglobin. Specifically, we aimed to investigate potential radiation damage, changes around the iron centres, and the overall effect on the protein structure. To ensure a proper comparison between the two pulse durations, the attenuator was set for the 10 femtoseconds pulse duration to achieve the same fluence as the 3 femtoseconds pulse duration. However, this was not successfully implemented, as shown in Figure 6.8, and we had to split the stream file for the 10 femtoseconds pulse duration into sub-streams based on higher and lower intensity.

For the structure comparison, we focused on a subset of samples from the same batch (HF2) that experienced the same crystallization conditions. These samples exhibited compatible statistics regarding the number of indexed crystals and intensity distribution for both pulse durations. Specifically, runs 187-195 corresponded to the 10 femtoseconds pulse duration, while runs 230-246 belonged to the 3 femtoseconds pulse duration. See some results at Fig. 6.9.

The data were processed with `CrystFEL` [141], using the `indexamajig` program (version 0.9.1 + 0e48c77b). The `peakfinder8` algorithm was used to detect Bragg peaks with the following parameters: `--min-snr=6 --threshold=200 --min-pix-count=1`.

The identified "hits" were indexed using `XGANDALF` [147] with the `--muti` and `--no-check-peaks` options. Figure 6.11 compares unit cell distribution for selected subsets with 10 femtoseconds and 3 femtoseconds pulse durations.

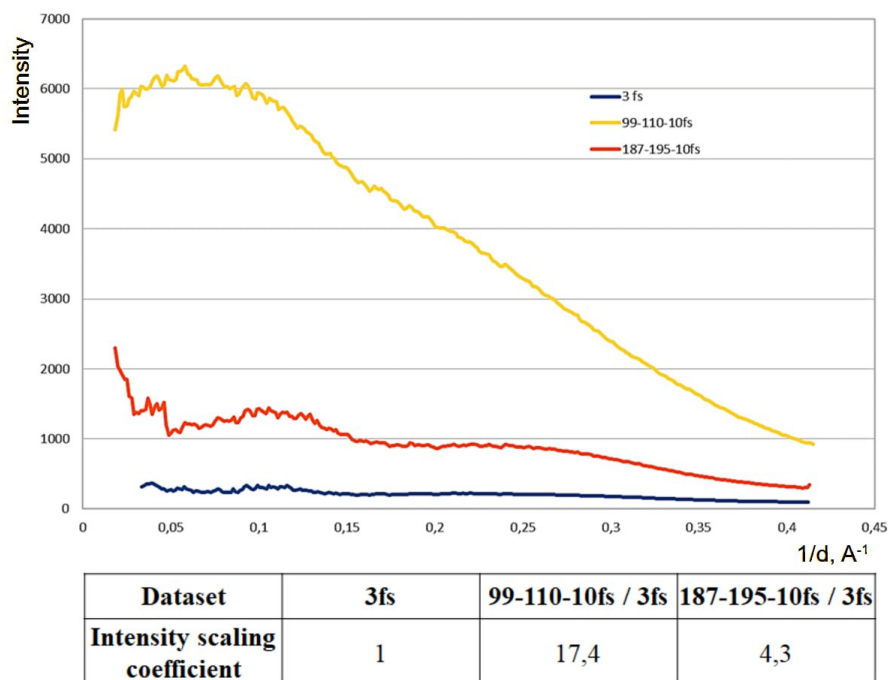


Figure 6.8: The intensity scaling coefficients for the three datasets are 1 for the 3fs dataset, 4.3 for the low-intensity 10fs dataset, and 17.4 for the high-intensity 10fs dataset. These coefficients represent the differences in scattered intensity specifically for the Bragg peaks observed in each dataset.

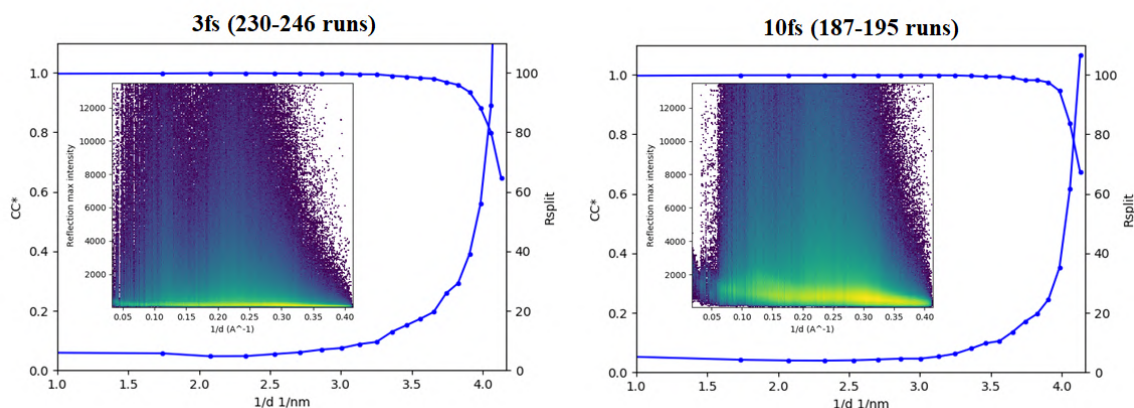


Figure 6.9: The figure consists of two plots of such quality metrics  $CC^*/R_{split}$  versus  $1/d$ , where  $d$  represents the resolution in nanometers. These plots specifically focus on subsets of data collected with pulse duration of 10fs and 3fs. The intensity distributions of these subsets are comparable, allowing for a direct comparison of the data quality at different resolutions.

To scale and merge the data into the point group  $mmm(222)$ , we utilized the `partialator` suite (version `0.9.1 + 0e48c77b`) of `CrystFEL`. Three iterations were performed with the parameters `-push-res=1.0` and `-model=xsphere`. The quality assessment of the data involved the calculation of figures of merit using `CrystFEL`'s `compare_hkl` (including  $R_{split}$ ,  $CC_{1/2}$ , and  $CC^*$ ) and `check_hkl` (including  $SNR$ , multiplicity and completeness). MTZ files for crystallographic data processing were generated from the `CrystFEL hkl` files using `f2mtz` from CCP4 [214].

The results are summarised in supplementary materials in Table D.18 and visualised in Figure 6.9 and Figure 6.10. Table D.18 presents the overall statistics for all merged data for each pulse duration, with the results for selected datasets shown in parentheses. Molecular replacement was performed using `Phaser` [275] with



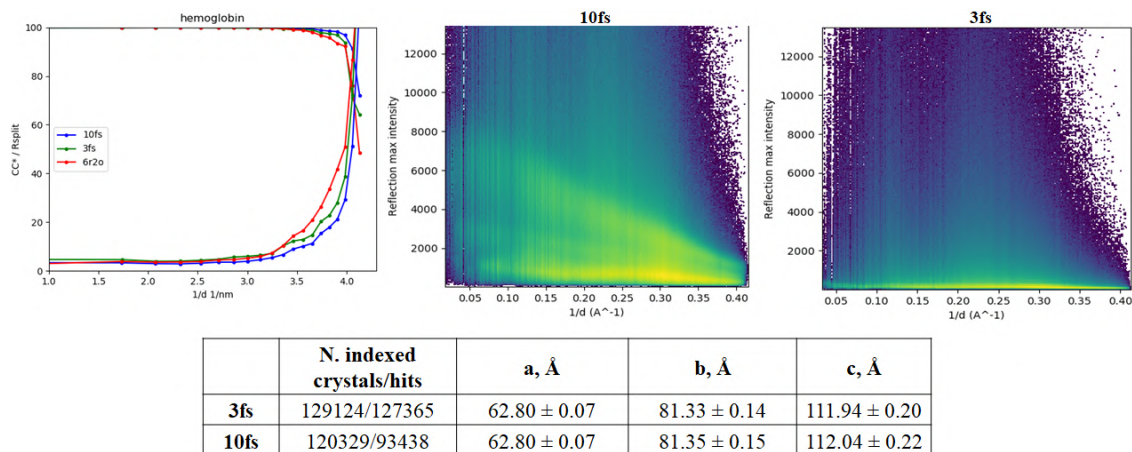


Figure 6.10: The figure illustrates various comparisons and data statistics of reprocessing experimental data. The top-right section compares the  $CC^*/R_{split}$  plots of the reprocessed data using pulse duration of 3fs and 10fs, along with the published data corresponding to the PDB ID 6R2O. The top-left portion of the figure presents peakograms for the complete dataset measured with 10fs and 3fs pulse duration, visually representing the distribution of diffraction peaks. The table below provides important statistics, including the number of indexed crystals and the count of hits with determined unit cell parameters. These statistics offer valuable insights into successfully identifying and characterising crystals within the dataset.

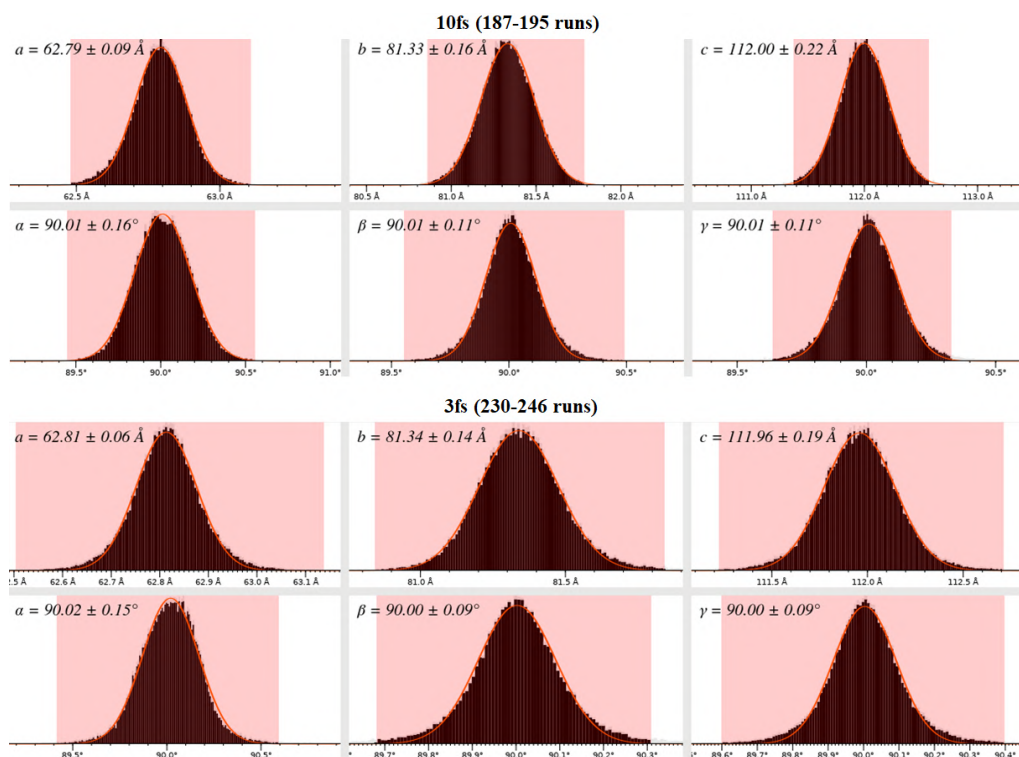


Figure 6.11: The figure displays the distribution of unit cell parameters for subsets of the hemoglobin dataset collected with pulse duration of 10 fs and 3 fs. The unit cell parameters provide essential information about the size and symmetry of the crystals.

PDB-ID: 6R2O as the search model. The resulting structures were iteratively refined using `phenix.refine` [215] and `Coot` [276]. The  $2F_o - F_c$  electron density maps for different chains of hemoglobin are presented in Fig. 6.12. Additionally, the overall structures for the 3 femtoseconds and 10 femtoseconds pulse durations can

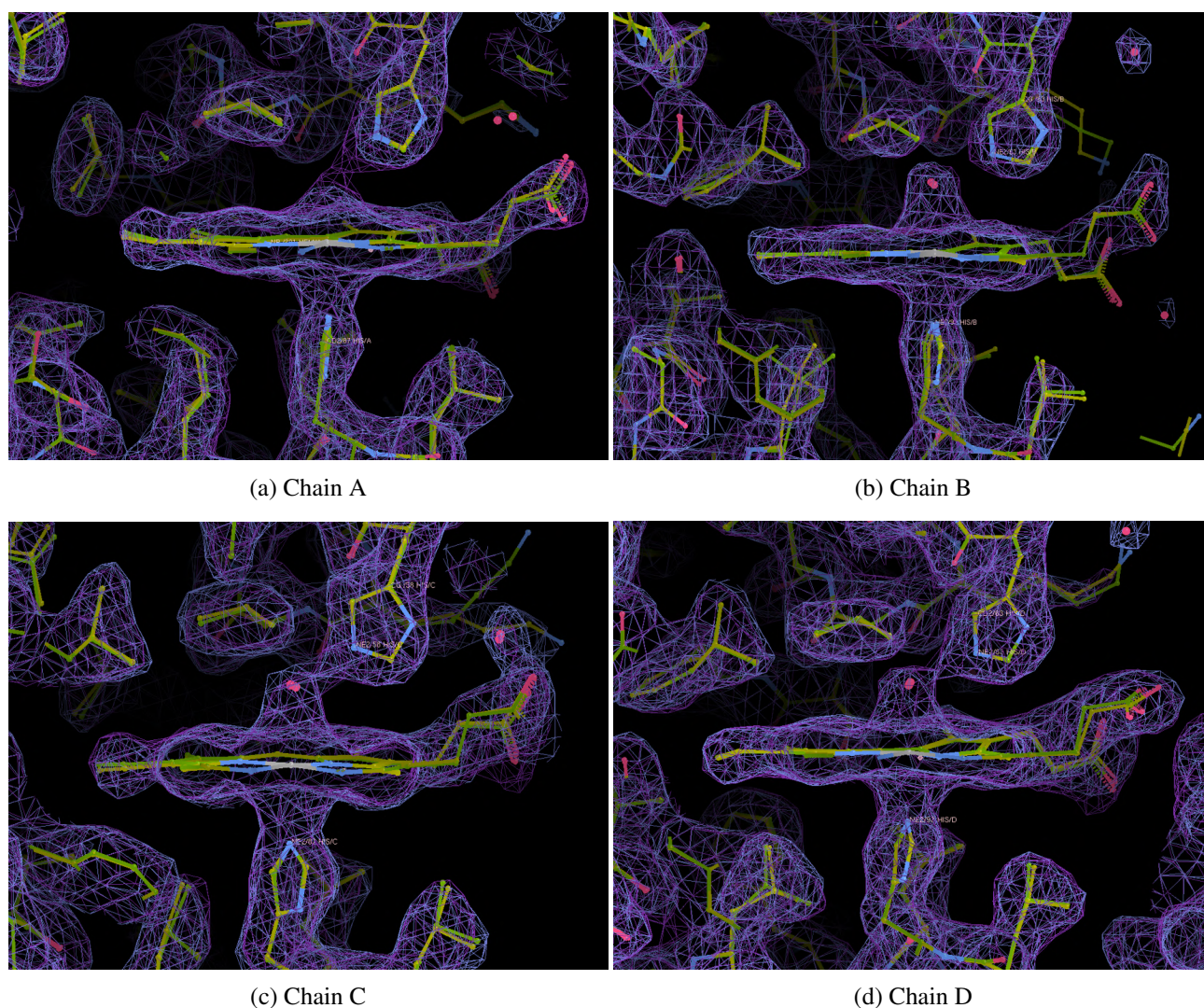


Figure 6.12:  $2F_o - F_c$  electron maps (blue 3 fs, violet 10 fs) counteracted at  $1.1 \sigma$  for different chains

be found in Figure 6.13.

Based on visual inspection, elucidating the specific reasons for the observed changes around the iron centre presents a challenging task. Moreover, the overall structure remains relatively unaffected, as illustrated in Figure 6.13. Interestingly, the changes discovered in the structure around the iron centre align with findings from a prior study on radiation-induced effects of Fe [277]. In that study, the authors investigated the time-resolved femtosecond evolution of iron's K-shell X-ray emission spectra under high-intensity illumination of X-rays in a micron-sized focused hard X-ray free electron laser (XFEL) beam. They reported rapid spectral energy shifts and broadening within the first 10 fs of X-ray illumination, attributed to the rapid evolution of high-density photo-electron-mediated secondary collisional ionisation processes following the absorption of the incident XFEL radiation.

## 6.2 Conclusion

To obtain reliable and accurate structural information, data quality is paramount in serial femtosecond crystallography (SFX) experiments. Data quality assessment involves using specific metrics, which are discussed in detail in Section 5.1, to provide valuable information on the reliability and usability of collected data. Key metrics



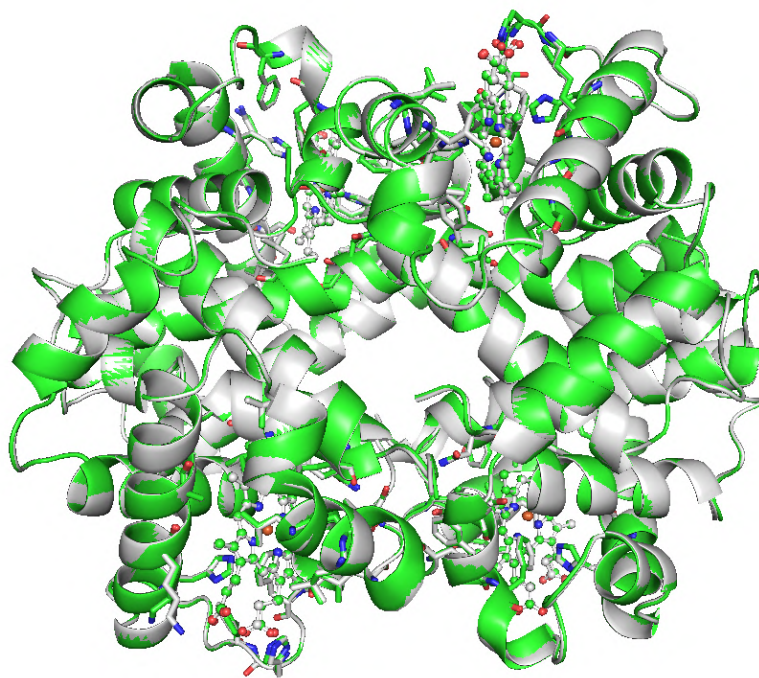


Figure 6.13: Overall structure of hemoglobin with 3 (grey) and 10 fs (green).

such as the signal-to-noise ratio ( $SNR$ ),  $R_{split}$ , and  $CC^*$  are crucial in evaluating the data's overall quality and reliability, indicating the measurements' precision and reproducibility. Additionally, the  $R_{free}/R_{work}$  values indicate the agreement between the observed data and the model, assessing the accuracy of the obtained structural information. A detectable anomalous signal in the data allows an *ab initio* structure to be reconstructed, which can additionally be used as an indicator for a thorough assessment of data quality. Analysis of the anomalous signal can be critical in evaluating the reliability and applicability of new algorithms and data processing strategies.

However, it is essential to emphasise that these data quality metrics should be used with a visual inspection of the reconstructed electron density. Visual inspection allows researchers to assess the overall quality and interpretability of the electron density maps, ensuring that the structural information obtained is reliable and meaningful.

In this chapter, we have presented different cases of re-processing previously collected data from various facilities. Comparison between old and recent data treatments highlights the importance of evaluating each step and refining parameters for improved data processing. Specifically, we reprocessed the data utilising the anomalous signal from bromine for *ab initio* structure reconstruction. This case highlights the improvement in the data processing pipeline and our enhanced understanding of preprocessing steps, such as masking unreliable detector regions, optimising detector geometry, peak finding parameters and due to better intensity integration process, namely partiality model and scaling in each individual diffraction pattern. The datasets collected in 2011 at LCLS showcased a drastic improvement in data quality metrics compared to the published results after reprocessing with new, well-optimised parameters. In the case of hemoglobin data, we achieved superior results compared to the published ones and successfully reconstructed the structure for another pulse duration [266]. Reprocessing old data emphasises the necessity to preserve primary raw data to evaluate new algorithms and obtain better results.

This chapter also focuses on the critical role of data quality assessment and the benefits of reprocessing

previously collected data. Researchers can enhance their structural data reliability, utility, and interpretability by employing robust data quality metrics, optimising processing parameters, and preserving primary raw data.

---

# Compression and data reduction in serial crystallography

Serial crystallography (SX) is now an established technique for protein structure determination with particular application for the study of small or radiation-sensitive crystals and the study of fast or irreversible protein dynamics. Newly developed multi-megapixel X-ray area detectors capable of recording at a frame rate of more than 1000 images per second have been highly beneficial but, at the same time, significantly increased the volume of collected data. Today, up to 2 PB of data per experiment could be easily obtained under efficient operating conditions. The cumulative cost of storing data from many experiments creates a strong motivation for developing strategies that reduce the volume of data retained on disk without affecting the quality of science outcomes. Lossless data compression methods, by definition, do not reduce the information content but usually fail to achieve a high compression ratio when applied to experimental data that contain any noise. On the other hand, lossy compression methods can significantly reduce the data volume; however, careful evaluation of the resulting effects on data quality and scientific conclusions is required since lossy compression, by definition, discards information. Indeed, the use of appropriate data quality metrics is important to answer the question of whether lossy compression adversely affects the data (which is discussed in detail in Section 5.1). In this chapter, we evaluate different approaches for lossless and lossy compression applied to SX data and pay attention to the metrics appropriate for SX data quality assessment.

## 7.1 Introduction

In recent years, serial crystallography (SX) has established itself as a technique for the determination of protein structures with a particular application in the study of small or radiation-sensitive crystals and for the study of fast or irreversible protein dynamics [64, 65, 278, 279]. This has been enabled by the development of new generation X-ray sources such as X-ray Free Electron Lasers (FELs) and 4<sup>th</sup> generation synchrotrons, which produce very bright and coherent X-ray beams, combined with improvements in focusing optics, which increase flux density at the sample and thereby decrease the exposure time required to obtain a measurable signal. Equally important has been a revolution in detector technology, enabling the development of multi-megapixel detectors capable of accurately measuring weak X-ray images at frame rates approaching or exceeding 1 kHz.

Modern detectors such as Eiger [280], JUNGFRÄU [281–283], Lambda [284], ePix [285], AGIPD [286], LPD [287, 288], or DSSC [289–292] have the capability to capture thousands of images per second. The development of these detectors, coupled with the aforementioned high-intensity photon sources, enables the

collection of valuable images at a kilohertz frame rate [18]. Combined with the tiling of detector modules to increase the number of pixels (up to 16 million pixels for Eiger or JUNGFRÄU), this leads to very high overall data rates [283]. For example, the Eiger2 XE 16M detector, developed by Dectris for synchrotron facilities, generates 16-megapixel images at a frame rate of 400 images per second. When uncompressed, one gets a staggering data rate of 13.5 GB/s. Considering the continuous operation typical in serial crystallography (SX), this amounts to approximately 1 PB/day of data. In the case of XFEL facilities, each of the two JUNGFRÄU 16M detectors installed at Switzerland's X-ray free-electron laser at the Paul Scherrer Institute (SwissFEL) can operate at a remarkable 2 kHz, generating data rates of up to 60 GB/s. This equates to a potential accumulation of close to 4 PB/day by each detector. However, SwissFEL cannot be operated at such an extreme speed, which alleviates the burden on data storage systems. New detectors like ePixHR [293] at LCLS-II and AGIPD 4M at European XFEL are expected to generate data at similar rates. While it is technically feasible to save such data streams, the cumulative cost of doing so imposes a substantial burden on operational budgets. As a result, there is a compelling motivation to explore data reduction strategies that preserve data while ensuring the quality of scientific outcomes remains unaffected.

SX experiments routinely use the maximum frame rate that the detector can sustain for long periods of time. Therefore, the need to store, retain, and process multiple petabytes of data per day is readily apparent, as is the pressing need to reduce data volumes without compromising science output. Although it is technically possible to save raw data at full quality, the cost of storing all data from multiple experiments at a given instrument rapidly becomes prohibitive. Therefore, there is a strong motivation to develop compression and data reduction strategies that allow raw data to be retained without affecting the quality of scientific results. Data reduction is a comprehensive concept encompassing a range of techniques designed to decrease the size or complexity of a dataset while retaining essential information. These methods include data compression, summation, filtering, feature selection, and dimensionality reduction. Often, data reduction strategies are divided into two big subgroups, such as lossless and lossy algorithms.

To efficiently apply any data reduction method, it is crucial to understand the data being processed. A typical diffraction pattern in SX comprises bright and sharply defined Bragg peaks, which originate from the studied crystals, and a relatively smooth background that arises from various factors, such as the sample delivery medium, disordered structure and solvent within the crystal, and parasitic scattering from the beamline. The intensities observed in various regions of the diffraction pattern can vary significantly, often differing by several orders of magnitude. Additionally, the useful signal represented by the Bragg peaks at high scattering angles may be comparable to the background noise. These features of diffraction patterns in SX affect the applicability of different compression algorithms.

Lossless compression techniques are frequently employed to reduce the size of scientific data, particularly when the signals recorded in each pixel of the detector are mostly zero or constant. However, constant signals are rarely observed in typical SX diffraction patterns. As a consequence, the effectiveness of lossless compression schemes is significantly diminished in this particular case. On the other hand, applying standard image compression techniques directly to SX data is challenging due to the significant signal variation observed in neighbouring pixels, particularly near Bragg peaks. Thus, to achieve efficient data compression in SX, alternative compression approaches that are specifically designed to handle the sparse nature, high dynamic range, high noise level, and sharp intensity changes observed in diffraction patterns are needed.

The main focus of this chapter is to evaluate different lossless and lossy data compression methods and determine the appropriate metrics, described in Section 5.1, for evaluating the impact of lossy compression on the final SX data quality. The imperative outcome of this part of the work is that an effective strategy for

data reduction in the case of strongly diffracting crystals is to selectively save only the images that exhibit a substantial number of Bragg peaks. For sure, all metadata required for further processing the reduced dataset has to be saved [259]. As described in Chapter 6 in Section 6.1.2, this approach demonstrates remarkable efficacy, even when reprocessing previously collected data using new algorithms. Furthermore, the analysis shows that a non-linear reduction in the precision of the measured diffraction pattern intensities is the second most effective strategy for achieving lossy compression. Moreover, it has been shown that binning, which effectively enlarges the pixel size, is highly effective, especially when applied to diffraction data obtained from crystals with small unit cells and measured using multiple-megapixel detectors.

The presented in this chapter research underscores the importance of considering the potential risks associated with particular lossy data reduction schemes. Specifically, strategies that involve reducing the number of collected patterns or selectively saving determined Bragg peaks may lead to notable deterioration in the data quality. Therefore, it is essential to carefully weigh the trade-offs between data reduction and preserving crucial scientific information when implementing these schemes.

## 7.2 Review of existing data reduction methods in science

Various compression methods have been explored in recent years, as storing vast amounts of data has become a problematic topic in many areas. Therefore, researchers are looking for suitable data compression and data reduction schemes. Detailed reviews of mostly lossless methods applied to different scientific datasets can be found in these papers [294, 295]. Compression schemes typically reduce data volume by exploiting symmetries or redundancies in the data. Thus, the best type of compression to use for a given application depends on the nature of the data being compressed and the information deemed important to retain. Identifying a universal lossy compression scheme is thus difficult since the choice depends on what information must be retained and what can be discarded. Here, the key focus is on data compression methods applicable to or available in the context of serial crystallography.

In [294], the authors evaluated existing compression and clipping algorithms on the oceanographic and meteorological data sets and also introduced the digit rounding algorithm similar to the SZ error-controlled quantisation approach [296]. In [295], the authors focused on the underlying theories and the application of mechanisms to reduce data volumes in high-performance computing (high-performance computing (HPC)) systems and discussed related hardware acceleration. The main objective was to accommodate the existing compression algorithms to the growing volume of accumulated data. In computed X-ray tomography, an azimuthal regrouping of input images followed by applying CBF compression was used in [297] as a data reduction pipeline, thereby exploiting the known rotational properties of a tomographic data set.

As mentioned previously, the growing data rates in various scientific fields necessitate adopting lossy data reduction methods, as lossless compression alone is insufficient. One such field is electron microscopy, extensively employed for studying protein structures. In a recent study [298], researchers proposed a data reduction and compression scheme specifically designed for electron microscopy. Their approach involved storing solely the information pertaining to identified electron puddles from the raw data, which was subsequently subjected to further compression.

For fluorescence microscopy datasets, real-time compression was introduced during data collection [299]. The compression scheme included lossless and noise-dependent lossy modes. The lossless scheme was extended to lossy compression by adding a variance stabilisation step before the prediction and quantising the prediction residuals to integers before Huffman coding.

Ptychographic diffraction imaging is another thriving storage-consuming method, and several different methods were proposed to reduce the data volume. In [300], the authors proposed two novel compression strategies. The first method compressed data by means of a truncated singular value decomposition (singular value decomposition (SVD)), and another approach reduced diffraction data by using constrained pixel sum compression (constrained pixel sum compression (CPSC)) - summation over a given region. Another interesting data compression technique for ptychography was presented in the work [301]: the authors implemented an online lossy compression algorithm that stored the measured intensities after the quantisation step and then saved them as 8-bit unsigned integer values. In the paper [302], a lossy compression scheme using adaptive coding quantisation with additional data quality estimation was proposed.

In medical X-ray imaging, the JPEG compression format is still widely used [303, 304]. But even the appearance of artefacts after applying lossy compression will not affect the final diagnosis because of the preservation of the common features of such images. In [305] the influence of JPEG 2000 [306] (<https://jpeg.org/jpeg2000/index.html>) and JPEG XR (<https://jpeg.org/jpegxr/index.html>) compression methods of the original X-ray projections on the final tomographic reconstructions was investigated. They could achieve a compression ratio (CR) of about a factor of 6-8.

In (<https://github.com/FilipeMaia/h5h264>), the author developed the plugin compatible with HDF5 files to perform H.264 lossless compression through ffmpeg (<http://ffmpeg.org>) and tested it on Flash X-ray Imaging diffraction data from LCLS. It compressed the data by a CR factor 3 compared to 1.5 CR after gzip.

In [283], the authors presented different compression schemes performed on JUNGFRU detector images. They demonstrated the evaluation of the impact on quality metrics of the compressed data collected at the Swiss Light Source X06SA beamline with the JUNGFRU 4M using the rotation method. They tested a rounding algorithm, conversion up to several photons, with various combinations of such lossless compression as Bitshuffle filter [307], LZ4 (<https://github.com/lz4/lz4>), Zstd (<https://github.com/facebook/zstd>) and Gzip (<https://www.rfc-editor.org/rfc/rfc1952.html>, <https://www.gnu.org/software/gzip/>). The authors also studied the SZ algorithm [308], where they emphasised the drawback of this lossy compression, which affects weakest reflections at high resolution. In this paper, they also mentioned the analysis of lossy compression for X-ray protein diffraction images collected using CCD detectors by J. Holton ([https://bl831.als.lbl.gov/~lijjamesh/lossy\\_compression/](https://bl831.als.lbl.gov/~lijjamesh/lossy_compression/)), where the idea was to preserve the features of the image by using lossless compression and to compress the background in a lossy way.

Exhaustively testing all possible compression schemes is impractical. Here, the following compression approaches are considered, which have been discussed as candidate compression schemes to reduce the volume of SX data stored on disk:

1. Lossless compression algorithms commonly available for the HDF5 library including: gzip, bzip2, zstd, lz4 with and without bit shuffle, and different combinations of blosc (lz4, lz4hc, blosclz, snappy, zlib, zstd).
2. Data reduction schemes including:
  - measuring less data (i.e., a reduced number of frames);
  - non-hit rejection (saving only diffraction patterns with detectable crystal diffraction);
  - saving only found peaks in diffraction patterns.



3. Lossy compression schemes including:

- binning (effectively increasing the pixel size and reducing the number of pixels at the detector);
- quantization (saving only several discrete levels of intensity with different choices of levels);
- quantization with data rearrangement (saving the value of each pixel rearranged as a single byte).

The effect of each compression type on the resultant data and the quality of the structural model in the context of the SX test datasets is studied.

## 7.3 Selection of test datasets

An ideal protein crystal diffraction pattern measured using a noiseless detector would be very sparse, consisting of bright Bragg peaks with zero background. Such a diffraction pattern is easily compressible by most existing lossless compression algorithms. By contrast, in real crystallography experiments the background recorded in each diffraction pattern is quite high and is often comparable to the strength of the measured Bragg peaks. Statistical noise in the background leads to significant intensity differences between neighbouring pixels. Furthermore, the integrating detectors, described in Section 3.5, used at XFELs do not count incoming photons but rather accumulate the charge deposited in a single femtosecond-duration exposure. Accumulated charge including intrinsic electronic noise sources is not directly converted to individual photon counts but estimated after the detector is read out. Experience shows that compression of experimental SX diffraction data rarely can reach the compression factor better than 5 using lossless algorithms.

To capture these challenges, four representative SX data sets have been selected, which cover a range of detector technologies, photon sources, and sample delivery methods described in Chapter 3 and Chapter 4. These test data sets compare:

1. Counting detectors with low background (Eiger 16M at Petra III with tape drive delivery system [106, 107]) - measured samples: lysozyme, lactamase, ferritin, MPro (unpublished);
2. Integrating detectors with photon conversion and high background (JUNGFRAU 16M at SwissFEL with Lipidic Cubic Phase (LCP) jet [58]) - measured samples: thaumatin [249];
3. Integrating detectors without photon conversion with the data stored as integers (CS-PAD at LCLS with liquid jet [53, 96]) - measured sample: lysozyme [12];
4. Integrating detectors without photon conversion with the data stored as floating point numbers (AGIPD at EuXFEL with liquid jet [53, 96]) – measured samples: granulovirus, lysozyme (unpublished).

These test cases cover a representative sample of current protein crystallography datasets, including cryo-crystallography, where the background is high, and a counting detector is usually used (similar to the 2<sup>nd</sup> dataset). In the discussion, comparisons between datasets have been selected that are considered to best illustrate the challenges posed for different algorithms, drawing upon practical experience working with various compression schemes and datasets. This avoids the combinatorial explosion of testing all algorithms against all datasets, enabling to focus on the key issues rather than presenting large tables of exhaustive comparisons.

## 7.4 Applying different lossless and lossy compressions

As was mentioned before, serial crystallography (SX) experiments require a lot of diffraction snapshots to get 3D structural information of the studied protein because carrying on such experiments means that there will be a probability of not hitting the crystal or partially hitting it, thus, in the end, the full dataset will contain patterns with non-hits. Moreover, combined with new X-ray detectors and performing experiments at modern synchrotrons and free electron lasers (FELs), SX invariably leads to high storage consumption. For instance, during the experiment with the Eiger2 XE 16M detector, in the worst case, we can collect up to 756 TB/24 hours. One of the possible ways to overcome this problem is to delete the raw data after a short period of time and keep only the averaged intensity of Bragg peaks. Such an approach can be justified only in normal protein crystallography (MX) because of the well-optimised pipeline. Unfortunately, for FELs and for more complicated synchrotron SX experiments, it is not possible yet – reprocessing of raw data can greatly improve the result due to the advances in the processing pipeline and better detector calibration. Therefore, we are keen on reducing data by using lossless or lossy compression approaches applied to raw data to save more storage. Lossless compression is always preferable because it does not influence data quality, but lossy compression can give us a higher compression rate. Here, we are going to talk about compression techniques.

### 7.4.1 Existing lossless compressions and its evaluation

Lossless compression approaches are commonly used for compressing scientific data. By definition no information is lost and the original data can be restored verbatim. The only question is therefore the achievable compression rate (CR) and speed, which are both highly dependent on the statistical properties of the data to be compressed. Lossless compressions vary according to the algorithms they use. Thus, we can distinguish lossless compressions as entropy-based, dictionary-based and other algorithms that use several methods such as, for instance, run-length encoding, Huffman coding, move-to-front technique, shuffle, etc. All compression filters treat various data types differently: some can work only with integer data types, while others can be adapted to compress a floating point.

We observe that lossless compression schemes vary significantly in effectiveness depending on the experimental SX data to which they are applied. From our sample data sets, two extreme cases for lossless compression were the Eiger 16M detector and the AGIPD. The Eiger 16M is a counting detector that registers zero counts without incident photons and integer values corresponding to the number of photons incident on a pixel. On the other hand, the AGIPD is an integrating detector with 3 different gain stages for which calibrated data is stored in a floating-point format, and the value in a pixel might be even negative due to the subtraction of non-constant dark signal. For integer data (Eiger 16M, CS-PAD, or AGIPD rounded to integer values) compression ratios of higher than 4 can be achieved using zstd or bzip2, and in some cases can be higher than 10. Commonly available in HDF5 gzip compression (with compression level 6) demonstrates quite good result. Conversely, the achievable compression ratio for floating point type data (AGIPD detector) reaches only 1.3 with gzip level 6.

A complete table with the results of the lossless compression algorithms tested against our reference data sets can be found in the supplementary materials (see Table E.1). We observe that conversion to photons (integers) is an important factor in determining compression efficiency even if this conversion is itself a form of lossy compression, as expected.

All tests mentioned in Table E.1 were performed for offline data reduction. Another important parameter for lossless compression is its speed, which is especially critical for online data processing and real-time compression. Tests performed using blocks of 1000 frames from our two extreme cases above, Eiger 16M and AGIPD 1M

detectors, indicate that the best compromise of the compression/decompression speed versus compression ratio was observed for such algorithms as blosc, zstd and, bit-shuffle. For more information see Fig. 7.1 and Fig. 7.2.

## 7.4.2 Lossy compression

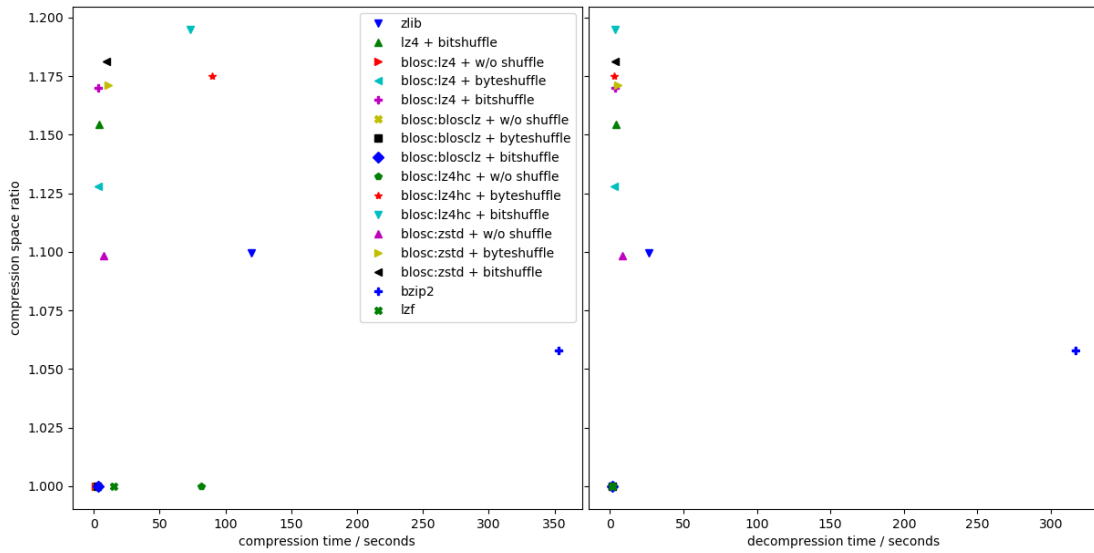
### 7.4.2.1 Non-hits rejection

In serial crystallography, it is common that not all detector frames to contain crystal diffraction. This is due to the sample being passed across the X-ray beam and intersecting with the beam at random. Whether any individual detector frame contains diffraction is a matter of chance. The hit rate, namely the fraction of measured frames containing useful crystal diffraction, is one of the important characteristics of an SX experiment and is related to the sample concentration, flow rate, and relative size of the X-ray beam, among other factors. In practice the hit rate during experiments is frequently rather low - on the order of 0.1 – 10%. This observation leads to the obvious conclusion that the volume of data can be reduced by discarding data frames without any crystal diffraction. However, we have to understand if it is enough to store only hits or if we should store all the data, hoping that more advanced data processing algorithms could help to get better results in the future. Section 6.1.2 addresses the answer to this question. There it was concluded that for strongly diffracting crystals it is indeed a good compromise to save only frames with clear diffraction peaks but to save this data in the “raw” data format so that improvements in detector calibration can be applied from the raw data later. For the case of weakly diffracting crystals the situation is more complicated because there is always the potential that weakly diffracting frames may not be found [273] and additional information may be identified in the diffraction patterns after the experiment, such as diffuse scattering outside of Bragg peaks [274]. The decision as to whether this justifies the retention of all weakly diffracting data is one that each experiment team will have to make itself.

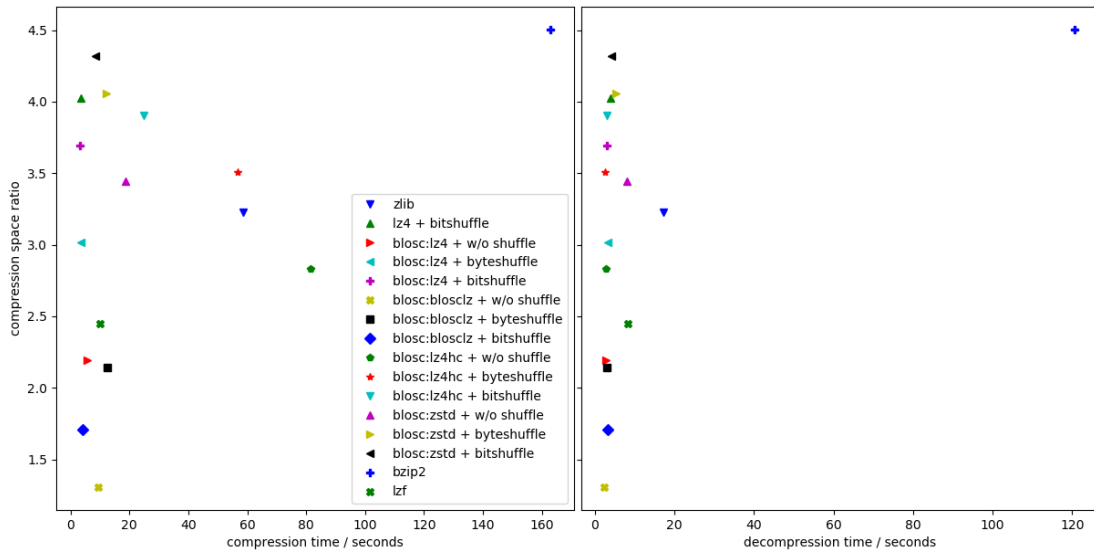
### 7.4.2.2 Measuring less data

Measuring more data frames in an SX experiment generally leads to higher-quality results due to averaging a greater number of observations. But an associated question during any experiment is: When has sufficient data been collected to answer the scientific question? Measuring only enough data to answer a scientific question reduces experiment time and minimises the amount of data collected to only the amount needed. Indeed, one of the common questions during SX beamtime is when to stop data collection. In order to address this question, it is necessary to assess the impact of reducing the number of measured frames [18, 106].

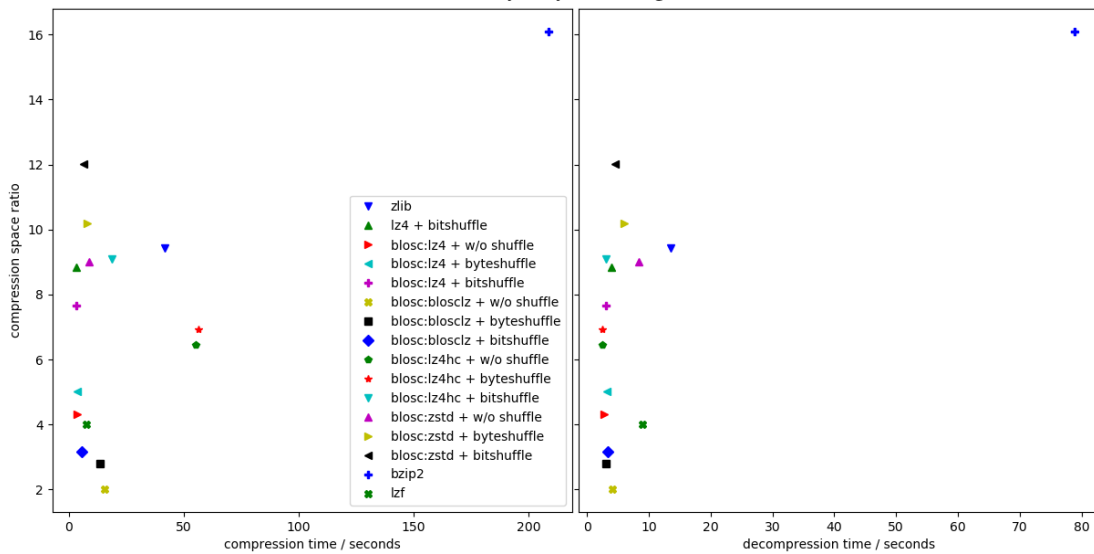
The effect of measuring fewer data frames can readily be checked for any single dataset by integrating progressively smaller data subsets. Fig. 7.3 shows the quality metrics  $CC^*$  and  $R_{split}$  versus resolution as well as  $R_{free}/R_{work}$  metrics for the different subsets of lactamase data measured during one of the tape-drive [105, 106, 108] SX experiments at the P11 beamline of PETRA III [106, 108]. The initial dataset consists of 200,000 diffraction patterns and is processed as smaller subsets equivalent to less measurement time (for more details, see Table 7.1). This dataset was collected as a single 25 minutes acquisition with an Eiger2 X 16M detector operated at 133 Hz.



(a) AGIPD, lysozyme, floating-point data

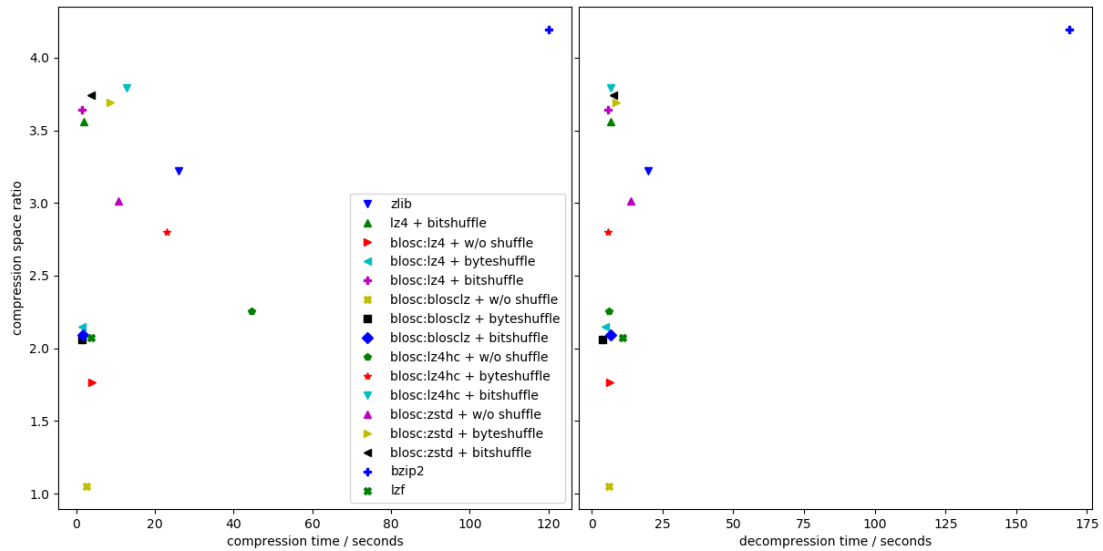


(b) AGIPD, lysozyme, integer data

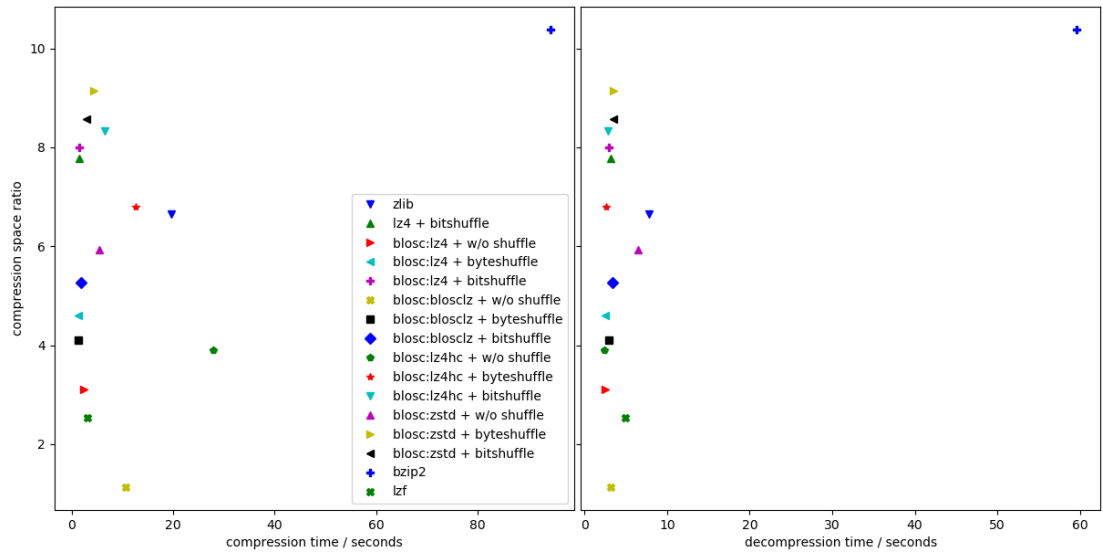


(c) AGIPD, lysozyme, integer truncated to 1 bit

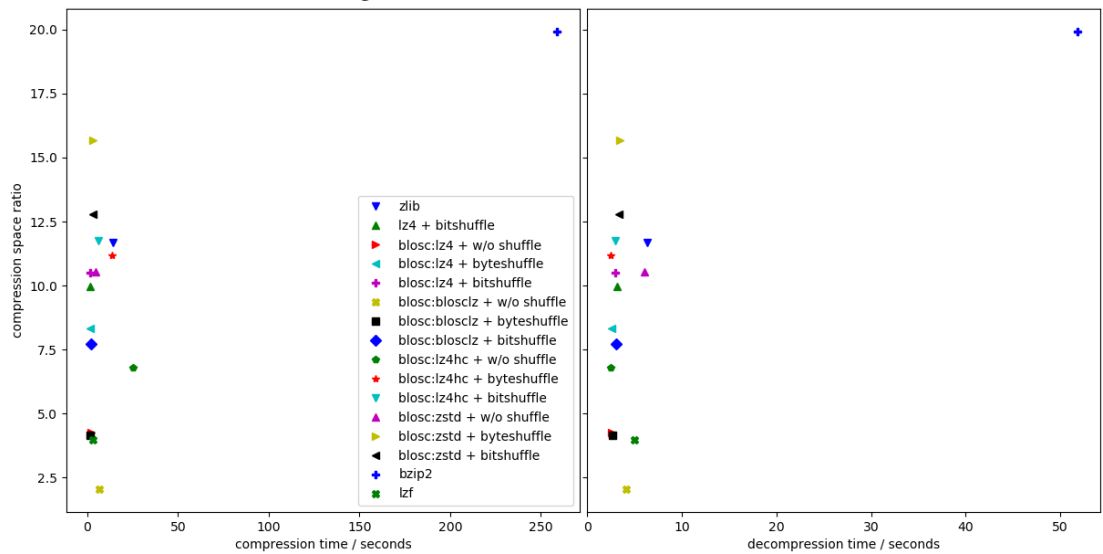
Figure 7.1: Relationship between compression ratio and compression/decompression speed for 1000 diffraction patterns for AGIPD dataset of lysozyme, including data subjected to lossy data reduction.



(a) Eiger 2X 16M, lactamase, integer data



(b) Eiger 2X 16M, lactamase, truncated to 3 bits



(c) Eiger 2X 16M, lactamase, truncated to 1 bit

Figure 7.2: Relationship between compression ratio and compression/decompression speed for 1000 diffraction patterns for Eiger 2X 16M dataset of lactamase, including data subjected to lossy data reduction.

Table 7.1: The results of the pattern rejection technique

Part	Num. of patterns/ hits	Indexed patterns/ crystals	$R_{\text{free}}/$ $R_{\text{work}}$
<b>1</b>	199606/	187826/	0.1561/
	198088	505329	0.1881
<b>1/4</b>	49902/	46947/	0.1576/
	49531	126301	0.1866
<b>1/8</b>	24951/	23477/	0.1603/
	24759	63191	0.1936
<b>1/16</b>	12475/	11759/	0.1688/
	12387	31731	0.1944
<b>1/32</b>	6238/	5888/	0.1728/
	6193	15859	0.2048
<b>1/64</b>	3119/	2929/	0.1794/
	3098	7895	0.2122
<b>1/128</b>	1559/	1450/	0.1932/
	1550	3968	0.2202

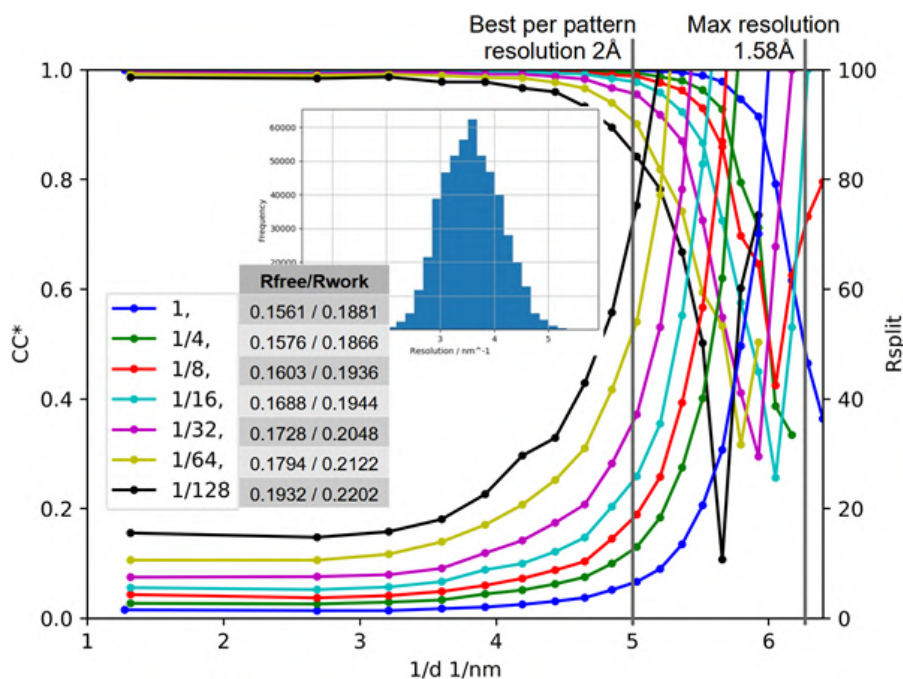


Figure 7.3: Data quality metrics  $CC^*$  and  $R_{\text{split}}$  for the whole dataset and its fractions. The insets show the table of  $R_{\text{free}}/R_{\text{work}}$  metrics and the histogram of achievable resolution for each pattern.

The degradation in data quality with fewer patterns is rather obvious from Fig. 7.3 and is to be expected given the fact that redundant measurements improve such statistical metrics as  $CC^*$  and  $R_{\text{split}}$  and thus the

quality of the obtained data. We can conclude that while the strategy of limiting measurement time is quite understandable, measuring more data always improves data quality in line with known statistics. Therefore, the lossy reduction idea to save space just by measuring fewer data is, in fact, not the best because it results in lower quality (see Fig. 7.3). On the other hand, the improvement of the resolution achievable using 1563 patterns (1/128) vs. all 200000 patterns is from 1.8 Å to 1.58 Å (0.22 Å difference). Therefore, the decision to halt data acquisition should be made based on the specific scientific inquiries of the study.

To evaluate data quality corresponding to the number of indexed lattices, the stream after *ambigator* was split randomly at 1/4, 1/8, 1/16, 1/32, 1/64 and 1/128 and then subjected to scaling and merging. Phenix [215] (`phenix.reflection_file_editor`) was used to add the same set of  $R_{free}$ -flags to each resulting dataset, and all datasets were refined with `phenix.refine`, using the same starting model, parameters, and resolution cut-off (as set by the highest resolution shell still containing useful data for the 1/128 dataset). `Polygon` [309] and `MolProbity` [310] and `Coot` were used for validation of the final model.

### 7.4.2.3 Storing only detectable Bragg peaks

Another proposed data reduction scheme is to save only the information around peaks found in each measured diffraction pattern. The idea is that only Bragg peak information affects the structure, so it should only be necessary to save information around the Bragg peaks.

Fig. 7.3 shows that adopting such a strategy will limit the resolution which can be achieved. For this dataset, if we limit ourselves to only using the found peaks, the achievable per pattern resolution would reach 2 Å ( $5\text{ nm}^{-1}$ ) per of patterns (see the resolution histogram in the inset), while the entire dataset achieves a resolution of 1.58 Å according to the  $CC^*$  cut-off decision. It is by now well known that redundant measurement of weak data improves the overall resolution achieved beyond the resolution at which peaks can be detected before integrating [311], see Fig. 7.4. As a consequence, compression schemes based on saving full detector data only around detectable peaks [312], will artificially limit the resolution. For example, processing the stream file from Table 7.2 to include only reflections from each pattern found in the initial peak search, the resulting resolution dropped to 1.62 Å and  $R_{work}/R_{free}$  increased to 0.236/0.292. In other words, we conclude retaining data from only detected peaks noticeably decreases the structure quality.

Table 7.2: Test of storing only hits on the lysozyme dataset collected at LCLS in 2011 [12].

Data	$R_{free}/R_{work}$	Resolution limit, Å
Published results	0.196/0.229	1.9
Re-processed	0.172/0.192	1.52
Stored only the found peaks	0.236/0.292	1.62

### 7.4.2.4 Binning to lower the number of detector pixels

Reducing the pixel count by binning data to fewer pixels is a lossy compression scheme widely used for different types of data, particularly when it is known that the detector has a finer pixel pitch than is strictly necessary. For

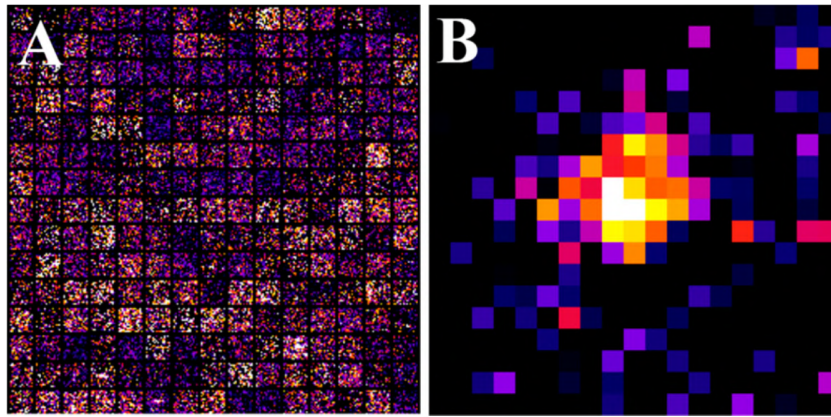


Figure 7.4: From [311]: (A) Two hundred and twenty-five randomly selected single images at the predicted location of the 21 26 29 reflections (corresponding to 2.3 Å resolution). (B) Averaged Bragg intensity of the 21 26 29 reflection intensities from 3,176 observed reflections after first rotating them into the common frame of reference of the lattice.

example, a 4M detector would suffice for the current experiment, but the beamline has a 16M detector. This data reduction scheme can be applied when the features of the diffraction pattern are much bigger than the pixel size. In protein diffraction, the Bragg peaks can be quite narrow. However, it is the integrated reflection intensity that contributes to the data; thus, not much information is discarded by binning provided the separation between peaks remains adequate for data processing and the shape of the Bragg peaks need not be resolved. For experiments with a monochromatic X-ray beam, such minimum distance should be 5-10 pixels. Therefore, many datasets can be binned, especially for proteins with small unit cell parameters.

Binned data, see details in Table E.2 and in Fig. E.1, indicate that 2x2 pixel binning for the tested datasets measured with 16M detector did not degrade the data quality for the samples we used, but the data volume was reduced by a factor of 3-4. The one caveat is that it might be more difficult to detect the peaks after the binning, therefore, we have developed a procedure in which the positions of the peaks are found before the binning, recalculated into the coordinates of the binned image and saved within the output HDF5 file. Those found positions can be used later at the integration step.

#### 7.4.2.5 Quantization of detector output

The quantisation of data refers to the reduction of the bit depth of the saved data to a fewer number of discrete values, reducing unnecessary precision in the stored values such as remapping 32-bit integers to 16- or 8-bit values, or converting floating point data to integers. In photon science, a common form of quantisation is converting the electrical signal to photon counts. This is performed in the electronics of counting detectors (Pilatus, Eiger), where each pixel directly counts the number of incident photons at high speed. As previously noted, such data compresses well using lossless compression schemes compared to data saved in floating point format. And in general data with fewer discrete values compresses better using most of the lossless compression schemes.

Since counting detectors are not suited for the short pulse lengths found at XFEL sources, integrating detectors that integrate the deposited charge in each pixel during the exposure are used. Converting the deposited charge into the number of incident photons helps to reduce the data precision required, however, this operation relies on good calibration of detectors and is not necessarily a trivial task, there is a tendency to save actual digitizer readout for later photon conversion



Our tests on quantization indicate that reducing the data precision of integrating detectors can be highly effective at enabling data compression. Results are presented in Table 7.3, where we test not only conversion to photons but also more aggressive reduction of data precision. For the AGIPD detector, even a quite high quantization level (1024 ADUs per quantum, which corresponds to approx. 14 photons at 9 keV) still achieves reasonably good data quality:  $R_{\text{free}}/R_{\text{work}}$  of 0.1753/0.1543 with a compression ratio of 64, compared to  $R_{\text{free}}/R_{\text{work}}$  of 0.1670/0.1497 for the original data.

Table 7.3: The result of a quantization approach with constant steps performed on the AGIPD lysozyme dataset (in this case 1 photon was equal to 73 ADUs), consisting of only hits. The data from the detector is calibrated and usually saved as “native float” (so-called “processed data”).

Level of rounding	Num. patterns/ hits	Indexed patterns/ crystals	$R_{\text{free}}/$ $R_{\text{work}}$	CR, gzip, w/o shuffle	CR, gzip, with shuffle
<b>float (original)</b>	189.9k/ 189.9k	166.8k/ 236k	0.1670/ 0.1497	1.102	-
<b>integer</b>	189.9k/ 189.9k	166.8k/ 236k	0.1689/ 0.1501	3.25	4.105
<b>rounded to 16 ADUs</b>	189.9k/ 189.9k	166.8k/ 236k	0.1666/ 0.1499	-	5.926
<b>rounded to 64 ADUs</b>	189.9k/ 189.9k	166.8k/ 236k	0.1677/ 0.1504	-	8.869
<b>rounded to 256 ADUs</b>	189.9k/ 189.9k	166.8k/ 236k	0.1690/ 0.1509	-	21.167
<b>rounded to 1024 ADUs</b>	189.9k/ 189.9k	166.7k/ 235.8k	0.1753/ 0.1543	46.829	63.578
<b>rounded to 4096 ADUs</b>	189.9k/ 189.9k	159k/ 225.8k	0.2431/ 0.1993	-	650.586

To compare the influence of the quantization to the described earlier approach of measuring less data in Section 7.4.2.2, we have performed the following test: we have applied rounding to 64 and 1024 ADUs for the full dataset (with 190k diffraction patterns) and to its 1/16 fraction (see Table 7.4 and Fig. 7.5). While the resulting size of the data was similar (for the full dataset rounded to 1024 ADUs and the 1/16 of the data converted to integer), the quality of the data was better for the rounded full dataset: the achievable resolution is higher and the statistics even at low resolution is better (see Fig. 7.5). This result suggests that the statistics in the measured data is more important than the precision of saving the measured intensities. Therefore, researchers should prioritise acquiring a sufficient number of patterns to ensure reliable and accurate structural information.

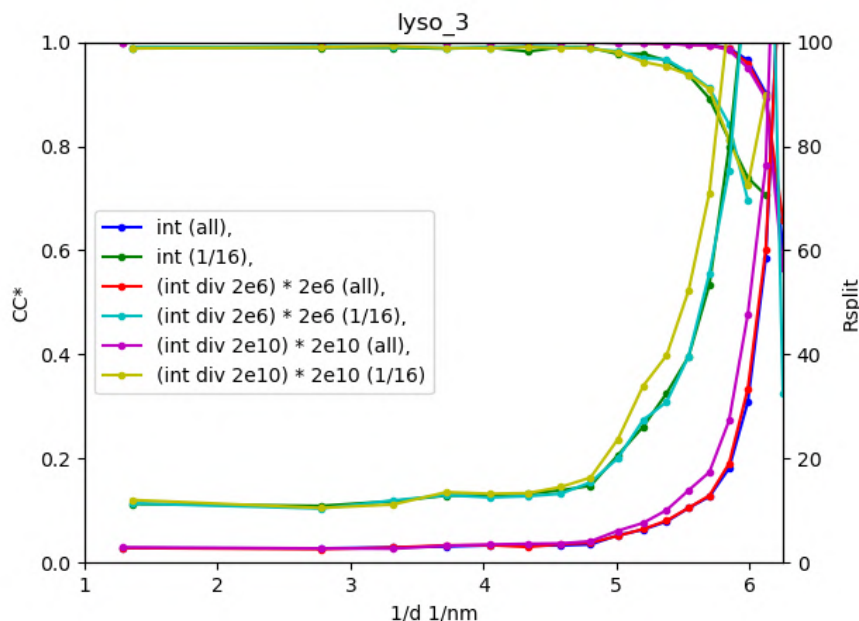


Figure 7.5: Data quality ( $CC^*$  and  $R_{split}$ ) for the datasets rounded to 1024 ADUs and for a small subset (1/16) of the same data. Diffraction from lysozyme crystals measure at EuXFEL using AGIPD.

Table 7.4: Results of influence of two lossy compression schemes: rounding to 64 and 1024 ADUs and using less data (1/16 of the original dataset). Diffraction from lysozyme crystals measure at EuXFEL using AGIPD.

Part	Num. patterns/hits	Indexed patterns/crystals	$R_{free}/R_{work}$
int (1)	189.9k/189.9k	166.8k/236k	0.1689/0.1501
int (1/16)	11.8k/11.8k	10.5k/14.7k	0.1846/0.1625
rounded to 64 ADUs (all)	189.9k/189.9k	166.8k/236k	0.1677/0.1504
rounded to 64 ADUs (1/16)	11.8k/11.8k	10.5k/14.7k	0.1881/0.1619
rounded to 1024 ADUs (1)	189.9k/189.9k	166.8k/236k	0.1753/0.1543
rounded to 1024 ADUs (1/16)	11.8k/11.8k	10.5k/14.7k	0.1856/0.1618

#### 7.4.2.6 Non-uniform quantisation

An even higher compression ratio can be achieved by selecting the levels for quantization in a non-uniform way. Diffraction from crystals usually consists of some background (typically smooth) and rather sharp Bragg peaks. As mentioned earlier, high dynamic range is usually needed to record such diffraction, with the intensity of the Bragg peaks varying from rather high (at low resolution) to very low (at high resolution). However, while single photon counting may be useful in the weak reflections it is not necessarily needed in the bright Bragg peaks. For this reason, special X-ray detectors (AGIPD, JUNGFRÄU, ePIX) were developed that have variable gains per each pixel to be able to record single photons at low flux as well as very high intensities (up to 10000 photons per pixel) at high flux in a single image [281, 282, 285, 286].

Table 7.5: Examples of rounding integer values to the three most significant bits, including the floating-point representation.

Initial number	Binary representation of the initial number	Binary representation of the resulting number	8-bit floating-like representation	Resulting number
81	0101 0001	0101 0000	00011101	80
87	0101 0111	0101 0000	00011101	80
88	0101 1000	0110 0000	00011110	96
258	0001 0000 0010	0001 0000 0000	00100100	256
1316	0101 0010 0100	0101 0000 0000	00101101	1280
1450	0101 1010 1010	0110 0000 0000	0101110	1536

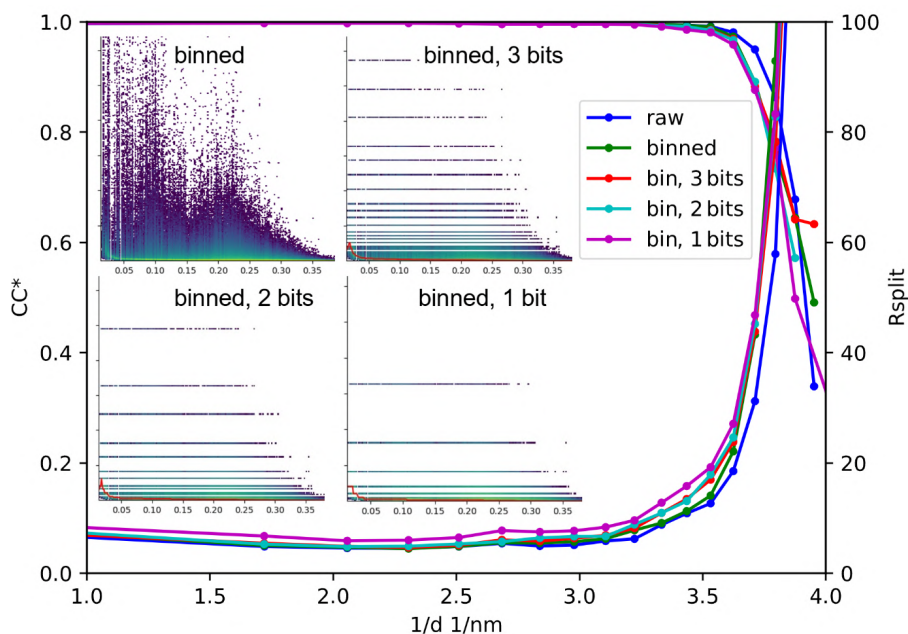


Figure 7.6: The quality metrics for the original data of thaumatin (PDB ID: 6S19), binned and rounded to 1, 2, 3 of the most significant bits. The histograms of data values over  $1/d$  (peakograms),  $d$  – resolution length in  $nm$ , for different datasets are shown in the insets. SAD dataset of thaumatin collected at 4.57 keV at SwissFEL [249].

The tolerable relative error of the peak intensity drives the required precision. Thus, at low photon counts the quantization levels are rather dense, while at the higher fluxes, the levels are comparatively sparse, in proportion to the counting noise. In the measured data this can be achieved by keeping the value of just a few of the most significant bits (starting from the first non-zero bit) in the integer representation of the measured intensity. For positive values, the simplest method is to preserve the most significant bit with the value of ‘1’ and set all other bits to 0. To get better results, rounding to the nearest value of 1, 2, or 3 most significant bits is utilised instead of truncation. An example figure representing the pixel intensities for different

Table 7.6: Lossy compression of SAD data of thaumatin. CR was achieved using gzip (compression level 6) and shuffle.

Type	CCano	$R_{\text{split}}/CC^*$	$R_{\text{work}}/R_{\text{free}}$	Number of residues	CR	CR after 8bit compressor
raw	0.327	5.97%/0.9983	0.2232/0.2821	209	1	-
binned	0.320	6.35%/0.9984	0.2165/0.2831	208	5	-
binned, 3 significant bits	0.247	6.65%/0.9984	0.2379/0.2993	205	32.2	36.3
binned, 2 significant bits	0.271	6.81%/0.9980	0.2700/0.3337	205	38.8	45.1
binned, 1 significant bit	0.251	7.94%/0.9980	0.5314/0.5514	181	49	59.3

distances from the centre of the detector is presented in the insets of Fig. 7.6. Such truncation will not make the data smaller, but modified data will be well-compressible by most of the lossless compression algorithms (Table 7.6). In Table 7.5, some examples of the numbers before and after the rounding are presented and the floating-point representation is demonstrated. More tests can be done using the code deposited on GitHub (<https://github.com/galchenm/binningANDcompression.git>). It is important to note that this rounding technique alone does not decrease the data size. However, the modified data becomes highly compressible using various lossless compression algorithms (as indicated in Table 7.6).

The proposed compression applied to the data discussed previously had almost no influence on data quality metrics such as  $CC^*$  or  $R_{\text{free}}/R_{\text{work}}$ , see Table 7.7. Therefore, for this test, we have chosen the technique more sensitive to data quality – SAD. We have used the thaumatin dataset (PDB ID: 6S19) measured at SwissFEL with JungFrau 16M detector [249]. The structures for different datasets after applying lossy compression algorithms of thaumatin were solved by SAD-phasing and refinement using the CRANCK2 pipeline [268] with default settings and without manual refinement. As seen from the Fig. 7.6, the data quality did not degrade much after applying rounding to several of the most significant bits. The results presented in Table 7.6 demonstrate that even the SAD data, which is very sensitive to the data quality, can still be successfully used if only 2 significant bits are saved (more statistics can be found in the Supplementary Table E.3). At the same time, saving just the single most significant bit is not enough for the same dataset – as can be seen from the last row of Table 7.6 the  $R_{\text{free}}/R_{\text{work}}$  are very high in this case.

Table 7.7: Influence of different quantization lossy compression on data (lysozyme, AGIPD) quality. The data was rounded to 64 and 73 (one photon) ADUs and also leaving only 3 and 1 the most significant bits was tested.

Type	Num. patterns/ hits	Indexed patterns / crystals	$R_{\text{free}}/$ $R_{\text{work}}$ (10 Å - 1.69 Å)	CR, gzip + shuffle	CR, bzip2 + shuffle
<b>Integer</b>	82798/ 82798	34720/ 34821	0.2048/ 0.1653	3.946	3.577
<b>Rounding to 64 ADUs</b>	82798/ 82798	34715/ 34830	0.2072/ 0.1632	5.137	5.926
<b>Photon conversion (to 73 ADUs)</b>	82798/ 82798	34685/ 34801	0.2077/ 0.1680	5.319	5.855
<b>Rounding to 3 bits</b>	82798/ 82798	34698/ 34812	0.2065/ 0.1655	5.421	6.124
<b>Rounding to 1 bit</b>	82798/ 82798	34447/ 34565	0.2048/ 0.1663	8.751	10.280

Retaining only the most significant bits is quite similar to the way the data is represented by floating-point numbers, thus we have also represented the integer data in a floating-point-like way - we have converted the 32-bit integers into 8-bit floating-point values: one bit for the sign, 5 bits for the exponent and 2 more bits for the mantissa. From the numbers in 7.6, one can see that such conversion allows to compress data even better. One very important benefit of the proposed lossy compression scheme is its speed. The truncation of the least significant bits requires very little computation. Indeed, the conversion may lend itself to implementation directly in hardware, such that it could be even realised within the detector.

## 7.5 Discussion

Some data reduction techniques described in this chapter have been successfully applied to SX data measured at different facilities. For example, users usually copy only diffraction patterns that contain diffraction (hits) after the SFX experiments at LCLS or EuXFEL, while the raw data are currently backed up for 10 years. At synchrotrons for some SSX experiments, raw data is just deleted without even archiving to tape: for well-diffracting crystals, we save only 'hits' to save space. For example, for pink beam experiments at APS or ESRF, we often collect data using the JUNGFR AU 1M detector at 1kHz speed, which results in up to 50Tb of raw files. After hit-finding and lossless compression, we copy only 2-5Tb of the data and delete everything else. This strategy is well justified by both our tests, showing that storing only hits is enough in some cases and the fact that the cost of the data storage is quite high. Therefore, it is often worth keeping only data reduced using lossy compression.

For 16M detectors, such as Eiger 16M or JUNGFRÄU 16M, we use binning together with saving only hits, such as a recipe for our SSX experiments performed at the P11 beamline of the Petra III synchrotron. For now, the optimisation of hit-finding parameters includes some human interaction, but it should be done automatically in the future. For experiments with the tape drive delivery system, [108], we routinely achieve a compression ratio of 5-7 on top of the compression factor of 5-6 achieved by bit-shuffle-filtered LZ4 compression, used by Dectris. This gives a total compression ratio up to 40 times compared to saving raw data.

As was demonstrated in this chapter, even higher compression rates are possible using the quantisation approach. This method works well for SX data because of good statistics: as we have shown, it is better to make the intensity values less precise than to measure fewer patterns. The reason is that the error in the determination of the exact integrated intensities of the Bragg peaks is rather high for the SX, mainly due to the unknown partiality and a possible error in the determination of the precise crystal orientation [82, 152, 313]. Additionally, one can do quantisation in a non-uniform way, having more levels at low intensity and fewer levels for strong signals. This way, the low-signal data is saved almost without losing precision, which is important for data measured at high resolution (closer to the detector edges).

Even higher compression rates can be achieved if only some intermediate data processing results are saved. For example, in rotation crystallography, one can save only the resulting merged reflection data; thus, the CR can easily reach thousands and more - instead of many gigabytes of raw files, only single-digit megabytes are saved. This approach is usually applied during or after the experiment (with a delay of minutes to months). But such an approach requires a very well-established pipeline: the geometry of the experiment has to be very well known, and the indexing and intensity integrating algorithms should work reliably. Unfortunately, this is not the case for SX experiments yet - as demonstrated in this chapter, reprocessing the previously measured data using modern tools can lead to much better results.

Rather, a similar idea for huge space reduction is on-the-fly data processing. With modern algorithms and computing power available, for example, at EuXFEL, and considering that the processing of the SX data is very well parallelizable, it is often possible to do real-time data processing even during the experiment. Such processing is great for quick feedback on the data quality, but, as was demonstrated in this chapter, careful reprocessing of the data can deliver much better results, for details see Section 6.1.2. The reprocessed lysozyme data set measured in 2011 shows better resolution than, for example, the data set measured in 2018 at EuXFEL [314] (in both experiments, the resolution was limited by the detector). Also, we re-analysed the Photosystem 1 dataset from the same experiment performed in 2011, and the achievable resolution is 2.9 Å. For comparison, the structure reconstructed using the data set measured in 2019 [315] had a resolution of 2.95 Å.

All compression schemes described in this chapter should also be applied to protein crystallography with electron or neutron diffraction. Some methods can be useful for other techniques that use diffraction and 2D detectors. However, the main demand for using any lossy compression remains the same: the data quality has to be carefully checked after compression.

## 7.6 Conclusion

Modern X-ray detectors combined with very bright sources, like FELs, can produce huge data volumes. Transferring and storing these data is rather expensive and technically challenging; therefore, data reduction must be implemented.

Protein crystallography is one of the most storage space-consuming techniques, especially its serial version (SX). We applied different data reduction algorithms for SX data measured at different facilities with various

detectors. Diverse lossless compressions were tested to find the highest compression rate at a reasonable speed. Unfortunately, the resulting CR for all lossless compressions is usually low (from 1.2 to 10, depending on the data). Thus, some lossy data reduction schemes have to be applied. The main requirement for any lossy compression is the resulting data quality preservation, so some quality metrics must be utilised to check the data after any lossy compression. We used the same data quality metrics that are usually used for further data processing; fortunately, different quality metrics in protein crystallography are well-known and widely used. We check such metrics as data quality and reproducibility ( $SNR$ ,  $R_{split}$ ,  $CC^*$ ), the quality of the reconstructed structure factors ( $R_{free}/R_{work}$ ) and the possibility of using an anomalous signal for *ab initio* structure reconstruction (SAD / MAD phasing). Importantly, each of the mentioned data quality metrics should be used properly. For example, the signal-to-noise level of the structure factors measured at high resolution (far from the detector centre) is usually rather low; therefore, these data are quite sensitive to data quality reduction. So, the model metrics  $R_{free} / R_{work}$  must be checked while reconstructing the structure to the highest possible resolution. Also, what is very important is that the geometry of the experiment should not limit the resolution of the data. In addition, the anomalous signal has to be analysed for the dataset, where the signal is rather low, even for the raw data. This can be easily achieved by, for example, reducing the statistics of the measured data and simply selecting a small portion of the data that is just enough for the method to work [249].

The following lossy data reduction schemes were tested: non-hits rejection, binning, quantisation with different step sizes, quantisation with nonuniform step, and saving only found peaks. Non-hit rejection is often used for SX experiments due to the low hit rate (0.1-10%). We have demonstrated by reprocessing a 12-year-old dataset that this scheme is well justified for the datasets with strong hits. Binning is often applied for data measured with multiple-megapixel detectors, especially if the crystals are measured with small cell parameters. For this method, the most important quantity is the minimal distance between the measured Bragg peaks, which has to be at least 5 pixels after the binning. Quantisation greatly helps, especially for data originally saved in the floating-point format. Quantisation to photons is actively used already for different detectors, but we have demonstrated that even quantisation to several photons can still preserve reasonably good data quality. Even better results can be achieved using quantisation with the non-uniform step: having more levels for the low intensity than for the high-intensity data. The easiest implementation is to truncate (we use rounding) each value to just a few significant bits. This scheme makes the data much more compressible by any lossless compression, and the data quality, while rounding to 3 bits, is fully preserved in the tests we have performed. This lossy compression is quite similar to the multiple-gain mode used in modern detectors for capturing high dynamic range while keeping high sensitivity for low signals. One more advantage of such a scheme is that it is very computationally efficient, and it can be easily implemented in an FPGA inside the detector to make the compression on-the-fly.

Some data reduction methods proposed and even used by other groups have shown a reduction in data quality after the application. For example, reducing the statistics (measuring fewer data) leads to resolution and an overall reduction in data quality. The same applies to saving only the peaks determined during the hit finding. Saving only the resulting mtz file, as well as on-the-fly data processing, has the drawback that the data cannot be reprocessed later, so it cannot benefit from the improved algorithms and better detector corrections.

Based on the results of our test, we recommend the following strategy for saving SX data:

- For datasets with strong diffraction (easily detectable Bragg peaks) storing only “hits” (diffraction patterns with detected crystal diffraction)
- For big (in terms of the number of pixels) detectors, we recommend to bin the data to the level when the

distance between the neighbouring peaks is bigger than 5 pixels

- Apply non-uniform quantization, rounding data to just 3 the most significant bits, applying some lossless compression (gzip, lz4 or bzip2) to the rounded data

Some proposed data reduction schemes (non-hits rejection, rounding to the nearest integer, binning) are already successfully applied to the data we measure at different facilities, while the raw data are not even backed up. In the near future, similar strategies have to be applied at different facilities worldwide. Otherwise, we will be drowned in the data flood we are generating.



---

## Summary and outlook

Protein crystallography is one of the most successful methods for determining biological structures. This technique requires the acquisition of numerous diffraction snapshots to obtain 3D structural information on the studied protein. In macromolecular crystallography, the conventional approach involves acquiring diffraction patterns from a crystal while it undergoes rotation along one or more axes. It is a well-developed technique with established data processing pipelines that efficiently transform raw diffraction images into structure factors, reducing the volume of useful data from gigabytes to hundreds of kilobytes. Another rapidly developing method suitable for structure determination from small or radiation-sensitive crystals and for investigating fast or irreversible protein dynamics is known as serial crystallography (SX) [9, 12, 67]. A notable advantage of serial crystallography is the ability to apply the full tolerable X-ray exposure to each individual crystal instead of distributing it across a rotation series of a single crystal. This approach can avoid the need for cryogenic cooling, thus allowing the measurements at room temperature.

Recent advancements in X-ray facilities, including 3rd and 4th-generation synchrotrons and Free Electron Lasers (FELs), in combination with state-of-the-art X-ray detectors, have enabled conducting sophisticated experiments at a remarkable rate, capturing more than 1000 images per second. However, the increased acquisition rate comes with a trade-off - an enormous volume of data, with some experiments yielding up to 2 PB of data. Moreover, this number keeps growing yearly due to the development of even faster detectors and the emergence of brighter sources. At the same time, the cost of storage (an example of the price per PB can be found here <https://wasabi.com/blog/on-premises-vs-cloud-storage/>) has been relatively stable over the past decade. Therefore, the central focus of this thesis revolved around data compression and reduction techniques for serial crystallography.

Chapter 7 extensively covers data reduction techniques, such as lossless and lossy compression methods and data dimensionality reduction. Since lossless algorithms usually fail to achieve a sufficient compression rate for SX data, it must be combined with some lossy reduction technique. At the same time, lossy compression may spoil the data, thus affecting the scientific outcome of the experiment. That is why special attention in the Thesis is dedicated to the data quality check after any reduction scheme. Different quality metrics are described, evaluated and applied for evaluation of various data reduction schemes. Importantly, the proper application of the quality metrics is described in detail. The proposed quality check methods are vital for any existing lossy compression schemes and can be used for future developments of new data reduction methods.

Comparing the data acquisition rate increase by 100 times (CSPAD 120 Hz in 2011 to ePixHR 10k up to 10 kHz in 2024) over the last decade and the much slower growth of the storage capacity (increased only by 10: 2 Tb per disk in 2012 to 20 Tb per disk in 2021), it is obvious that the storage of all raw data soon would become

impossible or, at least, not feasible. Thus, the lossy data reduction has to become standard at any facility that operates at these high acquisition rates. Ideally, this task should be automatically accomplished by the facility. But first, the user community must accept the applied lossy reduction methods. That's why it is important to demonstrate that the data quality after reduction is sufficient to get a similar scientific outcome compared to the processing of original raw data.

Various research groups have already used data reduction schemes, indicating some acceptance of the compromises involved. For instance, after performing Serial Femtosecond Crystallography (SFX) experiments at facilities like LCLS or EuXFEL, users typically copy only the diffraction patterns that contain actual diffraction ("hits"). The strategy eliminates non-informative frames while preserving the essential information required for downstream analysis. This practice is based on the understanding that the facility retains all raw data for a period of 10 years. However, at synchrotron facilities, raw data is often deleted after a certain period of time and is not archived on tape at the facility. Therefore, it is essential to find a compromise between the cost of storing all data and the possibility of reproducing or improving the results later. To check if it is sufficient to store only hits, we have reprocessed an old dataset measured in 2011 (both only hits and whole raw data) and concluded that storing only hits is well justified in the case of strongly scattering crystals. All details of this evaluation are described in Section 7.4.2.1.

By combining non-hits rejection and lossless compression, the overall data storage and management can be significantly optimized without adversely impacting the integrity or scientific value of the data. The demonstrated test results provide compelling evidence supporting the efficacy and validity of this approach, reinforcing its applicability in the context of serial crystallography experiments.

In the case of multi-megapixel detectors like the Eiger 16M or JUNGFRÄU 16M, a practical solution for reducing data size is to bin the data to a smaller detector, as long as the crystal unit cell and the achievable resolution allows sufficient separation between Bragg peaks. This approach is applied in conjunction with saving only the hits for Serial Synchrotron Crystallography (SSX) experiments conducted at the P11 beamline of the Petra III synchrotron. After the experiments, the raw data was substituted by reduced data and deleted. Through binning and non-hits rejection, we routinely achieved a compression ratio of 5-7 on top of the compression factor 5-6 achieved by the lossless compression like gzip, as described in Section 7.4.2.4. This combined compression strategy results in an impressive total compression ratio of up to 40 times compared to saving uncompressed raw data while maintaining a high level of scientific output.

Even higher compression rates are possible by quantizing the detector output into fewer discrete levels, presented in Section 7.4.2.5. This method is particularly effective for Serial Crystallography (SX) data, where the focus is on statistical measurements rather than precise intensity values in each pattern. Additionally, the quantization in a nonlinear way is described in Section 7.4.2.6. This quantization allows finer increments at low intensity and coarser increments for strong signals. In this way, the low signal data is saved almost without losing the precision, which is very important for data measured at high resolution close to the detector edges. The quantization with non-uniform increments offers a promising approach to achieve significant data compression while retaining the essential information required for further data analysis. This approach provides a promising avenue for reducing data size without sacrificing the quality and reliability of the scientific results obtained from serial crystallography experiments.

A set of data metrics capable of assessing the loss of information due to applying various compression schemes is used to evaluate the effect of any lossy compression schemes. This required a careful understanding of the specific analysis techniques but is nevertheless an imperative step in evaluating different compression algorithms. Such metrics as data quality and reproducibility ( $SNR$ ,  $R_{split}$ ,  $CC^*$ ), the quality of the reconstructed

structure factors ( $\mathbf{R}_{\text{free}}/\mathbf{R}_{\text{work}}$ ), and the possibility of using the anomalous signal for ab-initio structure reconstruction (SAD phasing) were employed for proper data quality evaluation. Properly employing each of the mentioned data quality metrics is of utmost importance. For example, one has to ensure that the achievable resolution is not limited by the geometry of the experiment (detector edge resolution) or that the anomalous dataset has overwhelming statistics to tolerate the introduced loss of information. Failure to address such limitations may render certain quality metrics insensitive to potential degradation in data quality.

An essential responsibility involves showcasing the limitations of certain proposed data reduction methods and cautioning potential users regarding these constraints. For example, saving data only around found Bragg peaks may significantly lose achievable resolution. This happens because the resolution of the whole dataset often extends beyond found peaks due to the presence of a weak signal at Bragg peak locations in individual diffraction patterns, which nevertheless integrates above noise levels when many observations of the same reflection are averaged. Similarly, reducing the measurement time and, thus, the number of measured diffraction patterns reduces data statistics: fewer measured patterns means fewer observations of each reflection, thus a reduction in signal-to-noise of the merged reflection intensity. Section 7.4.2.2 demonstrates the data quality degradation after applying the above-mentioned data reduction schemes.

Even better data reduction can be achieved if the original diffraction patterns (raw data) are discarded and only intermediate calculation results are retained. For example, in rotational crystallography, it is common to look at only the resulting merged reflection data and the original diffraction patterns are almost never revisited. Similarly, efforts are underway in SX to perform all indexing and integration in real time, obviating the need to save individual diffraction patterns. If this can be done, the compression ratios achieved can be enormous - instead of many gigabytes (tens of terabytes in the case of SX) of raw files, less than 10 megabytes are saved. This approach is usually applied during or after the experiment (minutes to months delay). In this case, however, revisiting the original data later is impossible. Such approaches can only be adopted when a well-established pipeline exists and all calibration factors, including the geometry of the experiment [26] and the detector response, are very well known. While this is not the case yet for SX experiments, an investment in robust geometry and detector calibration combined with an established analysis pipeline could significantly reduce saved data volumes in the future.

To demonstrate the influence of the improved data processing software in SX experiments, careful re-processing of previously collected data was performed and is presented in Chapter 6. In particular, Section 6.1.2 showcases the reprocessing of various datasets collected in 2011, which resulted in higher resolution outcomes compared to the original analysis, provided that the raw frames containing crystal diffraction were preserved. Even some structural features not observed during the initial analysis were resolved after the reprocessing. This underscores the importance of storing measured diffraction patterns for extended periods, necessitating effective data compression schemes.

It is important to highlight that the compression schemes discussed in this thesis can also be applied to protein crystallography involving electron or neutron diffraction techniques. The methods presented here offer potential benefits for other experimental approaches that utilise diffraction and 2D detectors. However, it is essential to acknowledge that each analysis chain is unique, making it challenging to generalise the impact of compression or data reduction across different techniques. Therefore, the data quality check, similar to the one introduced in Chapter 5, is required at each stage of any data processing pipeline. By incorporating robust data quality evaluation strategies into the analysis pipeline, researchers can confidently interpret the outcomes derived from processing the reduced data or reprocessing previously collected data using novel algorithms. This approach enables researchers to explore data reduction methods or adapt existing pipelines without compromising the

scientific outcome.

Even though serial crystallography appeared to be very promising, the user community of this method is still much smaller than for traditional data collection in the form of rotation series. This happens partially because SX requires different and often more complicated sample delivery and preparation methods. But the main roadblock for SX is its complicated data analysis: while in MX, most of the analysis can be done automatically using such software packages as XDS or `autoproc`, in SX, a lot of tasks have to be done by a user or beamline staff. Such steps as the geometry refinement, bad regions masking, and even plotting the statistics for each dataset are still not automated. Also, the new sample delivery methods often require dedicated strategies for efficiently using the sample and the beamtime. This Thesis partially solves all the mentioned bottlenecks and makes the analysis of SX data more user-friendly.

Two commonly employed sample delivery systems for SX include jets and fixed targets. Although jets can be utilized for measurements at an MHz rate, which is necessary at European XFEL, fixed-target techniques offer numerous benefits. These advantages encompass low sample consumption, clog-free delivery, often lower background, and the ability to control crystal-on-chip density for optimal hit rates. Chapter 5 describes the optimised approach of measuring crystals deposited on fixed-target supports (chips) using a two-step scanning mode: first, the chip is scanned at a low dose and high speed to determine the positions of crystals on the chip. Only these positions are subsequently measured as mini rotation series. Optimising the scanning process offers multiple benefits, including faster data collection and reduced data volume. By avoiding the collection of empty frames from crystal-lacking positions, unnecessary data is prevented from being measured and stored, resulting in significant resource savings. The method exhibits its most significant improvement when the chip is loaded with only a few crystals, often for certain proteins that are particularly challenging to crystallise. In such instances, the proposed method becomes instrumental in maximising the utilisation of all available crystals, greatly enhancing the likelihood of successfully obtaining the structure of the measured protein. The method proposed in this dissertation is applicable both at synchrotrons and laboratory sources. Using an attenuated beam, the method can be utilised at FELs to measure rapid dynamical processes, such as light-activated phenomena, in protein crystals.

The next step after data collection is processing the measured diffraction patterns. As discussed in Section 5.1, precise knowledge of the detector geometry with sub-pixel accuracy and optimised data processing parameters are crucial to extracting the maximum information from the collected data.

The situation becomes even more challenging when dealing with beam drift or the need to reposition the detector due to the different unit cell parameters of the measured samples. All these mentioned challenges highlight the necessity of developing a robust data processing pipeline capable of overcoming these difficulties and enhancing the quality of the obtained results. Such a pipeline would automate the data analysis process, allowing for efficient and reliable processing by generating figures of merits at each data processing step, even in complex experimental scenarios.

To make the SX data analysis easier for users and to address the mentioned issues, an offline data processing pipeline was developed and presented in Section 5.4. This pipeline fully analyses SFX data from raw images to merged *hkl* intensities with the corresponding calculated data quality metrics that are automatically deposited to a Google spreadsheet and summarised in the table format required for publishing articles. Additionally, in Section 5.4, the detailed user manual for the developed pipeline is provided. The main advantage of this pipeline is the ease of use and reliability of the results at each data processing stage. The pipeline consists of separate blocks that are called sequentially. Such a structure is beneficial because these blocks could be used as standalone programs if the user is interested in a specific processing step. Furthermore, this versatile pipeline

facilitates both on- and offline data handling, accommodating conventional and serial crystallographic data. It enables the efficient processing of diverse datasets, thereby enhancing the overall workflow of crystallographic experiments. The pipeline and individual programs described in Chapter 5 were rigorously tested during experiments conducted at various facilities, including PETRA III, EuXFEL, LCLS, SwissFEL, and ESRF. These tests ensured the robustness and effectiveness of the developed data processing tools across different experimental setups, including various sample delivery systems and detectors.

Specifically, as outlined in Section 5.6, the established data processing pipeline has been adapted and seamlessly integrated into the beamline control system of the P09 drug screening beamline at Petra III. This integration ensures full automation of all data handling stages. The experiments conducted at P09 employ four different detectors: Lambda (1.5M), Pilatus CdTe (2M), Pilatus 6M and Eiger 4M. The pipeline is designed to support all four detectors, automatically determining which detector was used during the experiment. This pipeline, deployed at the P09 beamline at PETRA III, underwent thorough testing in early March 2023. The integration of the analysis pipeline marked a significant milestone in automating data handling processes and improving the overall efficiency of experiments conducted at the P09 beamline.

The developed data processing pipeline, optimised data acquisition using chips, and data reduction strategies allow for solving some of the urgent existing problems in protein crystallography. The developed methods are compatible with various control software at different facilities. They enable to conduct experiments with complicated scenarios. Moreover, they found their application at the drug-screening beamline P09, PETRA III.



# Introduction to X-rays

## A.1 Correction Terms for the Atomic Scattering Factor

It is necessary to consider how the scattering of radiation by the electrons of an atom is affected by the fact of the bounded electrons.

The phenomenon of absorption of radiation has not to be ignored: in reality, electrons are bound to atoms and assumed discrete energy levels determined by quantum mechanics, or in the case of valence electrons for atoms in condensed matter, energy bands. Thus, the electrons have defined binding energies. The response of bound electrons depends on the relationship between the binding energy and the energy of the incident photon.

There are three distinct scenarios to consider when examining the interaction between photons and electrons:

1. Firstly, when the photon energy is significantly smaller than the binding energy of the electron.
2. Secondly, when the photon energy is significantly larger than the binding energy.
3. Lastly, when the photon energy precisely matches the binding energy, it results in a condition called resonance.

The atomic form factor of the elements as a function of the scattering vector  $r^*$  as follows

$$f^0(\sin \theta/\lambda) = \sum_{i=1}^4 a_i \exp(-b_i \sin^2 \theta/\lambda^2) + c \quad (\text{A.1})$$

The expression Eqn. A.1 is the simplest description of the atomic form factor, which assumes the electrons are unhindered by their response to the incoming X-rays, by the fact that they are bound to atomic nuclei. The photon energy is much larger than the electrons binding energy, this is a valid approximation.

Electrons in atoms assume well-defined energies enforced by the laws of quantum mechanics. The core level 1s or  $K$  electrons are the most strongly bound, followed by  $L$  electrons, and so on.

The response of a bound electron to instant electromagnetic radiation can be modelled as a damped oscillator responding to an oscillatory driving force. The system has a natural oscillation frequency given by  $\omega_B = E_B/\hbar$ . If the driving frequency  $\omega$  is much smaller than  $\omega_B$ , the response amplitude and associated cross-section are strongly suppressed by the electron being bounded. The cross-section drops off as approximately  $(\frac{h\nu}{E_B})^4$  which is called Rayleigh scattering. For hard X-rays, a good fraction of the electrons will have binding energies  $E_B$  that are much smaller than the photon energy  $h\nu$ . Thus, for photon energies far above resonance, the electrons respond to excitation by the photons essentially as if they were free, and the cross-section approaches that of the

free electron, the Thomson cross-section. The re-emitted radiation has a phase of 180 degrees relative to the instant beam due to the electron response being exactly out of phase with the instant radiation. At resonance, the cross-section of the electron is enhanced, and the phase of the re-emitted radiation is 90 degrees, which is associated with absorption.

For X-rays with photon energies of the order of a keV or higher, most electrons and all the elements can be considered to be quasi-free regarding their response to and cross-sections for interactions with those photons. Nonetheless, the impact of resonance is crucial for many phenomena.

First, it is needed to formalise the impact of resonances on the atomic form factor  $f$ . The expression of the atomic form factor needs to be modified by two extra terms that depend on the photon energy called  $f'$  and  $f''$ . Because of their dependence on photon energy, these are called dispersive terms.  $f'$  is a negative correction to the atomic form factor, which describes that the bound electrons have a damped resonance.

$f_1(r^*, \hbar\omega) = f^0(r^*) + f'(\hbar\omega)$ , because  $f'$  is negative,  $f_1$  is smaller than  $f^0$ . Therefore,  $f'$  accounts for the dumping in the electrons' oscillation amplitude due to it being bounded to the atom. The effect of  $f'$  is, therefore, to make the atom appear to have a lower electron density than it has, at least from the perspective of the X-rays. One can calculate the apparent electron density as seen by the X-rays by performing X-ray reflectivity measurements.

Close to an electron's binding energy, not only is  $f'$  enhanced, but a second term becomes important called  $f''$ . Because the resonant electron has a phase of 90 degrees to the instant radiation, this term is multiplied by  $i = \sqrt{-1}$ .  $f''$ , also called  $f_2$ , is the imaginary component of the total form factor and  $f_1$  is the real component.  $f_2$  results in energy dissipation due to photo absorption and is equal to  $\frac{\sigma_a}{2\lambda r_0}$ , where  $\sigma_a$  is the absorption cross-section and  $r_0$  is the Thomson scattering length equals  $2.82 \times 10^{-5}$  Å. Thus, the expression for the total form factor can be represented as:

$$f(r^*, \hbar\omega) = f^0(r^*) + f'(\hbar\omega) + if_2(\hbar\omega) \quad (\text{A.2})$$

## A.2 Refraction, reflection and absorption

We begin with refraction. The wavelength of visible light is reduced because the velocity is smaller in a transparent medium by a factor  $n$  known as the refractive index in the visible regime. The light beams are bent to steeper angles when entering the medium. The refractive index  $n$  has both a real  $n_R$  and an imaginary  $n_I$  component:

$$n = n_R + in_I \quad (\text{A.3})$$

The real part  $n_R$  is a more familiar number that describes refraction, while the imaginary part  $n_I$  refers to absorption. The refractive index changes with photon energy. The refractive index of a medium is described by a model in which the bound electrons are forced into damped oscillations by a driving electromagnetic field and could be expressed with the following formula:

$$n^2 = 1 + \left( \frac{e^2 \rho}{\epsilon_0 m_e \omega^2} \right) \frac{1}{\left( \frac{\omega_B}{\omega} \right)^2 - 1 - \frac{i\Gamma}{\omega}} \quad (\text{A.4})$$

Below resonance, at  $E_B = \hbar\omega_B$  we are in the visible regime. For photon energies well above resonance, the refractive index can be simplified to

$$n = 1 - \delta + i\beta \quad (\text{A.5})$$



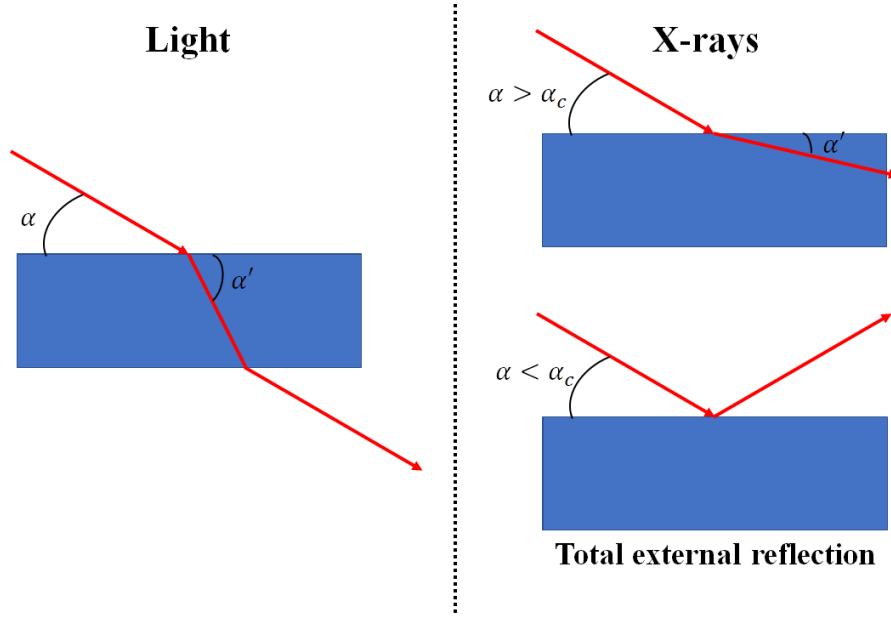


Figure A.1: The figure is an adaptation from [39]. The refraction of visible light by a transparent medium is greater than unity. In contrast, the refractive index for X-rays is marginally less than one. At grazing angles smaller than critical angle  $\alpha_c$ , the total external reflection will occur.

where both  $\delta$  and  $\beta$  are small numbers. Or, in other words,  $n_R = 1 - \delta$  and  $n_I = \beta$ .  $\delta$  is called the refractive index decrement, while  $\beta$  is the absorption index. Below the resonance, the real part of the refractive index is greater than one, while above the resonance, it is less than unity.

Snell's law gives the changes in the angle of light for a beam of radiation travelling through a heterogeneous interface from a medium with a smaller real part of the refractive index to one with a larger  $n_R$ :

$$\frac{\cos \alpha}{\cos \alpha'} = \frac{n_{R'}}{n_R}. \quad (\text{A.6})$$

Because matter has refractive indices greater than one in the visible and less than one in the X-ray regime, we can expect different refraction effects. Visible light is bent to steeper angles when passing from a lighter to a denser medium. For X-rays, the opposite is true: light is bent to shallower angles when entering a denser medium, as the refractive index of the matter is less than unity (see Fig. A.1).

This implies that the phase velocity  $v_P = c n_R$  of X-rays in the medium is greater than  $c$ , - the speed of light. In a vacuum, the crests and troughs of the electromagnetic wave travel at the same speed as the envelope Gaussian pulse. However, in a medium, the crests and troughs can move faster than the envelope. The speed of the crests and troughs is the phase velocity, while that of the Gaussian envelope is the group velocity. For X-rays, the group velocity carries the energy, which is equal to  $n_R \times c$ , which is smaller than the speed of light.

It turns out that the refractive index in the X-ray regime can be expressed as follows:

$$n = 1 - \frac{r_0}{2\pi} \lambda^2 \sum_i N_i f_i(0) \quad (\text{A.7})$$

where  $f_i(0)$  is the complex atomic scattering factor of the  $i^{\text{th}}$  atom in the forward direction,  $N_i$  - is the number density of atom type  $i$ . Recalling that  $n = 1 - \delta + i\beta$ , the expression for

$$\delta = \frac{r_0}{2\pi} \lambda^2 \sum_i N_i f_{1,i}(0) \quad (\text{A.8})$$

where  $f_{1,i}(0) = Z_i + f'_i$  is a real part of the atomic form factor.  $\sum_i N_i f_{1,i}(0)$  is simply the electron density as seen by the X-rays. And

$$\beta = \frac{r_0}{2\pi} \lambda^2 \sum_i N_i f_{2,i}(0) \quad (\text{A.9})$$

Where  $f_{2,i}(0)$  is the magnitude of the imaginary components of the atomic form factor.

Total internal reflection is a phenomenon in the visible regime that occurs when the light within one medium of refractive index  $n_2$  approaches the interface to another medium of lower refractive index  $n_1$  at a sufficiently glancing angle at or below the so-called critical angle. This phenomenon is a simple consequence of Snell's law:

$$\frac{\cos \alpha_1}{\cos \alpha_2} = \frac{n_2}{n_1} \quad (\text{A.10})$$

At a certain angle, the refracted ray becomes parallel to the surface; in other words,  $\alpha_2 = 0$ . As the angle of incidence decreases beyond this critical angle, the condition of refraction can no longer be satisfied. There is no refracted ray, and the ray incident on the interface is internally reflected.

In the X-ray regime, the refractive index of the matter is smaller than one, but only by a very small amount. Now, a beam of X-rays entering a medium of refractive index  $n_2$  from, for example, a vacuum, will be refracted at marginally shallower angles. Following the same arguments as before, there will be a minimum angle of incidence at which the refracted beam will be parallel to the surface. This angle is the critical angle for total external reflection and has a value determined by  $\cos \alpha_C = n_2$ . For incident angles below  $\alpha_C$ , the x-rays will therefore be externally reflected.

We know that  $n_2 \approx 1 - \delta$ , where  $\delta$  is the refractive index decrement. For now, we will ignore the other term that describes the refractive index, namely  $i\beta$ , a value typically between 1-3 orders of magnitude lower than  $\delta$ . Because  $\delta$  is small,  $\cos \alpha_C$  must be close to one, which means that  $\alpha_C$  is a very small angle, which magnitude can be estimated as follows  $\alpha_C \approx \sqrt{2\delta}$ . This phenomenon is used to construct X-ray mirrors. X-rays incident on surfaces at angles below the critical angle will be externally reflected. Although synchrotron beams generally have low divergences, they may be of the order of a millimetre or more in height at a position where a mirror could be installed at a beamline.

In general, the reflection intensity is equal to the square of the reflection amplitude, which depends on the incident and transmission angle,  $\alpha$  and  $\alpha'$ , and thus, the reflectivity can be expressed as

$$R = r^2 = \left( \frac{\alpha - \alpha'}{\alpha + \alpha'} \right)^2 \quad (\text{A.11})$$

At or below the critical angle,  $\alpha'$  equals zero; hence, the reflectivity is 100 %. Using Snell's law

$$\frac{\cos \alpha}{\cos \alpha'} = 1 - \delta \quad (\text{A.12})$$

the equation for  $R$  can be expressed in terms of  $\delta$  and  $\alpha$  instead of  $\alpha$  and  $\alpha'$ :

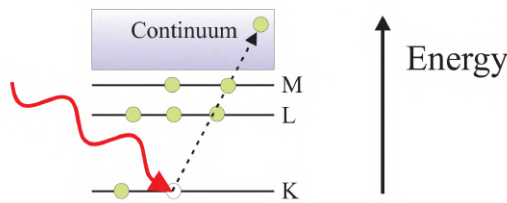
$$R = \left[ \frac{1 - (1 - 2\delta/\alpha^2)^{1/2}}{1 + (1 - 2\delta/\alpha^2)^{1/2}} \right]^2 \quad (\text{A.13})$$

At angles substantially larger than  $\alpha_C$ , the reflectivity drop-off is the inverse fourth power of the incident angle. Although below the critical angle, the reflectivity is 100 %, there must be some interaction going on for a reflection to happen. This wave that penetrates below the surface is called an evanescent wave. Below the critical angle, the X-rays penetrate the material only by a few angstroms, making it a sensitive probe for surface and interface properties.

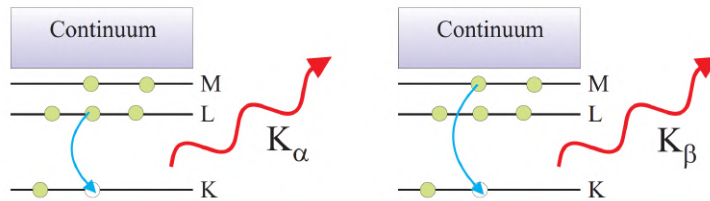
The penetration depth  $\Lambda_0$  depends only on  $1/\sqrt{\rho}$ , where  $\rho$  is the perceived electron density. At angles well above the critical angle, the penetration depth is proportional to the incident angle  $\alpha$  and the X-ray wavelength  $\lambda$  and inversely proportional to the extinction coefficient  $\beta$ . At the critical angle, the penetration depth is proportional to the X-ray wavelength  $\lambda$  and  $1/\sqrt{\beta}$ .

Photo-absorption can occur as a result of promoting an electron from where it resides in, by definition, an occupied state, to either an unoccupied but bound state, normally an excited state, or indeed, into a so-called vacuum state for which the electron is completely free and has escaped the system. Photo-electron spectroscopy investigates the energy and momenta of such electrons. Conservation of energy dictates that the absorbed photon has energy equal to the difference in energy between the final and initial states of the electron.

(a) Photoelectric absorption



(b) Fluorescent X-ray emission



(c) Auger electron emission

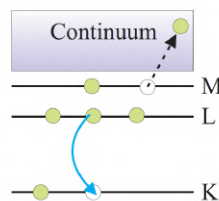


Figure A.2: Schematic energy level diagram of an atom (taken from [39]). (a) The photoelectric absorption process. An X-ray photon is absorbed, and an electron is ejected from the atom. The hole created in the inner shell can be filled by one of two distinct processes: (b) Fluorescent X-ray emission. One of the electrons in an outer shell fills the hole, creating a photon. (c) Auger electron emission. The atom may also relax to its ground state energy by liberating an electron.

We consider a plane wave of X-rays incident on an absorbing block of material with a refractive index:  $n = 1 - \delta + i\beta$ . The refractive index causes the wavelength to increase by a factor of  $\frac{1}{1-\delta}$  while travelling through the block. Moreover, the amplitude of the wave falls exponentially due to absorption. The decay rate is determined by  $\beta$ . In a vacuum, the time and space-dependent electric field amplitude is given by:

$$E = E_0 \times \exp i(k_0 z - \omega t) \tag{A.14}$$

In the medium, however, we need to include the refractive index in the spatial term:

$$E = E_0 \times \exp i(nk_0z - \omega t) \quad (\text{A.15})$$

The temporal term  $\omega t$  is unaffected, as the frequency of the oscillations of field amplitude  $E$  cannot change. Inserting the expression for the complex refractive index into the equation for propagation through matter, we arrive at an expression that contains a pre-factor describing the exponential decay  $\exp(-\beta k_0z)$  plus the oscillatory term, which the wave vector  $k$  decreases by a factor  $1 - \delta$ , thus describing the increased wavelength in the medium, and, therefore, the entire expression is

$$E(z, t) = E_0 \exp(-\beta k_0z) \exp(i(1 - \delta)k_0z - \omega t) \quad (\text{A.16})$$

In reality, when measuring X-rays, we can only directly detect the intensity: the oscillations of the amplitude of the X-ray radiation (which are at approximately  $10^{18}$  Hz) are many orders of magnitude too rapid to follow using even the fastest modern electronics. The detected intensity is proportional to the squared amplitude. Because the amplitude decays as  $\exp(-\beta k_0z)$ , the intensity will drop off as the square of this, as  $\exp(-2\beta k_0z)$ . The drop in intensity can be equated to a decay length or absorption length, which is

$$\Lambda = 1 / 2\beta k_0 \quad (\text{A.17})$$

$\beta$  is inversely proportional to the fourth power of the photon energy, while  $k_0$  is directly proportional to the photon energy. This leads to an absorption length proportional to  $(h\nu)^3$ . The absorption coefficient  $\mu = 1/\Lambda$  is therefore proportional to  $(h\nu)^{-3}$ . Thus, assuming that there are no absorption edges in between, the absorption length of the material will increase by a factor of 8, if the photon energy is doubled.

The absorption of a photon adds energy to the system, which is first channelled into the photo-electron. In principle, the electron can relax again down to its initial state, resulting in the emission of a photon with an energy equal to that of the initial radiation. However, the atom may often subtly reconfigure when the electron is excited. The state to which it then relaxes is not identical to the initial absorbing state, resulting in energy loss and inelastic scattering. This process is the basis for the Resonant Inelastic X-ray Scattering technique (RIXS).

If we assume another scenario, where the photo-electron is completely removed from the atom, leaving a core hole behind, another electron can relax to this hole. The excess energy can be released either as a fluorescent photon or through the ejection of yet another electron in a process called Auger emission. Finally, the ejected electrons can cause the secondary ejection of electrons from outside the parent atom in an avalanche process, resulting in the formation of an electron cascade, characterised by their having low kinetic energies, typically well below 1 eV.

Photo-electrons, fluorescence, and Auger emission are closely interrelated, see Fig. A.2. A photon is absorbed in parting enough kinetic energy to the core electron, which is fully ejected from the atom. This is a photo-electron. The core hole left behind can be filled by an electron from a shell further out, resulting in the emission of a photon with an energy equal to the difference between the relaxing electron and the core-level energies. This energy will be lower than that of the initial radiation. Auger emission is a three-electron process. As in fluorescence, a photo-electron is ejected, and an outer electron relaxes to fill the resulting core hole. Now, the difference in the energy gained by the relaxation of the second electron is not expended in fluorescence but instead in the ejection of a third electron from another orbital. The kinetic energy of this Auger electron is simply the relaxation energy of the second electron minus the binding energy of the Auger electron. It is, therefore, independent of the initial photon energy, unlike the directly ejected photo-electron.

Auger electron emission and X-ray fluorescence are competitive processes. The probability of spontaneous emission of radiation through electronic relaxation was shown by Einstein in 1916 to be proportional to the cube of the relaxation energy:

$$p_{fl} \propto (h\nu)^3 \quad (\text{A.18})$$

The probability for Auger emission is simply 1 minus that for fluorescence. The relatively high yield for Auger emission in light atoms has deleterious consequences in macromolecular crystallography. The dominant mechanism resulting in radiation damage of protein crystals is the secondary ionisation of atoms by Auger electrons, which have a cross-section significantly larger than the initial photo-electrons. The number of secondary electrons generated can exceed the number of photo-ionisation events by well over an order of magnitude, leading to the radical formation and the breaking of bonds. These secondary electrons partake in an avalanche-like process until the electrons become almost fully thermalized, and the cross-section for further capture and electron ejection becomes so small that the mechanism shuts down. Because so many more electrons can be released, the resulting current is often used as a sensitive probe for absorption processes, e.g., in many variants of X-ray absorption spectroscopy. Indeed, spatial variation in the amount of secondary electron emission across an irradiated sample can be imaged using an electron microscope in techniques like Photo-emission Electron Microscopy.



## X-ray sources

### B.1 Synchrotron radiation

#### B.1.1 Synchrotron radiation from a circular arc

A non-relativistic electron with momentum  $\mathbf{p} = m\mathbf{v}$  moving in a constant magnetic field  $\mathbf{B}$  experiences the Lorentz force  $\mathbf{F} = d\mathbf{p}/dt = -e\mathbf{v} \times \mathbf{B}$ . This force accelerates the electron, causing it to follow a circular trajectory with a radius  $\rho$  in a plane perpendicular to  $\mathbf{B}$ , resulting in bending magnet radiation. For non-relativistic particles, the Lorentz force equals the centripetal acceleration  $\frac{v^2}{\rho}m$ , yielding  $p = \rho eB$ . This is also valid for relativistic particles, where  $p = \gamma mv$ , with  $\gamma = E_e/mc^2$  representing the electron energy in units of the rest mass energy. In the case of super-relativistic particles, as encountered in synchrotrons, we have:

$$\gamma mc = \rho eB \tag{B.1}$$

so that in practical units, the radius of an electron orbiting in a synchrotron is defined as

$$\rho[m] = 3.3 \frac{E_e[GeV]}{B[T]} \tag{B.2}$$

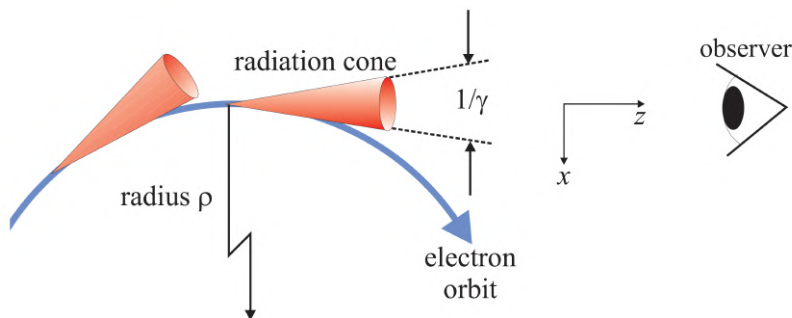


Figure B.1: It is an original figure taken from [39]. A relativistic electron is moving along a circular path of radius  $\rho$ . The emitted radiation is limited to a narrow cone with an opening angle of  $1/\gamma$  around the instantaneous velocity.

The electric field radiated by an accelerating charge is directly proportional to the apparent acceleration. Consequently, when an electron moves in a circular path, it experiences constant acceleration and radiates continuously throughout its orbit. However, relativistic charged particles in circular orbits produce highly

collimated radiation cones, as shown in Fig. B.1. Two key parameters determine the radiation's characteristics: the cyclic frequency  $\omega_0$  of the orbiting electron and  $\gamma = E_e/mc^2$ .

The radiation cone's direction aligns with the electron's instantaneous velocity, and its opening angle is  $\gamma^{-1} = mc^2/E_e$ , typically around  $10^{-4}$ . The emitted spectrum is wide, spanning from far infrared to hard X-rays. Nevertheless, it rapidly diminishes for photon frequencies above  $\gamma^3\omega_0$ . An electron's angular frequency  $\omega_0$  in the storage ring is approximately  $10^6$  cycles per second.

### B.1.2 The natural opening angle of synchrotron radiation

We will analyze the scenario where an electron travels at velocity  $v$  along a path composed of short straight segments, with abrupt bends at points A, B, C, etc., as depicted in Fig. B.2. Then, we will consider the limit in which these straight sections approach infinitesimal size, transforming the path into an arc of a circle.

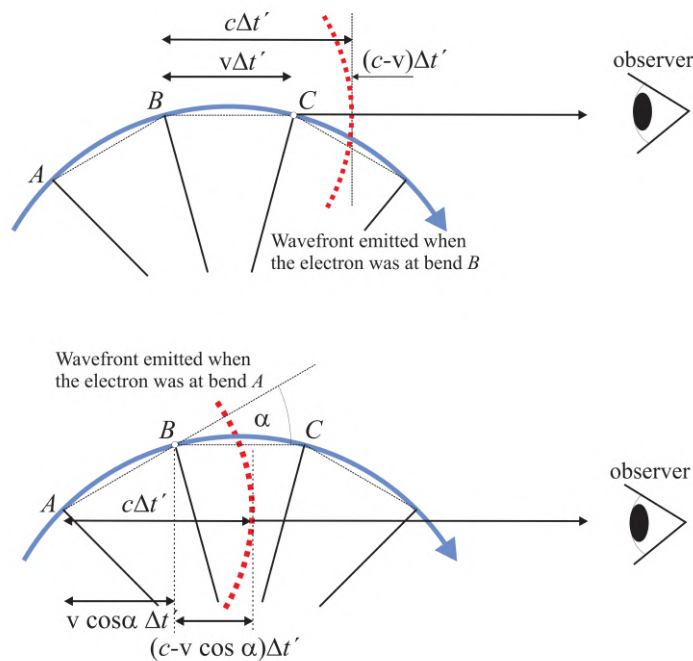


Figure B.2: This original figure is sourced from [39], where straight segments interconnected by bends at points A, B, C, etc., approximate a circular arc. The thick dotted line represents a wavefront. When the electron passes a bend, it emits a wavefront that propagates at velocity  $c$ . The electron travels at velocity  $v$ , taking a time interval  $\Delta t$  to move from one bend to the next. The observer experiences a time interval between wavefronts of  $\Delta t = (c - v \cos \alpha)\Delta t/c$ , with  $\alpha$  denoting the angle between the electron velocity and the direction toward the observer.

As the electron passes a bend, it emits a wavefront propagating at velocity  $c$ . Let  $\Delta t'$  represent the time for the electron to move from one bend to another, with the observer located along the BC direction. When the electron travels from B to C, the wavefront moves a distance of  $c\Delta t'$  toward the observer. A new wavefront is emitted from C, which is  $v\Delta t'$  closer to the observer than B. The same reasoning applies to the pair of wavefronts emitted when the electron was at A and B. The only difference is that the distance the electron travels towards the observer is  $v\Delta t' \cos \alpha$ , where  $\alpha$  is the angle between the velocity and the observer's direction. Therefore, the wavefront from A is not  $(c - v)\Delta t'$  ahead of the wavefront from B (as in the CB case), but at a distance of  $(c - v \cos \alpha)\Delta t'$ . In other words, the time compression of wavefronts, known as the Doppler effect, appears less pronounced to the observer. The observer perceives a time interval of  $\Delta t = (1 - \beta_e \cos \alpha)\Delta t'$



between wavefronts, where  $\alpha$  is the angle between the electron velocity and the direction toward the observer and  $\beta_e = v/c$ . As both  $\beta_e$  and  $\cos \alpha$  are close to unity, we can expand them, resulting in the following formula:

$$\Delta t \cong \left[ \frac{1 + (\alpha\gamma)^2}{2\gamma^2} \right] \Delta t' \quad (\text{B.3})$$

Here,  $\gamma = \frac{1}{\sqrt{1-\beta_e^2}}$ . The Doppler effect reaches its maximum when  $\alpha = 0$  and reduces by a factor of two when  $\alpha = 1/\gamma$ . This explains why the inherent opening angle of synchrotron radiation is approximately  $\gamma^{-1}$ .

While an electron emits radiation throughout its orbit, an observer positioned in the direction tangent to point B detects a significant amount of radiation only as the electron moves from A to C. This is because the amplitude of the far-field radiation is proportional to the apparent acceleration, which is exceptionally high when the electron is near B due to the substantial time compression described by Eqn. B.3.

### B.1.3 Characteristic frequency of synchrotron radiation

As an electron moves along the arc AC, it generates an intense pulse of radiation of finite duration  $\Delta t$ , which implies, from the general property of Fourier transforms, that there is a characteristic, cut-off frequency  $\omega_c \sim 1/\Delta t$ .

From the observer's perspective, the electron's motion resembles half the period  $T$  of a complete oscillation. The time taken by the electron to travel from A to C is  $[\gamma^{-1}/(2\pi)]T = 1/(\gamma\omega_0)$ , whereas the observer experiences this time as being roughly  $\gamma^2$  shorter (see Eqn. B.3). Consequently, the characteristic frequency  $\omega_c$  is approximately  $\gamma^3\omega_0$ , or more specifically,  $\omega_c = \frac{3}{2}\gamma^3\omega_0$ . Since  $\omega_0 = 2\pi/T = 2\pi/(2\pi\rho/c) = c/\rho$  and taking into account Eqn. B.2, the corresponding characteristic photon energy can be expressed in practical units as:

$$\hbar\omega_c[\text{keV}] = 0.665E_e^2[\text{GeV}]B[\text{T}] \quad (\text{B.4})$$

### B.1.4 Emitted power

The energy  $E_{rad}$ , radiated by the electron as it transits from A to C, equals to

$$E_{rad} = \frac{1}{4\pi} \frac{e^2}{4\pi\epsilon_0} \frac{\gamma^3}{\rho} \quad (\text{B.5})$$

The number of photons emitted by a single electron, denoted as  $N_{rad}$ , is roughly on the order of  $E_{rad}/\hbar\omega_c$ . Given that the characteristic energy  $\hbar\omega_c \sim \hbar(\gamma^3c/\rho)$ , we arrive at the following formula:

$$N_{rad} \sim \frac{1}{4\pi} \frac{e^2/(4\pi\epsilon_0)}{\hbar c} \quad (\text{B.6})$$

For a current  $I$  of electrons passing point A per second, the photon flux is approximately  $\frac{e^2}{4\pi\epsilon_0\hbar c} \frac{I}{e}$ . This reveals that a current of relativistic electrons passing through a bending magnet emits an immense flux of photons, on the order of  $10^{17}$  per Ampere, within an exceedingly narrow cone with an opening angle of  $1/\gamma$ .

Eqn. B.5 refers to the energy emitted from an electron path length of  $\rho/\gamma$ , so per unit length, the energy is  $\sim \gamma^4/\rho^2$ . Considering Eqn. B.2, we find that  $\rho \propto E_e/B$ , and as  $\gamma \propto E_e$ , this leads to a dependence of  $E_e^2B^2$ . The total radiated power in practical units can be expressed as:

$$P[\text{kW}] = 1.266E_e^2[\text{GeV}]B^2[\text{T}]L[\text{m}]I[\text{A}] \quad (\text{B.7})$$

Where  $L$  is the length of the electron trajectory through the bending magnet.

The spectral distribution of bending magnet radiation in the horizontal plane can be expressed as:

$$\frac{\text{Photons/second}}{(\text{mrad}^2)(0.1\%BW)} = 1.33 \times 10^{13} E^2 [\text{GeV}] I [\text{A}] x^2 K_{2/3}^2(x/2) \quad (\text{B.8})$$

Where  $x = \omega/\omega_c$  and  $K_{2/3}(x/2)$  is a modified Bessel function.

Consider bending magnet radiation at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France. The electron energy in the ESRF storage ring is  $E_e = 6$  GeV, the ring's electron current is approximately 200 mA, and the bending magnets generate a field of 0.8 T. The opening angle of the synchrotron beam from an ESRF bending magnet equals  $1/\gamma = m_e c^2 / E_e = 5.11 \times 10^5 / 6 \times 10^9 = 0.08$  mrad. Utilising Eqn. B.2, we calculate the electron orbit radius through the bending magnet:

$$\rho = 3.3 \times \frac{E_e [\text{GeV}]}{B [\text{T}]} = 3.3 \times \frac{6}{0.8} = 24.75 [\text{m}] \quad (\text{B.9})$$

Additionally, applying formula Eqn. B.4, we find the characteristic energy:

$$\hbar\omega_c [\text{keV}] = 0.665 \times E_e^2 [\text{GeV}] \times B [\text{T}] = 0.665 \times 6^2 \times 0.8 [\text{keV}] = 19.15 [\text{keV}] \quad (\text{B.10})$$

To determine the peak flux at the characteristic energy, we must multiply by the solid angle of the  $1 \times 1$  mm<sup>2</sup> aperture, the square of the electron energy, and the current. Assuming the aperture is located 20 m from the tangent point of the arc, its angular acceptance is 0.05 mrad. Consequently, the peak flux is given by:

$$\text{Flux} = 1.95 \times 10^{13} \times \left(\frac{1}{20}\right)^2 \times 6^2 \times 0.2 = 3.5 \times 10^{11} \text{photons/s}/0.1\%BW \quad (\text{B.11})$$

According to Eqn. B.7, the observed radiated power of the bending magnet depends on the length  $L$  of the electron orbit viewed through the aperture. When observing from the tangent point, we have  $L = \rho \times (\text{the aperture's horizontal plane acceptance angle}) = 24.8 [\text{m}] \times 0.05 [\text{mrad}] = 1.24 \times 10^{-3} [\text{m}]$ . Therefore, the radiated power is:

$$P = 1.266 \times E_e^2 \times B^2 \times L \times I = 1.266 \times 6^2 \times 0.8^2 \times 1.24 \times 10^{-3} \times 0.2 \times 10^{-3} = 7.3 [\text{W}] \quad (\text{B.12})$$

The observed power is lower than this value for a couple of reasons. First, the above value is an integration over the vertical direction, so we need to correct for the finite angular acceptance of the slit. Factors like beryllium vacuum windows in the synchrotron beamline and other components like filters may also dissipate power.

Summarising, the general properties of the radiation of a circular arc are as follows:

1. The radiation power is especially high at the moment when the instantaneous velocity of the electron points directly towards the observer since, at this moment, the Doppler effect is maximum.
2. This glimpse of radiation disappears when the angle between the direction towards the observer and the electron's velocity becomes of the order of  $\gamma^{-1}$ .
3. The typical frequency in the spectrum is  $\gamma^3$  times the cyclic frequency of an electron orbiting in a storage ring.
4. The on-axis radiation is linearly polarised in the horizontal plane, whereas the circular component is obtained out of the plane of the orbit with opposite spirals above and below the plane.
5. The radiation is pulsed, and the duration of the pulse, when viewed through a pinhole, is equal to the length of the bunch of electrons divided by  $c$ .

## B.2 Equipment for modern X-ray sources

### B.2.1 The magnet lattice

Storage rings employ three primary magnet types within their lattice: bending magnets, which deflect electron paths, and quadrupole and sextupole magnets, which are responsible for focusing and correcting chromatic aberrations. The precision and design of these magnets significantly impact the storage ring's brilliance. Typically, magnet lattices in storage rings are periodic and symmetric, forming multiple magnet cells that compose a "super period." The complete ring is comprised of several such super-periods.

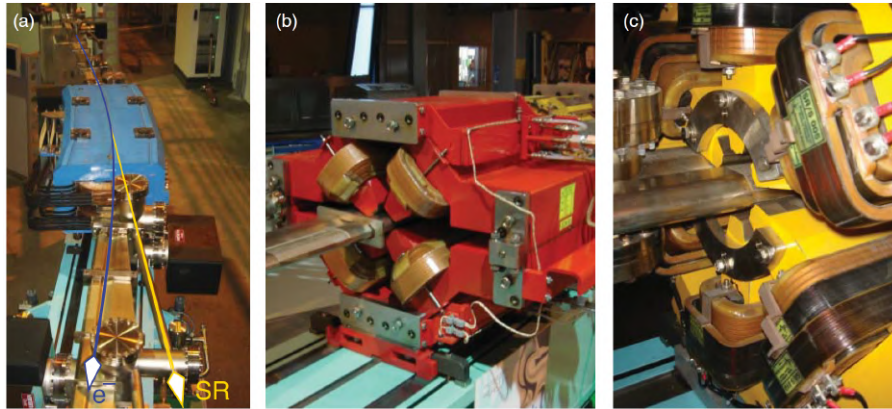


Figure B.3: The original photo is taken from [40]. The photo shows different elements of the magnet lattice: (a) bending or dipole magnets, (b) a focusing quadrupole magnet, and (c) a correcting sextupole magnet.

#### B.2.1.1 Bending Magnets and Superbends

The main purpose of bending magnets is to circulate the electron beam in the storage ring in a closed path. The typical magnetic field strength of bending magnets is about a Tesla. They produce radiation from a bending magnet in an oblate cone with a fan angle equal to the angle swept out in the path of the electrons due to the Lorentz forces to which they are subjected (plus a small additional amount due to the divergence of the photon beam, equal to  $\gamma^{-1}$ ). Due to the relatively large subtended angle of bending magnet radiation, measured in degrees, arranging more than one so-called 'bending magnet beamline' at a single bending magnet using two or more apertures is possible.

High brilliance and a high degree of monochromaticity are not needed for some types of experiment; for instance, in many computed tomography experiments, the sample might have to be illuminated by a uniform field of several square millimetres, while little is gained by highly monochromatizing the beam. In such cases, flux, not brilliance, matters a lot, and bending-magnet beamlines are very competitive. An even greater photon intensity can be obtained by attenuating the range of wavelengths the monochromator selects. This can be easily achieved using multilayer monochromators, which have a bandwidth of about ten to a hundred times greater than crystal monochromators.

#### B.2.1.2 Quadrupole and Sextupole Magnets

Quadrupoles consist of four magnets with alternating inward-pointing pole ends. In a perfectly aligned quadrupole, the magnetic field at the centre is zero but increases rapidly with radial distance from the central

axis. Quadrupoles set with a focus in the horizontal plane are called ‘F-quadrupoles’, and those placed so that the beam is defocused horizontally are called ‘D-quadrupoles’. If a D-quadrupole and an F-quadrupole are placed directly together without a gap, their fields cancel out, and the beam will not be focused. However, if they are separated by a carefully chosen distance, the combined effect is a total focus in both the horizontal and vertical planes.

The bending radius is proportional to the electron’s energy for a given transverse off-axis position inside the quadrupole. The effect of this is that those slightly higher/lower kinetic energy electrons will be less/more tightly focused, resulting in a longitudinal chromatic dispersion of the beam. The magnetic field of the sextupole magnets is such that all electrons that have sufficiently small deviations from the nominal energy of the storage ring are focused to the same point, thereby correcting the chromatic dispersion caused by the focusing pair of quadrupoles.

## B.2.2 Insertion devices

The third-generation synchrotrons were defined using insertion devices (IDs) placed in straight sections between arc segments of the bending magnet. IDs offered significantly higher fluxes and brilliance compared to bending magnets. They operate by inducing electron oscillations in the storage ring’s plane, employing dipoles that create alternating magnetic fields.

Two types of IDs can be distinguished based on the extent to which electrons deviate from a straight path. When angular deviations greatly exceed the natural opening angle  $\gamma^{-1}$ , the radiation cones from each ‘wiggle’ do not overlap. In such cases, the intensities add up, and the ID is called a wiggler.

On the other hand, an undulator features excursions on the order of  $\gamma^{-1}$ , causing the radiation cones from electrons to overlap as they slalom. Consequently, the dipoles’ radiations interfere with each other. In this scenario, field amplitudes are vectorially added (including phase differences), and their sum is squared to yield an intensity that peaks at wavelengths with constructive interference.

The maximum angular deviation  $\phi_{max}$  of electron oscillations in an ID is determined by the dimensionless ‘magnetic-deflection parameter’  $K$ , defined as:

$$\phi_{max} = K/\gamma \quad (\text{B.13})$$

Where  $K$  can be expressed in terms of the maximum magnetic field  $B_0$ :

$$K = \frac{eB_0}{m_e c k_{u,w}} = 0.934 \lambda_{u,w} [\text{cm}] B_0 [\text{T}] \quad (\text{B.14})$$

Where  $\lambda_u$  and  $\lambda_w$  are the periods of the oscillations in the undulator or wiggler, respectively, and  $k_{u,w} = \frac{2\pi}{\lambda_{u,w}}$ . For a wiggler,  $K$  is typically between 10 and 50, while for undulators,  $K$  is close to unity. The horizontal spread in the electron-beam divergence is:

$$\theta_x = 2K/\gamma \quad (\text{B.15})$$

The equations of motion of an electron with a trajectory  $\vec{\mathbf{r}}(t)$  in a magnetic field have the form

$$\frac{d\vec{\beta}(t)}{dt} = \frac{e}{mc\gamma} [\vec{\beta}(t) \times \vec{B}(\vec{\mathbf{r}}(t))] \quad (\text{B.16})$$

Here,  $e$ ,  $m$ ,  $\beta$ , and  $\gamma$  are the electron charge, mass, and reduced velocity and energy. Eqn. B.16 will be rewritten in terms of the components of the longitudinal coordinates:

$$\begin{cases} \frac{d\beta_x}{dt} = \frac{e}{mc\gamma} [B_z(z(t))\beta_y - B_y(z(t))\beta_z] \\ \frac{d\beta_y}{dt} = \frac{e}{mc\gamma} [B_x(z(t))\beta_z - B_z(z(t))\beta_x] \\ \frac{d\beta_z}{dt} = \frac{e}{mc\gamma} [B_y(z(t))\beta_x - B_x(z(t))\beta_y] \end{cases} \quad (\text{B.17})$$

Keep in mind the subsequent connection involving reduced velocities:

$$\begin{cases} \beta_x = \frac{v_x}{c} = \frac{1}{c} \frac{dx}{dt} \\ \beta_y = \frac{v_y}{c} = \frac{1}{c} \frac{dy}{dt} \\ \beta_z = \frac{v_z}{c} = \frac{1}{c} \frac{dz}{dt} \end{cases} \quad (\text{B.18})$$

By making certain mathematical rearrangements and incorporating Eqn. B.18, we can simplify Eqn. B.17 into the following expression:

$$\begin{cases} d\beta_x = \frac{e}{mc^2\gamma} [B_z dy - B_y dz] \\ d\beta_y = \frac{e}{mc^2\gamma} [B_x dz - B_z dx] \\ d\beta_z = \frac{e}{mc^2\gamma} [B_y dx - B_x dy] \end{cases} \quad (\text{B.19})$$

We can integrate Eqn. B.19 over the relevant range to determine how the reduced velocities vary with the corresponding coordinates.

Consider the insertion device with a magnetic field, described as

$$\vec{B} = (0, B_y(z), 0) \quad (\text{B.20})$$

where  $B_y(z) = B_0 \cos \frac{2\pi z}{\lambda_u}$ . After inserting Eqn. B.20 and Eqn. B.18 into Eqn. B.17, we will obtain the following system of differential equations:

$$\begin{cases} \frac{d^2x}{dt^2} = -\frac{e}{mc\gamma} B_0 \cos \frac{2\pi z}{\lambda_u} \frac{dz}{dt} \\ \frac{d^2y}{dt^2} = 0 \\ \frac{d^2z}{dt^2} = \frac{e}{mc\gamma} B_0 \cos \frac{2\pi z}{\lambda_u} \frac{dx}{dt} \end{cases} \quad (\text{B.21})$$

If we integrate  $\frac{d^2x}{dt^2} = -\frac{e}{mc\gamma} B_0 \cos \frac{2\pi z}{\lambda_u} \frac{dz}{dt}$ , we will get:

$$\frac{dx}{dt} = -\frac{eB_0\lambda_u}{2\pi mc\gamma} \sin \frac{2\pi z}{\lambda_u} \quad (\text{B.22})$$

Now we insert Eqn. B.22 into  $\frac{d^2z}{dt^2} = \frac{e}{mc\gamma} B_0 \cos \frac{2\pi z}{\lambda_u} \frac{dx}{dt}$  and obtain the following equation:

$$\frac{d^2z}{dt^2} + \frac{\lambda_u}{4\pi} \left( \frac{eB_0}{mc\gamma} \right)^2 \sin \frac{4\pi z}{\lambda_u} = 0 \quad (\text{B.23})$$

Equation Eqn. B.23 represents classical pendulum motion, solvable using the Jacobi function. By applying this solution from Eqn. B.23, we establish the relationship between the coordinate  $z$  and the time component. This relationship is then inserted into Eqn. B.22, and, with specific initial conditions, we derive the trajectory equation by eliminating the time variable:

$$\cos k_z z = \cosh k_z x - \frac{1}{\kappa} \sinh k_z x \quad (\text{B.24})$$

where  $\kappa^2 = \left( \frac{e}{mc^2\gamma\beta} B_0 \frac{\lambda_u}{4\pi} \right)^2$ . The trajectory equation given by the expression Eqn. B.24 differs little from a sinusoid.

### B.2.2.1 Wiggler

When analyzing an electron's trajectory, the wiggler can be considered a sequence of alternating left and right-turning circular arcs. This configuration enhances the observed radiation intensity by a factor of  $2N$ , with  $N$  representing the number of periods. Importantly, the maximum angular deviation from the wiggler axis surpasses the natural radiation opening angle  $\gamma^{-1}$ . The spectrum produced by a wiggler matches that of a bending magnet with the same field strength. Consequently, the equation describing emitted power closely resembles Eqn. B.7, but with one key difference. In the wiggler, the magnetic field strength fluctuates along the ID axis (dropping to zero between pairs of magnets), resulting in an absolute average field of  $B_0/\sqrt{2}$ , where  $B_0$  denotes the maximum magnetic-field strength. Consequently, Eqn. B.7 undergoes modification to:

$$P_w[kW] = 1.266E_e^2[GeV]B_0^2[T]L[m]I[A] \quad (\text{B.25})$$

The observed electron path, denoted as  $L$ , is approximately equal to the wiggler's length, typically around 1 m. Consequently, the radiated power can reach or exceed 1 kW. Such high thermal loads could potentially distort or damage the optical characteristics of the perfect crystals used to monochromate the X-ray beam. Various methods have been developed to preserve optical quality under these conditions.

The energy emitted by the wiggler increases as the gap between the top and bottom magnet sets decreases. However, as the gap size increases, the wiggler eventually transforms into an undulator. Nevertheless, wigglers are not designed for this purpose, and the radiative power, which is proportional to the square of the magnetic field (diminishing as the gap size increases), becomes impractically low.

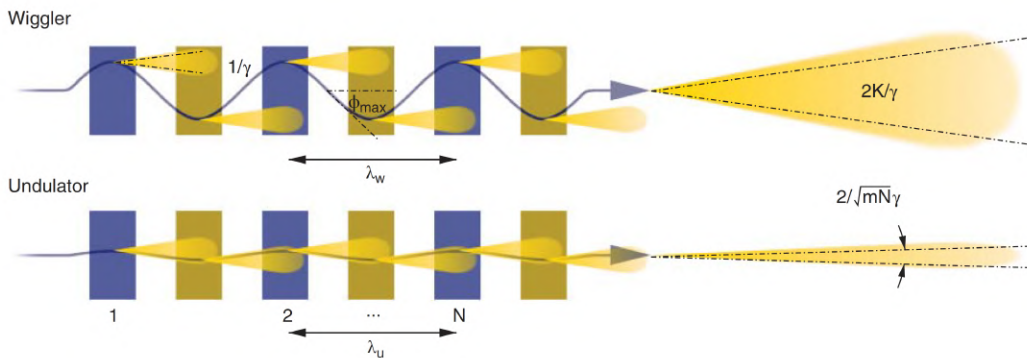


Figure B.4: The original image is sourced from [40], illustrating insertion devices such as wigglers and undulators. These devices primarily vary in the magnitude of angular deviations they induce in electron paths. Wigglers' maximum angular deviation  $\phi_{max}$  is considerably larger, scaled by a factor  $K$ , compared to the natural opening angle  $1/\gamma$ . In contrast, in undulators,  $\phi_{max}$  is proportional to  $1/\gamma$ .

### B.2.2.2 Undulator

An undulator is a specialised insertion device designed to ensure that the radiation emitted by an electron during one oscillation is synchronised with radiation from subsequent oscillations. This synchronisation causes the amplitudes of the radiated waves to add up first and then be squared to yield the resulting intensity. Due to overlapping and interference of radiation fields, undulators produce a spectral flux concentrated in evenly spaced narrow radiation bands compared to bending magnets and wigglers. Only specific wavelengths exhibit constructive interference, resulting in an undulator spectrum comprising a fundamental frequency and regularly spaced higher harmonics.

The key parameters for undulator radiation are the relativistic Lorentz factor  $\gamma$ , the undulator spatial period  $\lambda_u$ , and the undulator deflection parameter  $K$ , expressed as:

$$K = \frac{eB_0}{mck_u} \quad (\text{B.26})$$

Here,  $k_u = 2\pi/\lambda_u$ . Adjusting the gap size between the top and bottom magnet arrays allows for tuning the spectrum, ensuring that the spectral peak aligns with the desired photon energy.

The horizontal and vertical radiation divergences, in contrast to electron divergence, are determined by the expression:

$$\sigma_x^p = \sigma_y^p = \frac{1}{\gamma} \left[ \frac{1 + K^2/2}{2mN} \right]^{1/2} \approx 1/\sqrt{mN}\gamma \quad (\text{B.27})$$

Here,  $K \sim 1$ ,  $N$  is the number of periods in the undulator, and  $m$  is the harmonic number. For a typical undulator with one hundred or more poles, the horizontal spread in the electron-beam divergence, denoted as  $\theta_{x,y}^p$ , is approximately 10  $\mu\text{rad}$ .

However, the measured divergence at an undulator beamline exceeds that predicted by Eqn. B.27 because the observed emittance is a convolution of both the photon-beam and electron-beam emittances. For third-generation facilities, the electron emittance is typically much higher. In the case of DLSRs, electron emittance starts to dominate only at photon energies around the keV range.

The transition from wiggler to undulator radiation is not achieved by reducing lateral excursions through a decrease in magnetic field strength between pole pairs, as this would result in an unacceptable drop in flux. Instead, it is accomplished by decreasing the magnetic-pole spatial periodicity  $\lambda_u$ .

Now, let's derive the discrete wavelengths that lead to constructive interference in an undulator. In Fig. B.5, imagine that radiation is emitted at point A at time  $t' = 0$ . After a time equal to  $T'$ , the electron has moved to point B, completing one undulation downstream, and the radiation from A has travelled a distance of  $cT'$ . The difference between these distances is  $cT' - \lambda_u$ , and only radiation with a wavelength  $\lambda_m$  satisfying the following condition will experience constructive interference.

$$m\lambda_m = cT' - \lambda_u \quad (\text{B.28})$$

The fundamental wavelength  $\lambda_1$  satisfies Eqn. B.28 when  $m = 1$ .

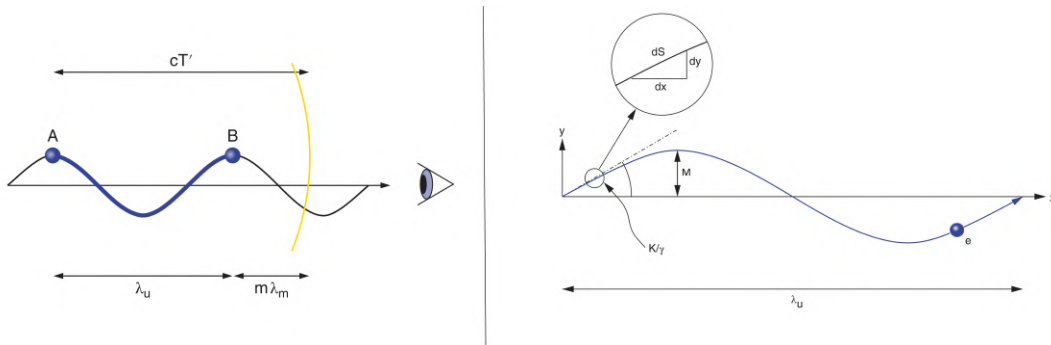


Figure B.5: This plot is sourced from [40]. The left image illustrates constructive interference between electromagnetic wavefronts originating from equivalent points on the undulations. The right image depicts the path  $S$  travelled by an electron with velocity  $v$  during one cycle in an undulator insertion device.

The path  $S = vT'$ , through which the electron passes in one cycle in the undulator, is equal to

$$S = \int_0^{\lambda_u} dS \quad (\text{B.29})$$

Using Pythagoras' theorem, we can derive the following

$$dS = \sqrt{dx^2 + dy^2} = \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{1/2} \approx \left[1 + \frac{1}{2}\left(\frac{dy}{dx}\right)^2\right] dx \quad (\text{B.30})$$

Also  $y = M \sin \frac{2\pi x}{\lambda_u}$ , thus,

$$\frac{dy}{dx} = \frac{2\pi M}{\lambda_u} \cos \frac{2\pi x}{\lambda_u} \quad (\text{B.31})$$

And  $dy/dx \ll 1$ . Recalling that for small angles, we have  $\tan \theta \approx \theta$ , hence, we obtain

$$\tan \frac{K}{\gamma} \approx \frac{K}{\gamma} = \frac{dy}{dx} \Big|_{x=0} = \frac{2\pi M}{\lambda_u} \cos \frac{2\pi x}{\lambda_u} \Big|_{x=0} \quad (\text{B.32})$$

Therefore, we get the equation for  $M$ :

$$M = \frac{K\lambda_u}{2\pi\gamma} \quad (\text{B.33})$$

Inserting Eqn. B.30, Eqn. B.31 and Eqn. B.33 into Eqn. B.29, we get

$$S = \int_0^{\lambda_u} \left[1 + \frac{K^2}{2\gamma^2} \cos^2 \left(\frac{2\pi x}{\lambda_u}\right)\right] dx \quad (\text{B.34})$$

Using trigonometrical formula

$$\cos^2 \alpha = \frac{1 + \cos 2\alpha}{2} \quad (\text{B.35})$$

we obtain

$$S = \lambda_u + \frac{K^2}{4\gamma^2} \int_0^{\lambda_u} \left[1 + \cos \left(\frac{4\pi x}{\lambda_u}\right)\right] dx = \lambda_u \left[1 + \frac{K^2}{4\gamma^2}\right] \quad (\text{B.36})$$

The condition for constructive interference is:

$$m\lambda_m = \frac{\lambda_u}{\beta} \left[1 + \frac{K^2}{4\gamma^2}\right] - \lambda_u \quad (\text{B.37})$$

Remembering that  $\beta \approx 1 - 1/2\gamma^2$  and  $\gamma^2 \gg 1$ , Eqn. B.37 can be simplified and look as follows

$$m\lambda_m = \frac{\lambda_u}{2\gamma^2} (1 + K^2/2) \quad (\text{B.38})$$

If we insert this into the equation that determines the wavelength of X-ray photons for a given photon energy ( $\lambda [\text{\AA}] = \frac{12.4}{E [\text{keV}]}$ ), we obtain

$$E_m [\text{keV}] = 0.95 \frac{mE^2 [\text{GeV}]}{(1 + K^2/2)\lambda_u [\text{cm}]} \quad (\text{B.39})$$

The interference spectrum at an angle  $\theta$  away from the central axis of the undulator is shifted toward lower energies and has the form:

$$m\lambda_m(\theta) = \frac{\lambda_u}{2\gamma^2} \left(1 + \frac{K^2}{2} + \gamma^2\theta^2\right) \quad (\text{B.40})$$

As mentioned, the undulator spectrum consists of narrow lines equally spaced in energy  $\Delta E$ , defined as follows.

$$\Delta E = \frac{2hc\gamma^2}{(1 + K^2/2)\lambda_u} \quad (\text{B.41})$$

The undulator spectrum can be adjusted by changing the parameter  $K$ , which is accomplished by altering the gap between the two sets of magnetic poles and, consequently, the magnetic field strength  $B_0$ . The spectral width of the undulator harmonics is influenced by the number of periods,  $N$ . Positive interference occurs when only a few waves are involved for a small relative frequency deviation from resonance. In contrast, at high values of  $N$ , constructive interference is confined to a very narrow frequency range. The monochromaticity, defined as  $\lambda_m/\Delta\lambda_m$ , is the inverse of the relative bandwidth and equals  $N$  multiplied by the harmonic number  $m$ .



## B.3 X-ray monochromators

Many experiments (except Laue diffraction) require a monochromatic beam in which well-defined values can be set for energy and bandwidth. In many cases, the radiation from insertion devices, even from a low-K undulator, is not monochromatic enough to be used as a source for experiments without further energy dispersion or monochromatisation (Fig. B.6).

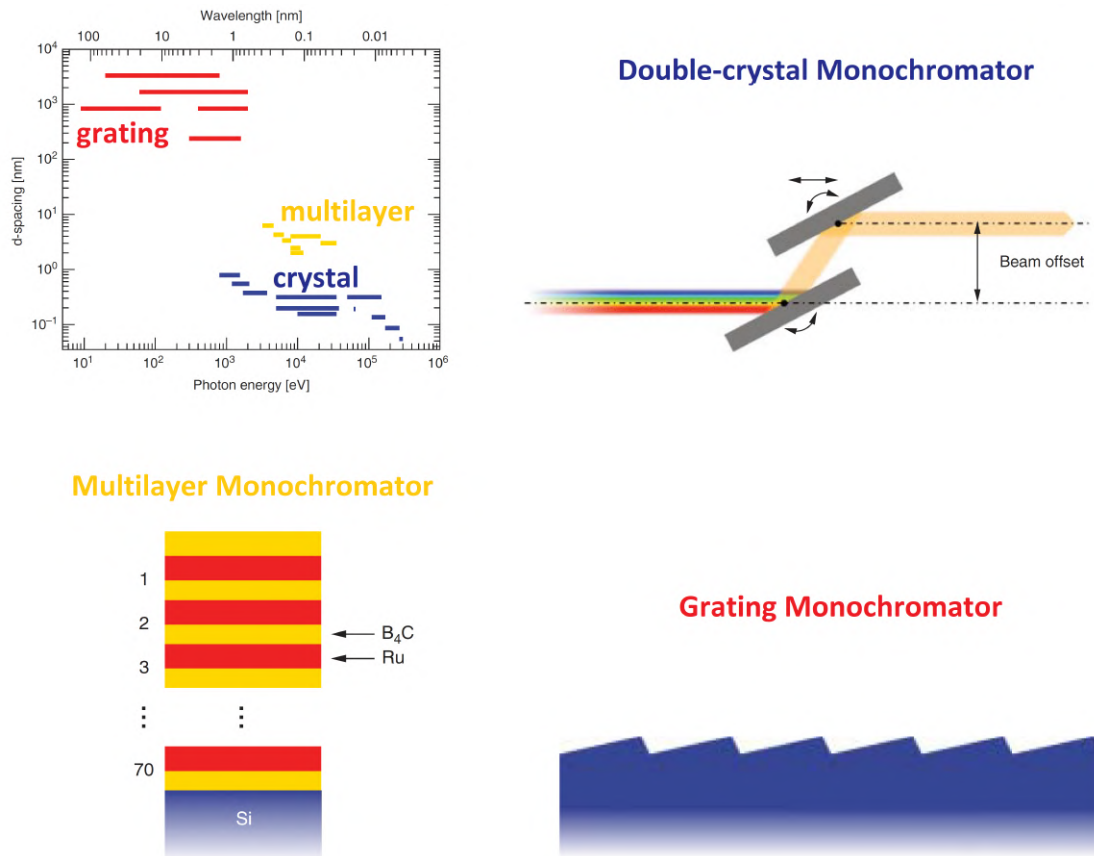


Figure B.6: This plot includes figures taken from [40]. A selection of grating (red), multilayer (yellow), and crystal (blue) monochromator element periodic spacings  $d$  and the typical energy range that they serve.

The primary function of an X-ray monochromator is to isolate specific X-ray energies from a continuous spectrum. This is typically achieved through X-ray diffraction, where the incident X-rays interact with the atomic lattice planes of a crystal. Bragg's law, which relates the angle of incidence, the wavelength of X-rays, and the spacing between lattice planes, governs this diffraction process. By selecting the appropriate angle, a monochromator can isolate a specific wavelength or energy from the dispersed X-ray spectrum. One of the significant advantages of X-ray monochromators is their tunability.

X-ray monochromators can be categorised into two main types: crystal monochromators and multilayer monochromators:

1. Grating monochromators: The profile of the grating grooves can vary. Blaze profiles are the most common, which display a saw-tooth-like profile. They work most efficiently when the incident and reflective beams reflect specularly on the blazing surface. For a grating evenly illuminated across  $N_f$  facets, which can be of the order of around  $10^4$ , the resolving power of the  $m^{\text{th}}$  harmonic is  $\frac{\lambda}{\Delta\lambda} = mN_f$ .

2. **Crystal Monochromators:** Crystal monochromators use perfect bulk-sized crystals. The most commonly used crystal material is silicon (Si) due to having reasonable thermal conductivity. It can be efficiently cooled by water or liquid nitrogen to minimise the mechanical strain induced by a local thermal bump where the incoming polychromatic beam impinges on the crystal surface. The most common crystallographic orientation is the (111) orientation. There are different types of crystal monochromators, for example:
- a) **Single Crystal Monochromators:** These monochromators use a single crystal to diffract X-rays and select a specific wavelength. The choice of the crystal depends on the desired X-ray energy range. Common crystal materials include silicon, germanium, and quartz.
  - b) **Double Crystal Monochromators:** In this setup, two crystals are used in tandem to achieve higher energy resolution and control. The first crystal selects a particular wavelength, while the second refines the beam.
  - c) **Channel-Cut Crystal Monochromators:** Channel-cut monochromator (CCM) are fabricated using just one single crystal, which has had a central channel milled out of it. This adds simplicity and speed to changing the photon energy, requiring a single rotation stage, a definite advantage when performing rapid energy scans in certain types of spectroscopic experiments. The disadvantage is that the exit beam height will change as the channel-cut monochromator is rotated.
3. **Multilayer Monochromators:** Crystals diffract X-rays through a scattering of planes of atoms. It is a three-dimensional system exploiting the periodicity in the direction parallel to the scattering vector. In the case of multilayers, scattering is produced by a contrast in the electron density at the interface between heavy reflection layers (such as ruthenium or tungsten) and a lighter space of layers (made of boron, carbon, or a combination of these). The only required periodicity is in one direction: the direction of the scattering vector. Some of the characteristics of multilayer monochromators can be seen below:
- a) The reflectivity can be very high. Coupled with a bandwidth which, maybe typically, is 100 times larger than that provided by single crystals, the flux output from multilayer monochromators can be very intense, which can be exploited by techniques such as X-ray computed Tomography.
  - b) The larger periodicity of multilayers can lead to refraction effects that can be exploited in suppressing harmonic contamination.
  - c) The strength of the reflection is determined partly by the ratio of the thicknesses of the thick layer to the thin layer. Two thick layers mean that the X-rays cannot penetrate deep enough into the multilayer, while two thin layers reduce the reflected intensity. Making the total thickness too large means that the X-rays cannot penetrate the bottom of the structure, obviating any advantages.
  - d) The layers should have atomically smooth interfaces. The selected wavelengths are approximately given by the modified Bragg equation  $m\lambda_m = 2\Lambda \sin \theta$ , where  $\Lambda$  is the multilayer periodicity.

The choice of monochromator depends on several factors, including the energy range of interest, the desired energy resolution, and the application's specific requirements. Crystal monochromators are often favoured for high-resolution experiments, while multilayer monochromators can offer broader tunability and efficiency.

---

# Improving data processing in protein crystallography

## C.1 Offline data processing pipeline for serial crystallography

I. Create a list of files for each run or block of runs and the folder structure by running the following script:

```
./auto_creating_list_of_files_with_folder_structure_for_processing.py
-i [path_to_raw_data]
-l [path_where_you_will_keep_all_list_of_files]
[-p pattern_in_filename, optional]
[-fe file_extension, optional]
[-b block_of_interest, optional]
[-r path_to_the_folder_for_creating_structure_for_processing]
```

II. Run the `turbo-index-slurm` script (<https://www.desy.de/~twhite/crystfel/scripts/turbo-index-slurm>) to process data with one geometry file for all datasets by running the following script:

```
./run_turbo_index.py
[path_to_the_folder_for_processing]
[path to the turbo-index-slurm script]
[-f block_of_interest]
[-r True to rerun indexing and pf with different parameters]
```

III. Use this script to check and merge streams. This script will also generate several plots for evaluation of such parameters as the detector centre:

```
./rerun-merge-detector-shift-v2.py
[path_to_the_folder_for_creating_structure_for_further_processing]
[-f block_of_interest]
[-pref prefix_for_merged_stream]
```

```
[-suf suffix_for_merged_stream]
[--s use this flag to skip folders with already merged stream]
[--r use this flag to rerun jobs on files that were not processed]
[--d Use this flag to run the detector-shift program]
```

The script above could also generate a geometry file for each run (block of runs) based on the results of the script `detector-shift`. In this case, datasets could be reprocessed with the corresponding geometry file as follows:

#### IV. Process data with the corresponding geometry file per each block of runs/single run:

```
./run_turbo_index-v2.py
[path_to_the_folder_for_creating_structure_for_further_processing]
[path_where_you_will_keep_all_list_of_files]
[path to the turbo-index script]
[-f block_of_interest]
[--r use this flag to rerun jobs on files that were not processed]
[-pg path_to_corresponding_geometries]
```

#### V. Use this script for running partiliator or create-mtz:

```
./partial-mtz.py
[path_to_the_folder_for_creating_structure_fthe or_further_processing]
[path_to_the_script_executed_partiliator_or_mtz]
[-f block_of_interest]
[--no-mtz use this flag to run partiliator]
```

#### VI. Accumulate all the results in Table 1 with overall statistics by running the following command:

```
./for_paper_table_generator.py
[folder with processed data]
[file_with_all_results]
```

#### VII. This script runs `compare_hkl` and `check_hkl` on the data and obtains results in Table 1 format.

```
./overallstatistics_with_new_cut_off.py
[absolute path to hkl file]
[-r high-res]
[-n number_of_shells]
```

An example of usage of `many_plots-upt-v2.py` program is here:

```

./many_plots-upt-v2.py
-i 1_CCstar.dat 2_CCstar.dat
-x '1/d'
-y 'CC*'
-o [name of your plot with the extension]
-add_nargs 1_Rsplit.dat 2_Rsplit.dat
-yad 'Rsplit/%'
-x_lim_dw number1
-x_lim_up number2
-t [put title]
-legend [put the legend here]
[--d, use this option if you want to show plots]

```

Before using the updated version of `vdsCsPadMaskMaker` program, the user must run the following command line.

```
g++ -shared -o SubLocalBG.so -fPIC SubLocalBG.c
```

And the following command is an example of how to execute `vdsCsPadMaskMaker`:

```

./maskMakerGUI.py
[filename for the HDF5 file]
[hdf5 path for the 2D cspad data]
[-g, the path to the CrystFEL geometry file for the image]
[-m, the path to the HDF5 file of the starting mask]
[-mp, path inside the HDF5 file of the starting mask]

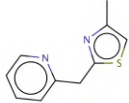
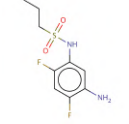
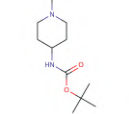
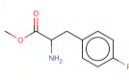
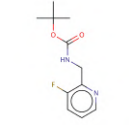
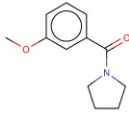
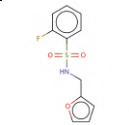
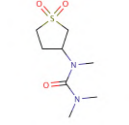
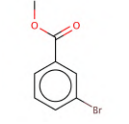
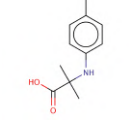
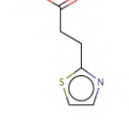
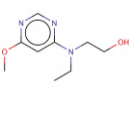
```

## C.2 Tool for generating per pattern mask for salt or ice reflections

Table C.2: Parameters of the experiment conducted at HiPhaX beamline PETRA III in July 2023

<b>run type</b>	grid fly
<b>step size, <math>\mu m</math></b>	15
<b>window size, <math>mm^2</math></b>	$4.22 \times 4.8$
<b>beam size, <math>\mu m</math></b>	7-8
<b>exposure time, ms</b>	40.0
<b>energy, keV</b>	16
<b>detector distance, mm</b>	213.6
<b>wavelength, <math>\text{\AA}</math></b>	0.775
<b>resolution limit, <math>\text{\AA}</math></b>	1.017
<b>Total number of windows per chip</b>	12

Table C.1: List of ligands of FAKP

Fragment 2D-structure	Well	Window	IUPAC Name
	D1	11	4-methyl-2-(pyridin-2-ylmethyl)-1,3-thiazole
	D2	10	N-(5-amino-2,4-difluorophenyl)propane-1-sulfonamide
	D3	9	tert-butyl N-(1-methylpiperidin-4-yl)carbamate
	D4	8	methyl 2-amino-3-(4-fluorophenyl)propanoate hydrochlorid
	D5	7	tert-butyl N-[(3-fluoropyridin-2-yl)methyl]carbamate
	D6	6	(3-methoxyphenyl)-pyrrolidin-1-ylmethanone
	D7	5	2-fluoro-N-(furan-2-ylmethyl)benzenesulfonamide
	D8	4	1-(1,1-dioxothiolan-3-yl)-1,3,3-trimethylurea
	D9	3	methyl 3-bromobenzoate
	D10	2	2-methyl-2-(4-methylanilino)propanoic acid
	D11	1	3-(1,3-thiazol-2-yl)propanoic acid
	D12	0	2-[ethyl-(6-methoxypyrimidin-4-yl)amino]ethanol

## C.3 Auto-processing pipeline for HiPhaX - a drug screening beamline P09, Petra III

### C.3.1 Google Sheets as an optimal database for monitoring results and saving metadata

In the following section, we present the fundamental steps to utilise the online interaction feature with Google Sheets. The complete manual is available at <https://github.com/galchenm/googleSheets>. Originally developed as a simple prototype for extracting Cheetah [133] results in real-time, this tool provides quick feedback on the processing status of the data, including hit rate, total frames, and hits. However, the `upt-cheetah-to-logbook-V2.py` script can be customised for specific requirements and implement additional features, as was done for the P09 beamline. To execute the main script, follow the instructions below:

```
./upt-cheetah-to-logbook-V2.py
[Google Sheet Name]
[/path_Cheetah/crawler.txt]
[file with information about fields]
```

Now we will inspect what all the parameters should look like. We start with [Google Sheet Name], see Fig. C.1.

	A	B	C	D	AL	AM	AN	AO	AP
1	Date	Time	Run #	Run duration [min]	RUN number	# Frames dark	HitsRate dark	# hits dark	indexed # dark
2	Parameters as discussed during kickoff meeting								
3									
4	5.11.2020		1	1	1		---	0	---
5			2	1	2		---	---	---
6			3	5	3		0.70	2972	---
7			4	7	4		0.93	8125	---
8			5	0					
9	06.11.2020		6		6		---	---	---
10			7		7		---	---	---
11			8		8		---	---	---
12			184		184		---	---	---
13			182		182		3.30	40567	---
14			108		108		1.93	493	---
15									
16									

Figure C.1: Google Sheet name here is UPTLOV@EUXFEL\_2020

Now we are going to look at the example of a generated file with results from Cheetah, see Fig. C.2

```
1 Run, Dataset, Rawdata, Cheetah, CrystFEL, H5Directory, Nprocessed, Nhits, Nindex, Hitrate%, Recipe, Calibration
2 1, dark-hg, Ready, Not started, ---, ---, 0, 0, ---, ---, ---
3 2, dark-mg, Ready, ---, ---, ---, ---, ---, ---, ---, ---
4 3, sample2_snr8, Ready, Finished, ---, r0003-sample2_snr8, 427326, 2972, ---, 0.70, agipd-snr8pix1.ini, calib-dk0002.ini
5 4, sample2_snr8, Ready, Finished, ---, r0004-sample2_snr8, 869928, 8125, ---, 0.93, agipd-snr8pix1.ini, calib-dk0002.ini
6 6, ---, Ready, ---, ---, ---, ---, ---, ---, ---, ---
7 7, ---, Ready, ---, ---, ---, ---, ---, ---, ---, ---
```

Figure C.2: Example of `/path_Cheetah/crawler.txt` file.

Another parameter required for `upt-cheetah-to-logbook-V2.py` is the [file with information about fields], the main idea of which is to pair the Cheetah field with Google Sheet field, an example of such a file

can be seen in Fig. C.3. It is necessary to create this file because users could name fields that they want to be filled with Cheetah results in different ways. So this part allows making this script more universal.

```
1 Run:Run #, RUN number
2 Nindex:indexed # dark
3 Hitrate%:HitsRate dark
4 Nhits:# hits dark
```

Figure C.3: Cheetah Field: Google Sheet field pairs

To work with Google Sheets, we need to have a token that we can obtain just by following the instructions given, for instance, here **Learn How to Use Python to Automate Google Sheets** or use the JSON file from <https://github.com/galchenm/googleSheets>. As a result, we will get the JSON file; see Fig. C.4.

```
1 {
2   "type": "service_account",
3   "project_id": "mypython-292614",
4   "private_key_id": "fea93cb3af60f8daf6fee619b3921ed56f07ebbe",
5   "private_key": "-----BEGIN PRIVATE KEY-----\nMIIEVgIBADANBgkqhkiG9w0B
6   "client_email": "test-980@mypython-292614.iam.gserviceaccount.com",
7   "client_id": "100557732721078573319",
8   "auth_uri": "https://accounts.google.com/o/oauth2/auth",
9   "token_uri": "https://oauth2.googleapis.com/token",
10  "auth_provider_x509_cert_url": "https://www.googleapis.com/oauth2/v1/
11  "client_x509_cert_url": "https://www.googleapis.com/robot/v1/metadata
12 }
```

Figure C.4: Cheetah Field: Google Sheet field pairs

This JSON file will allow us to have rights to work with Google Sheets, so we need to put the path to the JSON file in the code as follows in Fig. C.5:

```
29 scope = ["https://spreadsheets.google.com/feeds", 'https://www.googleapis.com/auth/spreadsheets']
30 creds = ServiceAccountCredentials.from_json_keyfile_name('client_secret.json', scope)
31 client = gspread.authorize(creds)
```

Figure C.5: Here, we have to put the absolute path to the JSON file

Before running the script, the user has to add `client_email` from the JSON file via the share feature in Google Sheets, shown at Fig. C.6.



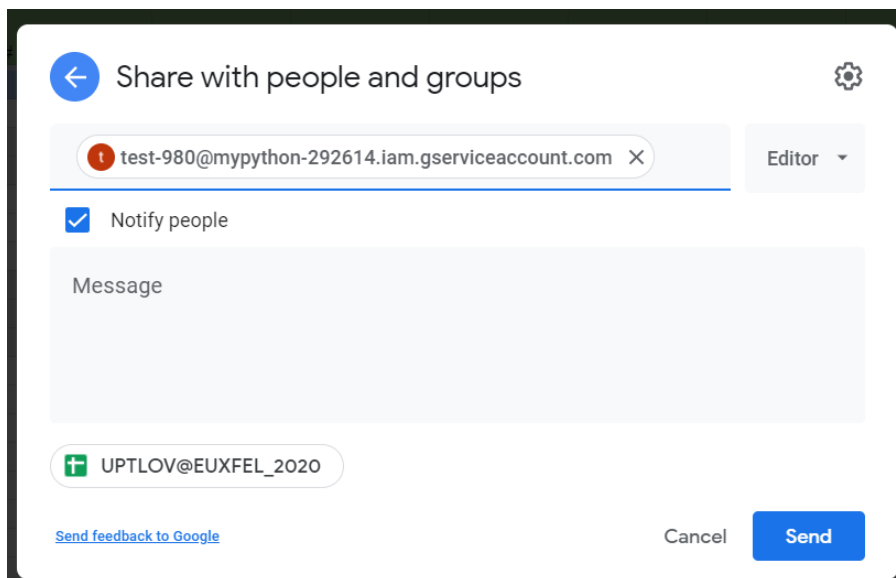


Figure C.6: An example of how to add `client_email` from the JSON file via the share feature in Google Sheets



---

# Enhancing data quality through modern data processing pipelines

## D.1 Re-processing old data, LCLS datasets

### D.1.1 Reprocessing previously collected data at LCLS in 2011

With new fast and megapixel x-ray detectors, the data rate becomes unbearable. The problem is that detectors and sources of radiation develop much faster than data storage systems. Modern facilities such as European X-ray Free Electron Laser (EuXFEL) or Swiss Free Electron Laser (SwissFEL), as well as even third generation synchrotrons, equipped with modern detectors such as Eiger 16M or Adaptive Gain Integrating Pixel Detector (AGIPD) [286], are already facing the problem of saving measured data. For example, around 12 years ago, an experiment at Linac Coherent Light Source (LCLS) with a 2.3 Mpix the Cornell-Stanford Pixel Area Detector (CSPAD) [269] produced up to 100Tb data during one 5 day experiment. The CSPAD, in particular, is a 64-segmented detector of application-specific integrated circuit (ASIC) modules which are bump-bonded into pairs. Now, an identical experiment performed at EuXFEL with 1 Mpix Adaptive Gain Integrating Pixel Detector (AGIPD) [286] with 176 active memory cells (half of the 352 designed cells) can produce 2.5PB of data. When AGIPD is used, the increase in data volume for one experiment compared to LCLS-I is about 50 times! And with the upcoming in the next couple of years AGIPD 4 Mpix or 2.1 Mpix (ePix) [285] 10k working at the 10kHz rate, the situation will become even worse – more than 10PB per experiment, which did not match the growth of the speed for the capacity of the hard drives. This trend suggests that raw measured data cannot be saved as before, and data conversion with reduction and compression has to be implemented. Considering how the data are read and converted from the detectors, it is usually not compressible in a lossless way. To get some lossless compression working, the data first has to be converted, for example, into single photons. This is a big problem because new detectors are usually not calibrated and characterised to reliably convert measured Arbitrary Detector Units (ADUs) into photons. So, an alternative is to use some lossy compression scheme.

And here, another concern appears. Experiments at FELs and often at modern synchrotrons are usually difficult and rather expensive. Therefore, people performing the experiments usually desire to save all the measured data, hoping to get the maximum out of it. Therefore, the experimental team usually objected to any deletion of the data, at least until some analysis was done or/and a paper was published. Even after this, some experimentalists hope to get more out of already measured data by using improved algorithms and better detector corrections.

All these questions are answered on the basis of the analysis of the data measured at the Coherent X-ray Imaging (CXI) beamline at LCLS in February 2011 and briefly described in Chapter 6. Many different samples were measured during the experiment, from simple samples with a small unit cell (UC) to membrane proteins with large UC parameters. This allows us to draw universal conclusions for different types of protein crystals.

To test the influence of modern processing pipelines, we have reprocessed the data measured during the experiment in 2011 (facility: LCLS, station: CXI, proposal: cxi21010, PI: S. Boutet). The experiment was performed at 120 Hz repetition rate, and the pulse length was 10fs and 40fs. A liquid jet was used as the sample delivery method. CSPAD 2.3M detector was utilised to capture diffraction patterns at full speed of LCLS. Several samples were measured during that experiment, mostly lysozyme, photosystem 1, cathepsin B, and a reaction centre. We reprocessed the data for all the mentioned samples, and the results were consistently better than those published shortly after the experiment [12, 264, 265].

The measured data was processed using the Cheetah program [133] to subtract the background and select only the diffraction patterns containing crystal diffraction (hit finding). The processed data set was deposited in CXI-DB <https://www.cxidb.org/id-17.html>. This data set was reprocessed using modern versions of CrystFEL v0.10.1 and Phenix/1.13. Also, the raw data were recovered from tape and re-processed using the modern version of Cheetah to get more hits than during the initial analysis. In addition, the subset of these data, which contains the same frames that were originally found, was chosen to test the modern pipeline. An example of the same pattern obtained using the original pipeline, deposited at CXI-DB (Fig. 6.5, a) and using the modern pipeline (Fig. 6.5, b) is presented. As can be seen from the Fig. 6.5.

The main problem was that this method was used for the whole pattern to remove the background that originated mainly from the liquid jet used as the sample delivery method. Such background filtering was quite justified, given that the Bragg peaks were rather sharp while the background was much smoother. But, as we can see now, this filtering degraded data quality (examples of patterns are presented in Fig. 6.5).

More interesting is comparing the same subset of diffraction patterns used in the original publication but converted with a recent pipeline to a new subset of patterns that can be obtained with a well-optimised modern hits-finding pipeline. As seen from Table 6.3, the latter method gives slightly more hits and indexed patterns, but the resulting data quality is almost the same. And, of course, all the results obtained using recent pipelines are much better than those produced in 2011 (see Table 6.2).

We have tested the same approach for other datasets measured over the same beam time and for some datasets from other experiments. It is mainly due to better knowledge of the detector geometry, better indexing, integration, and phase retrieval algorithms that we have obtained better results using a modern pipeline in all cases. All results are presented in Table D.1-Table D.17.

In summary, while for strongly diffracting crystals, it is worth saving only the hits, the 'raw' format can be used only in case the intensity calibration pipeline is improved. We could not draw a similar conclusion for weakly diffracting crystals; such a case requires additional studies.

Table D.1: Results of re-processed cathepsin B datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Ref.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.9.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>newLB,</b> <b>unity,</b> <b>0.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	6.8	0.998	0.999	11.072	100.000	1125.339	63M	56170	0.1795/ 0.1948	38.46
<b>newLB,</b> <b>unity,</b> <b>1</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	5.36	0.999	1.000	12.321	100.000	1390.224	78M	56170	0.1794/ 0.1988	37.69
<b>newLB,</b> <b>unity,</b> <b>1.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	4.7	0.999	1.000	12.959	100.000	1667.080	94M	56170	0.1783/ 0.1942	37.26
<b>newLB,</b> <b>xsphere,</b> <b>0.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	10.21	0.911	0.976	16.944	99.980	414.086	23M	56158	0.1796/ 0.2068	34.95
<b>newLB,</b> <b>xsphere,</b> <b>1</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	6.01	0.992	0.998	17.825	99.993	523.067	29M	56165	0.1797/ 0.2014	34.23
<b>newLB,</b> <b>xsphere,</b> <b>1.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	5.24	0.725	0.917	17.888	99.998	658.368	37M	56168	0.1775/ 0.1979	34.82

Table D.1: Results of re-processed cathepsin B datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Ref.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.9.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>noLB,</b> <b>unity,</b> <b>0.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	6.33	0.998	0.999	11.576	93.456	786.661	41M	52494	0.1830/ 0.1995	33.34
<b>noLB,</b> <b>unity,</b> <b>1</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	5.71	0.998	1.000	12.296	99.101	923.057	51M	55665	0.1808/ 0.1990	33.48
<b>noLB,</b> <b>unity,</b> <b>1.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	5.2	0.998	1.000	13.011	99.964	1104.667	62M	56150	0.1782/ 0.1940	33.16
<b>noLB,</b> <b>xsphere,</b> <b>0.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	4.38	0.999	1.000	19.302	84.066	334.137	15.7M	47219	0.1793/ 0.2002	33.91
<b>noLB,</b> <b>xsphere,</b> <b>1</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	4.21	0.999	1.000	18.099	95.380	376.045	20M	53574	0.1803/ 0.2017	32.47
<b>noLB,</b> <b>xsphere,</b> <b>1.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	4.1	0.999	1.000	17.704	99.389	441.538	24.6M	55826	0.1773/ 0.1897	32.85

Table D.1: Results of re-processed cathepsin B datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.9.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>oldLB,</b> <b>unity,</b> <b>0.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	8.43	0.995	0.999	11.454	99.820	927.548	52M	56069	0.2010/ 0.2107	37.33
<b>oldLB,</b> <b>unity,</b> <b>1</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	6.38	0.998	1.000	12.947	99.988	1163.995	65M	56163	0.1795/ 0.1888	33.54
<b>oldLB,</b> <b>unity,</b> <b>1.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	5.47	0.999	1.000	13.759	100.000	1414.968	79.5M	56170	0.1786/ 0.1916	32.69
<b>oldLB,</b> <b>xsphere,</b> <b>0.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	6.61	0.994	0.998	17.878	95.645	371.526	19.9M	53723	0.1921/ 0.2035	31.94
<b>oldLB,</b> <b>xsphere,</b> <b>1</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	5.84	0.997	0.999	18.197	99.087	466.389	25.9M	55656	0.1804/ 0.1978	29.86
<b>oldLB,</b> <b>xsphere,</b> <b>1.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	5.09	0.999	1.000	18.415	99.866	573.540	32M	56094	0.1783/ 0.1902	30.83

Table D.1: Results of re-processed cathepsin B datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.9.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>xfel,</b> <b>unity,</b> <b>0.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	7.26	0.996	0.999	12.673	97.184	1089.700	59M	54587	0.1832/ 0.2010	33.5
<b>xfel,</b> <b>unity,</b> <b>1</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	6.12	0.996	0.999	14.084	99.712	1337.493	74.9M	56007	0.1809/ 0.2030	32.2
<b>xfel,</b> <b>unity,</b> <b>1.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	5.18	0.998	1.000	14.985	99.989	1625.463	91M	56163	0.1783/ 0.1937	31.69
<b>xfel,</b> <b>xsphere,</b> <b>0.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	5.08	0.998	1.000	20.942	87.486	451.975	22M	49140	0.1837/ 0.2089	24.1
<b>xfel,</b> <b>xsphere,</b> <b>1</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	5	0.997	0.999	20.209	96.767	528.923	28.7M	54353	0.1803/ 0.1978	29.48
<b>xfel,</b> <b>xsphere,</b> <b>1.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	4.65	0.999	1.000	20.046	99.688	644.472	36M	55994	0.1779/ 0.1935	30.13



Table D.1: Results of re-processed cathepsin B datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.9.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>xfel,</b> <b>wo</b>	246k/	233k/	49.36										
<b>398-481,</b>	246k	252k	-	7.22	0.996	0.999	12.129	96.890	988.488	53.7M	54422	0.1837/ 0.2022	33.69
<b>unity,</b> <b>0.5</b>			1.60										
<b>xfel,</b> <b>wo</b>	246k/	233k/	49.36										
<b>398-481,</b>	246k	252k	-	6.18	0.997	0.999	13.522	99.655	1210.034	67.7M	55975	0.1813/ 0.2085	32.24
<b>unity,</b> <b>1</b>			1.60										
<b>xfel,</b> <b>wo</b>	246k/	233k/	49.36										
<b>398-481,</b>	246k	252k	-	5.3	0.998	1.000	14.397	99.984	1468.815	82M	56160	0.1787/ 0.1963	31.81
<b>unity,</b> <b>1.5</b>			1.60										



Table D.2: Results of re-processed cathepsin B datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.10.1</b> (model, push-res)													
<b>newLB,</b> <b>unity,</b> <b>0.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	6.950	0.998	0.999	11.404	100.002	1060.399	59.5M	56169	0.1814/ 0.2045	36.01
<b>newLB,</b> <b>unity,</b> <b>1</b>	282k/ 268k	251k/ /270k	49.36 - 1.60	5.310	0.999	1.000	12.769	100.002	1319.759	74M	56169	0.1806/ 0.1975	36.07
<b>newLB,</b> <b>unity,</b> <b>1.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	4.700	0.999	1.000	13.490	100.002	1593.059	89M	56169	0.1793/ 0.1987	36.08
<b>newLB,</b> <b>xsphere,</b> <b>0.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	12.380	0.715	0.913	18.053	99.868	389.161	21.8M	56094	0.1802/ 0.1930	31.71
<b>newLB,</b> <b>xsphere,</b> <b>1</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	7.480	0.788	0.939	18.968	99.979	492.728	27.6M	56156	0.1788/ 0.1996	32.75
<b>newLB,</b> <b>xsphere,</b> <b>1.5</b>	282k/ 268k	251k/ 270k	49.36 - 1.60	4.520	0.997	0.999	19.181	99.998	599.694	33.6M	56167	0.1795/ 0.1880	33.68

Table D.2: Results of re-processed cathepsin B datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.10.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>noLB,</b> <b>unity,</b> <b>0.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	6.040	0.998	0.999	12.170	90.327	777.766	39M	50735	0.1827/ 0.2069	32.13
<b>noLB,</b> <b>unity,</b> <b>1</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	5.580	0.998	1.000	12.739	98.576	888.201	49M	55368	0.1803/ 0.2120	33.14
<b>noLB,</b> <b>unity,</b> <b>1.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	5.120	0.998	1.000	13.348	99.945	1055.910	59M	56137	0.1766/ 0.2012	32.97
<b>noLB,</b> <b>xsphere,</b> <b>0.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	4.230	0.999	1.000	20.760	81.657	330.530	15M	45865	0.1801/ 0.2030	29.7
<b>noLB,</b> <b>xsphere,</b> <b>1</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	4.090	0.999	1.000	19.184	94.308	365.799	19M	52971	0.1800/ 0.1950	32
<b>noLB,</b> <b>xsphere,</b> <b>1.5</b>	157k/ 142k	141k/ 154k	49.36 - 1.60	4.020	0.999	1.000	18.387	99.179	426.521	23.7M	55707	0.1779/ 0.1879	32.45

Table D.2: Results of re-processed cathepsin B datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.10.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>oldLB,</b> <b>unity,</b> <b>0.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	8.070	0.992	0.998	11.746	99.247	903.677	50M	55745	0.1826/ 0.2061	31.86
<b>oldLB,</b> <b>unity,</b> <b>1</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	6.400	0.997	0.999	13.271	99.918	1131.551	63.5M	56122	0.1803/ 0.2065	32.52
<b>oldLB,</b> <b>unity,</b> <b>1.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	5.410	0.999	1.000	14.135	99.998	1378.421	77M	56167	0.1787/ 0.2058	32
<b>oldLB,</b> <b>xsphere,</b> <b>0.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	5.970	0.990	0.997	19.211	92.268	373.452	19M	51825	0.1824/ 0.1944	25.6
<b>oldLB,</b> <b>xsphere,</b> <b>1</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	5.440	0.998	1.000	18.987	98.326	453.471	25M	55228	0.1816/ 0.2052	28.71
<b>oldLB,</b> <b>xsphere,</b> <b>1.5</b>	307k/ 243k	229k/ 245k	49.36 - 1.60	4.960	0.999	1.000	19.136	99.785	554.352	31M	56047	0.1784/ 0.2007	29.47

Table D.2: Results of re-processed cathepsin B datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>xfel,</b> <b>unity,</b> <b>0.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	6.490	0.997	0.999	13.326	94.034	1082.106	57M	52817	0.1851/ 0.1976	31.31
<b>xfel,</b> <b>unity,</b> <b>1</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	5.760	0.998	1.000	14.518	99.307	1293.386	72M	55779	0.1808/ 0.1959	31.79
<b>xfel,</b> <b>unity,</b> <b>1.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	5.120	0.998	1.000	15.446	99.979	1569.118	88M	56156	0.1785/ 0.1954	31.3
<b>xfel,</b> <b>xsphere,</b> <b>0.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	4.740	0.998	1.000	22.787	83.970	453.206	21M	47164	0.1828/ 0.1981	25.88
<b>xfel,</b> <b>xsphere,</b> <b>1</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	4.760	0.995	0.999	21.356	95.424	516.453	27.6M	53598	0.1814/ 0.2005	29.35
<b>xfel,</b> <b>xsphere,</b> <b>1.5</b>	246k/ 246k	233k/ 252k	49.36 - 1.60	4.540	0.997	0.999	20.929	99.448	615.508	34M	55858	0.1781/ 0.1971	29.51

Table D.2: Results of re-processed cathepsin B datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL</b>	<b>Num.</b> frames/ hits	<b>Ind.</b> frames/ crystals	<b>Res.</b>	<b>R<sub>split</sub></b> (%)	<b>CC<sub>1/2</sub></b>	<b>CC*</b>	<b>SNR</b>	<b>Compl.</b> (%)	<b>Multipl.</b>	<b>Total</b> <b>Meas.</b>	<b>Unique</b> <b>Refl.</b>	<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	<b>Wilson</b> <b>B-factor</b>
<b>0.10.1,</b> <b>(model,</b> <b>push-res)</b>													
<b>xfel,</b> <b>wo</b>	246k/	233k/	49.36										
<b>398-481,</b>	246k	252k	-	6.530	0.997	0.999	12.868	93.457	987.099	51.8M	52493	0.1840/ 0.2131	31.51
<b>unity,</b> <b>0.5</b>			1.60										
<b>xfel,</b> <b>wo</b>	246k/	233k/	49.36										
<b>398-481,</b>	246k	252k	-	5.810	0.998	1.000	13.978	99.160	1173.240	65M	55696	0.1805/ 0.1919	31.91
<b>unity,</b> <b>1</b>			1.60										
<b>xfel,</b> <b>wo</b>	246k/	233k/	49.36										
<b>398-481,</b>	246k	252k	-	5.230	0.998	1.000	14.835	99.961	1420.493	79.7M	56146	0.1778/ 0.2029	31.43
<b>unity,</b> <b>1.5</b>			1.60										





Table D.3: Results of re-processed Dataset 1 from photosystem I collected in 2011 at LCLS with CSPAD detector

<b>Dataset 1, CrystFEL 0.10.1, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ hits</b>	115364/ 115363	115364/ 115363	115364/ 115363	115364/ 115363	115364/ 115363	115364/ 115363
<b>Ind. patterns/ crystals</b>	108633/ 117495	108633/ 117495	108633/ 117495	108633/ 117495	108633/ 117495	108633/ 117495
<b>Resolution</b>	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80
<b>R<sub>split</sub> (%)</b>	31.77	25.50	22.19	20.81	22.13	20.79
<b>CC<sub>1/2</sub></b>	0.723	0.885	0.922	0.897	0.916	0.930
<b>CC*</b>	0.916	0.969	0.979	0.972	0.978	0.982
<b>SNR</b>	5.042	5.173	5.539	7.915	6.087	6.250
<b>Completeness (%)</b>	90.549	99.999	100.000	75.706	99.623	99.999
<b>Multiplicity</b>	120.624	149.015	185.519	39.544	40.451	48.911
<b>Total Measurements</b>	40490051	55240193	68773497	11097854	14939017	18131300
<b>Unique Reflections</b>	335671	370703	370708	280649	369311	370703
<b>Wilson B-factor</b>	41.30	5.80	28.65	29.45	11.39	29.30

Table D.4: Results of re-processed Dataset 2 from photosystem I collected in 2011 at LCLS with CSPAD detector

<b>Dataset 2, CrystFEL 0.10.1, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ hits</b>	136009/ 136008	136009/ 136008	136009/ 136008	136009/ 136008	136009/ 136008	136009/ 136008
<b>Ind. patterns/ crystals</b>	127015/ 140248	127015/ 140248	127015/ 140248	127015/ 140248	127015/ 140248	127015/ 140248
<b>Resolution</b>	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80
<b>R<sub>split</sub> (%)</b>	33.19	20.48	19.15	22.47	19.32	18.40
<b>CC<sub>1/2</sub></b>	0.688	0.934	0.949	0.878	0.938	0.948
<b>CC*</b>	0.903	0.983	0.987	0.967	0.984	0.986
<b>SNR</b>	5.549	6.337	6.393	7.065	7.258	7.314
<b>Completeness (%)</b>	99.703	100.000	100.000	93.546	100.000	100.000
<b>Multiplicity</b>	209.566	279.789	314.301	67.070	97.800	115.584
<b>Total Measurements</b>	77456671	103720076	116513891	23258694	36255112	42847729
<b>Unique Reflections</b>	369605	370708	370708	346782	370706	370708
<b>Wilson B-factor</b>	36.05	1.12	27.92	47.89	7.94	25.71

Table D.5: Results of re-processed Dataset 3 from photosystem I collected in 2011 at LCLS with CSPAD detector

<b>Dataset 3, CrystFEL 0.10.1, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ hits</b>	62441/ 62441	62441/ 62441	62441/ 62441	62441/ 62441	62441/ 62441	62441/ 62441
<b>Ind. patterns/ crystals</b>	53197/ 58976	53197/ 58976	53197/ 58976	53197/ 58976	53197/ 58976	53197/ 58976
<b>Resolution</b>	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80	34.62 - 2.80
<b>R<sub>split</sub> (%)</b>	47.99	34.94	32.44	30.52	33.12	32.39
<b>CC<sub>1/2</sub></b>	0.603	0.821	0.859	0.8171254	0.8327917	0.849
<b>CC*</b>	0.867	0.950	0.961	0.948	0.953	0.958
<b>SNR</b>	3.639	4.168	4.229	5.106	4.372	4.382
<b>Completeness (%)</b>	99.825	100.000	100.000	81.802	99.815	100.000
<b>Multiplicity</b>	107.314	141.327	158.353	28.253	29.931	33.943
<b>Total Measurements</b>	39694990	52368192	58677002	8563852	11070256	12577140
<b>Unique Reflections</b>	369896	370545	370545	303111	369855	370542
<b>Wilson B-factor</b>	102.60	0.95	24.00	59.28	8.51	23.49

Table D.6: Results of re-processed old reaction centre (RC) datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL 0.9.1 old, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ Num. hits</b>	84k/ 23k	84k/ 23k	84k/ 23k	84k/ 23k	84k/ 23k	84k/ 23k
<b>Ind. patterns/ Ind. crystals</b>	2191/ 2439	2191/ 2439	2191/ 2439	2191/ 2439	2191/ 2439	2191/ 2439
<b>Resolution</b>	62.90 - 2.70	62.90 - 2.70	62.90 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70
<b>R<sub>split</sub> (%)</b>	49.04	48.31	48.18	45.47	44.93	44.91
<b>CC<sub>1/2</sub></b>	0.776	0.781	0.782	0.809	0.812	0.813
<b>CC*</b>	0.935	0.936	0.937	0.946	0.947	0.947
<b>SNR</b>	2.202	2.205	2.199	2.757	2.759	2.745
<b>Completeness (%)</b>	99.998	99.998	99.998	99.992	99.996	99.998
<b>Multiplicity</b>	68.150	72.234	74.016	35.529	37.599	38.284
<b>Total Measurements</b>	3.5M	3.7M	3.8M	1.8M	1.9M	1.9M
<b>Unique Reflections</b>	51952	51952	51952	51947	51949	51950
<b>R<sub>free</sub>/ R<sub>work</sub></b>	0.2362/ 0.3085	0.2403/ 0.3228	0.2389/ 0.3106	0.2368/ 0.3096	0.2372/ 0.3203	0.2359/ 0.3066
<b>Wilson B-factor</b>	61.94	58.08	61.31	60.46	56.29	59.32

Table D.7: Results of re-processed new reaction centre (RC) datasets with CrystFEL v0.9.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL 0.9.1 new, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ Num. hits</b>	18k/ 18k	18k/ 18k	18k/ 18k	18k/ 18k	18k/ 18k	18k/ 18k
<b>Ind. patterns/ Ind. crystals</b>	3744/ 4133	3744/ 4133	3744/ 4133	3744/ 4133	3744/ 4133	3744/ 4133
<b>Resolution</b>	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70
<b>R<sub>split</sub> (%)</b>	38.02	38.15	38.16	36.47	36.49	36.35
<b>CC<sub>1/2</sub></b>	0.861	0.860	0.860	0.866	0.866	0.868
<b>CC*</b>	0.962	0.962	0.962	0.964	0.963	0.964
<b>SNR</b>	3.158	3.120	3.115	3.853	3.833	3.831
<b>Completeness (%)</b>	100.0	100.0	100.0	100.0	100.0	100.0
<b>Multiplicity</b>	190.395	193.509	195.582	128.149	130.877	132.385
<b>Total Measurements</b>	9.9M	10.0M	10.1M	6.6M	6.8M	6.9M
<b>Unique Reflections</b>	51953	51953	51953	51953	51953	51953
<b>R<sub>free</sub>/ R<sub>work</sub></b>	0.2305/ 0.3088	0.2305/ 0.3058	0.2307/ 0.3043	0.2276/ 0.2947	0.2289/ 0.3084	0.2280/ 0.3134
<b>Wilson B-factor</b>	61.36	59.36	61.46	59.71	56.74	58.33

Table D.8: Results of re-processed old reaction centre (RC) datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL 0.10.1 old, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ Num. hits</b>	84k/ 23k	84k/ 23k	84k/ 23k	84k/ 23k	84k/ 23k	84k/ 23k
<b>Ind. patterns/ Ind. crystals</b>	2098/ 2312	2098/ 2312	2098/ 2312	2098/ 2312	2098/ 2312	2098/ 2312
<b>Resolution</b>	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70
<b>R<sub>split</sub> (%)</b>	62.00	61.90	62.33	62.56	62.45	63.00
<b>CC<sub>1/2</sub></b>	0.654	0.664	0.667	0.652	0.662	0.664
<b>CC*</b>	0.889	0.893	0.895	0.889	0.892	0.894
<b>SNR</b>	1.779	1.768	1.724	1.754	1.817	1.795
<b>Completeness (%)</b>	99.338	99.763	99.869	99.332	99.759	99.875
<b>Multiplicity</b>	14.505	15.618	16.187	14.566	15.707	16.231
<b>Total Measurements</b>	749k	809k	840k	752k	814k	842k
<b>Unique Reflections</b>	51606	51827	51882	51603	51825	51885
<b>R<sub>free</sub>/ R<sub>work</sub></b>	0.2641/ 0.3614	0.2623/ 0.3468	0.2589/ 0.3380	0.2617/ 0.3379	0.2603/ 0.3334	0.2583/ 0.3348
<b>Wilson B-factor</b>	60.94	53.85	56.46	58.12	51.13	53.81

Table D.9: Results of re-processed new reaction centre (RC) datasets with CrystFEL v0.10.1 with different partiality models and various resolution extensions (0.5, 1.0, 1.5)

<b>CrystFEL 0.10.1 new, (model, push-res)</b>	<b>unity, 0.5</b>	<b>unity, 1.0</b>	<b>unity, 1.5</b>	<b>xsphere, 0.5</b>	<b>xsphere, 1.0</b>	<b>xsphere, 1.5</b>
<b>Num. patterns/ Num. hits</b>	18k/ 18k	18k/ 18k	18k/ 18k	18k/ 18k	18k/ 18k	18k/ 18k
<b>Ind. patterns/ Ind. crystals</b>	3566/ 3900	3566/ 3900	3566/ 3900	3566/ 3900	3566/ 3900	3566/ 3900
<b>Resolution</b>	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70	57.50 - 2.70
<b>R<sub>split</sub> (%)</b>	51.87	51.40	51.55	52.48	51.66	51.80
<b>CC<sub>1/2</sub></b>	0.750	0.754	0.752	0.744	0.750	0.750
<b>CC*</b>	0.926	0.927	0.926	0.924	0.926	0.926
<b>SNR</b>	2.385	2.383	2.370	2.421	2.431	2.418
<b>Completeness (%)</b>	99.998	100.0	100.0	99.998	100.0	100.0
<b>Multiplicity</b>	34.419	35.454	35.590	34.254	35.680	35.978
<b>Total Measurements</b>	1.7M	1.8M	1.8M	1.7M	1.8M	1.8M
<b>Unique Reflections</b>	51949	51950	51950	51949	51950	51950
<b>R<sub>free</sub>/ R<sub>work</sub></b>	0.2452/ 0.3257	0.2454/ 0.3277	0.2447/ 0.3273	0.2449/ 0.3246	0.2458/ 0.3193	0.2465/ 0.3257
<b>Wilson B-factor</b>	55.32	51.87	52.38	52.52	49.44	49.75

Table D.10: Results of re-processed lysozyme, 5-fs pulses datasets with CrystFEL v0.10.1 with old geometry file, unity as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

model = unity, old geometry, (push-res)	10fs							
	new				old			
	0	0.5	1	1.5	0	0.5	1	1.5
<b>N. patterns/</b>	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/
<b>N. hits</b>	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k
<b>Ind. patterns/</b>	39k/	39k/	39k/	39k/	37k/	37k/	37k/	37k/
<b>Ind. crystals</b>	77k	77k	77k	77k	67k	67k	67k	67k
<b>Resolution</b>	35.33	35.33	35.33	35.33	35.33	35.33	35.33	35.33
	-	-	-	-	-	-	-	-
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub></b> (%)	5.42	5.38	5.36	5.4	5.69	5.65	5.61	5.66
<b>CC<sub>1/2</sub></b>	0.996	0.996	0.996	0.996	0.996	0.995	0.996	0.996
<b>CC*</b>	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
<b>SNR</b>	11.214	13.498	12.076	11.290	10.210	12.233	10.896	10.276
<b>Completeness</b> (%)	100.000	78.038	92.050	99.869	100.000	79.547	93.223	99.945
<b>Multiplicity</b>	971.450	956.945	931.407	932.364	902.395	915.741	888.845	889.710
<b>Total Measurements</b>	19M	14.8M	17M	18.5M	17.9M	14.5M	16.5M	17.6M
<b>Unique Reflections</b>	19875	15510	18295	19849	19875	15810	18528	19864
<b>R<sub>free</sub>/</b>	0.1836/	0.1763/	0.1843/	0.1861/	0.1928/	0.1862/	0.1931/	0.1938/
<b>R<sub>work</sub></b>	0.2101	0.2052	0.2129	0.2134	0.2210	0.2169	0.2247	0.2228

Table D.11: Results of re-processed lysozyme, 40-fs pulses datasets with CrystFEL v0.10.1 with old geometry file, unity as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

model = unity, old geometry, (push-res)	40fs							
	new				old			
	0	0.5	1	1.5	0	0.5	1	1.5
<b>N. patterns/</b>	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/
<b>N. hits</b>	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k



<b>Ind. patterns/ crystals</b>	65k/ 130.8k	65k/ 131k	65k/ 130.8k	65k/ 130.8k	65.8k/ 134.7k	65.8k/ 134.7k	65.8k/ 134.7k	65.8k/ 134.7k
<b>Resolution</b>	38.00	38.00	38.00	38.00	38.00	38.00	38.00	38.00
	-	-	-	-	-	-	-	-
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub> (%)</b>	4.19	4.65	4.4	4.24	4.39	4.58	4.45	4.38
<b>CC<sub>1/2</sub></b>	0.998	0.996	0.997	0.997	0.997	0.997	0.997	0.997
<b>CC*</b>	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
<b>SNR</b>	18.356	19.635	18.286	18.486	16.470	17.643	16.499	16.671
<b>Completeness (%)</b>	100.0	86.521	98.883	99.995	100.0	89.460	99.784	99.995
<b>Multiplicity</b>	1673.866	1413.931	1429.630	1552.446	1779.015	1535.492	1559.246	1677.108
<b>Total Measurements</b>	33M	24M	28M	30.8M	35M	27M	30.9M	33M
<b>Unique Reflections</b>	19876	17197	19654	19875	19876	17781	19833	19875
<b>R<sub>free</sub>/ R<sub>work</sub></b>	0.2038/ 0.2328	0.1946/ 0.2238	0.2042/ 0.2339	0.2018/ 0.2326	0.1911/ 0.2177	0.1874/ 0.2129	0.1947/ 0.2185	0.1952/ 0.2197

Table D.12: Results of re-processed lysozyme, 5-fs pulses datasets with `CrystFEL v0.10.1` with new geometry file, unity as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

<b>model = unity, new geometry, (push-res)</b>	<b>10fs</b>							
	<b>new</b>				<b>old</b>			
	<b>0</b>	<b>0.5</b>	<b>1</b>	<b>1.5</b>	<b>0</b>	<b>0.5</b>	<b>1</b>	<b>1.5</b>
<b>N. patterns/ N. hits</b>	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k
<b>Ind. patterns/ Ind. crystals</b>	39k/ 77k	39k/ 77k	39k/ 77k	39k/ 77k	37k/ 67k	37k/ 67k	37k/ 67k	37k/ 67k
<b>Resolution</b>	35.33	35.33	35.33	35.33	35.33	35.33	35.33	35.33
	-	-	-	-	-	-	-	-
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub> (%)</b>	5.130	5.110	5.030	5.080	5.850	6.320	5.960	5.860

<b>CC<sub>1/2</sub></b>	0.997	0.996	0.997	0.997	0.995	0.995	0.995	0.995
<b>CC*</b>	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
<b>SNR</b>	12.572	14.849	13.475	12.672	9.820	9.879	9.748	9.816
<b>Completeness (%)</b>	100.0	77.821	91.879	99.834	100.0	95.492	99.995	100.0
<b>Multiplicity</b>	884.898	883.135	857.760	856.556	1620.397	1360.657	1472.405	1576.929
<b>Total Measurements</b>	17.5M	13.6M	15.6M	16.9M	32M	25.8M	29M	31M
<b>Unique Reflections</b>	19875	15467	18261	19842	19876	18980	19875	19876
<b>R<sub>free</sub>/</b> <b>R<sub>work</sub></b>	0.1808/ 0.2016	0.1745/ 0.2009	0.1786/ 0.2025	0.1798/ 0.2008	0.1965/ 0.2192	0.1906/ 0.2207	0.2013/ 0.2265	0.1955/ 0.2193

Table D.13: Results of re-processed lysozyme, 40-fs pulses datasets with `CrystFEL v0.10.1` with new geometry file, unity as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

<b>model = unity, new geometry, (push-res)</b>	<b>40fs</b>							
	<b>new</b>				<b>old</b>			
	<b>0</b>	<b>0.5</b>	<b>1</b>	<b>1.5</b>	<b>0</b>	<b>0.5</b>	<b>1</b>	<b>1.5</b>
<b>N. patterns/</b>	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/
<b>N. hits</b>	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k
<b>Ind. patterns/</b>	65k/	65k/	65k/	65k/	65.8k/	65.8k/	65.8k/	65.8k/
<b>crystals</b>	130.8k	131k	130.8k	130.8k	134.7k	134.7k	134.7k	134.7k
<b>Resolution</b>	38.00	38.00	38.00	38.00	38.00	38.00	38.00	38.00
	-	-	-	-	-	-	-	-
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub></b> <b>(%)</b>	3.960	4.520	4.280	4.060	4.870	5.060	4.890	4.860
<b>CC<sub>1/2</sub></b>	0.998	0.996	0.997	0.998	0.996	0.996	0.996	0.996
<b>CC*</b>	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
<b>SNR</b>	21.003	21.673	20.749	21.027	14.949	14.657	14.847	14.912
<b>Completeness (%)</b>	100.000	86.859	98.954	99.995	100.000	100.000	100.000	100.000
<b>Multiplicity</b>	1597.837	1337.963	1356.425	1472.431	3209.134	2909.159	3068.636	3158.053

<b>Total Measurements</b>	31.7M	23M	26.6M	29M	63.7M	57.8M	60.9M	62.7M
<b>Unique Reflections</b>	19875	17264	19668	19875	19876	19876	19876	19876
$R_{\text{free}}/$ $R_{\text{work}}$	0.1965/ 0.2198	0.1935/ 0.2164	0.1965/ 0.2211	0.1925/ 0.2114	0.1983/ 0.2233	0.1990/ 0.2233	0.1952/ 0.2210	0.1980/ 0.2248

Table D.14: Results of re-processed lysozyme, 5-fs pulses datasets with CrystFEL v0.10.1 with old geometry file, xsphere as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

model =xsphere, old geometry, (push-res)	10fs							
	new				old			
	0	0.5	1	1.5	0	0.5	1	1.5
<b>N. patterns/</b> <b>N. hits</b>	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k	41.7k/ 41.6k
<b>Ind. patterns/</b> <b>Ind. crystals</b>	39k/ 77k	39k/ 77k	39k/ 77k	39k/ 77k	37k/ 67k	37k/ 67k	37k/ 67k	37k/ 67k
<b>Resolution</b>	35.33 - 1.50	35.33 - 1.50	35.33 - 1.50	35.33 - 1.50	35.33 - 1.50	35.33 - 1.50	35.33 - 1.50	35.33 - 1.50
$R_{\text{split}}$ (%)	4.570	4.310	4.330	4.410	4.790	4.540	4.590	4.670
$CC_{1/2}$	0.998	0.997	0.997	0.998	0.997	0.997	0.997	0.997
$CC^*$	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
<b>SNR</b>	12.588	16.309	14.297	12.901	11.539	14.900	12.936	11.689
<b>Completeness</b> (%)	99.995	74.687	88.302	98.596	99.995	75.653	89.127	98.933
<b>Multiplicity</b>	329.885	336.308	327.640	317.683	294.448	307.685	297.735	287.158
<b>Total Measurements</b>	6.5M	5M	5.7M	6M	5.8M	4.6M	5M	5.6M
<b>Unique Reflections</b>	19874	14844	17550	19596	19874	15036	17714	19663
$R_{\text{free}}/$ $R_{\text{work}}$	0.1732/ 0.2015	0.1674/ 0.1973	0.1736/ 0.2035	0.1784/ 0.2047	0.1789/ 0.2110	0.1742/ 0.2044	0.1822/ 0.2132	0.1899/ 0.2183

Table D.15: Results of re-processed lysozyme, 40-fs pulses datasets with `CrystFEL` v0.10.1 with old geometry file, `xsphere` as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

model = <code>xsphere</code> , old geometry, (push-res)	40fs								
	new				old				
	0	0.5	1	1.5	0	0.5	1	1.5	
<b>N. patterns/</b>	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/	75.9k/
<b>N. hits</b>	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k	75.6k
<b>Ind. patterns/</b>	65k/	65k/	65k/	65k/	65.8k/	65.8k/	65.8k/	65.8k/	65.8k/
<b>crystals</b>	130.8k	131k	130.8k	130.8k	134.7k	134.7k	134.7k	134.7k	134.7k
<b>Resolution</b>	38.00	38.00	38.00	38.00	38.00	38.00	38.00	38.00	38.00
	-	-	-	-	-	-	-	-	-
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub></b> (%)	3.630	3.730	3.620	3.630	3.600	3.630	3.550	3.550	3.550
<b>CC<sub>1/2</sub></b>	0.998	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998
<b>CC*</b>	0.999	0.999	0.999	0.999	1.000	0.999	1.000	1.000	1.000
<b>SNR</b>	20.887	23.573	21.454	21.148	19.704	22.369	20.342	20.121	20.121
<b>Completeness</b> (%)	100.000	83.769	96.710	99.935	100.000	86.294	98.566	99.985	99.985
<b>Multiplicity</b>	671.324	560.585	569.008	609.745	656.362	569.196	571.233	609.409	609.409
<b>Total Measurements</b>	13M	9M	10.9M	12M	13M	9.7M	11M	12M	12M
<b>Unique Reflections</b>	19876	16650	19222	19863	19875	17151	19590	19872	19872
<b>R<sub>free</sub>/</b>	0.1863/	0.1784/	0.1866/	0.1865/	0.1818/	0.1783/	0.1871/	0.1865/	0.1865/
<b>R<sub>work</sub></b>	0.2110	0.2043	0.2137	0.2131	0.2042	0.2025	0.2108	0.2090	0.2090

Table D.16: Results of re-processed lysozyme, 5-fs pulses datasets with `CrystFEL` v0.10.1 with new geometry file, `xsphere` as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

model = <code>xsphere</code> , new geometry, (push-res)	10fs								
	new				old				
	0	0.5	1	1.5	0	0.5	1	1.5	
<b>N. patterns/</b>	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/	41.7k/
<b>N. hits</b>	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k	41.6k

<b>Ind. patterns/ Ind. crystals</b>	39k/ 77k	39k/ 77k	39k/ 77k	39k/ 77k	37k/ 67k	37k/ 67k	37k/ 67k	37k/ 67k
<b>Resolution</b>	35.33 -	35.33 -	35.33 -	35.33 -	35.33 -	35.33 -	35.33 -	35.33 -
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub> (%)</b>	3.820	3.610	3.590	3.670	4.730	4.990	4.820	4.750
<b>CC<sub>1/2</sub></b>	0.998	0.998	0.998	0.998	0.997	0.997	0.997	0.997
<b>CC*</b>	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999
<b>SNR</b>	16.429	21.200	18.774	16.771	11.964	12.586	11.991	12.010
<b>Completeness (%)</b>	99.990	74.853	88.397	98.636	100.000	92.916	99.764	100.000
<b>Multiplicity</b>	323.338	329.136	320.419	310.920	596.514	525.213	545.039	577.128
<b>Total Measurements</b>	6M	4.8M	5.6M	6M	11.8M	9.6M	10.8M	11M
<b>Unique Reflections</b>	19873	14877	17569	19604	19876	18468	19829	19876
<b>R<sub>free</sub>/ R<sub>work</sub></b>	0.1694/ 0.1916	0.1609/ 0.1876	0.1656/ 0.1904	0.1682/ 0.1906	0.1896/ 0.2138	0.1890/ 0.2201	0.2019/ 0.2300	0.1888/ 0.2149

Table D.17: Results of re-processed lysozyme, 40-fs pulses datasets with `CrystFEL` v0.10.1 with new geometry file, `xsphere` as partiality model and various resolution extensions (0.0, 0.5, 1.0, 1.5)

<b>model = xsphere, new geometry, (push-res)</b>	<b>40fs</b>							
	<b>new</b>				<b>old</b>			
	<b>0</b>	<b>0.5</b>	<b>1</b>	<b>1.5</b>	<b>0</b>	<b>0.5</b>	<b>1</b>	<b>1.5</b>
<b>N. patterns/ N. hits</b>	75.9k/ 75.6k	75.9k/ 75.6k	75.9k/ 75.6k	75.9k/ 75.6k	75.9k/ 75.6k	75.9k/ 75.6k	75.9k/ 75.6k	75.9k/ 75.6k
<b>Ind. patterns/ crystals</b>	65k/ 130.8k	65k/ 131k	65k/ 130.8k	65k/ 130.8k	65.8k/ 134.7k	65.8k/ 134.7k	65.8k/ 134.7k	65.8k/ 134.7k
<b>Resolution</b>	38.00 -	38.00 -	38.00 -	38.00 -	38.00 -	38.00 -	38.00 -	38.00 -
	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
<b>R<sub>split</sub> (%)</b>	2.900	2.910	2.910	2.910	3.930	4.060	3.970	3.960

$CC_{1/2}$	0.999	0.998	0.999	0.999	0.998	0.998	0.998	0.998
$CC^*$	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999
<b>SNR</b>	28.005	32.799	29.171	28.665	18.778	18.600	18.789	18.854
<b>Completeness (%)</b>	100.000	84.322	96.850	99.889	100.000	99.995	100.000	100.000
<b>Multiplicity</b>	670.912	551.597	566.156	607.860	1302.244	1171.946	1233.848	1270.942
<b>Total Measurements</b>	13M	9M	10.8M	12M	25.8M	23M	24M	25M
<b>Unique Reflections</b>	19875	16759	19249	19853	19876	19875	19876	19876
$R_{\text{free}}/$ $R_{\text{work}}$	0.1749/ 0.1976	0.1727/ 0.1939	0.1717/ 0.1921	0.1731/ 0.1947	0.1893/ 0.2124	0.1906/ 0.2147	0.1907/ 0.2129	0.1906/ 0.2128

## D.1.2 X-ray Diffraction Analysis of Hemoglobin Samples at LCLS MFX

In Section 6.1.3 we discuss the collection and processing of crystallographic data. Here we will give more information about sample preparation and Table 1 with overall statistics.

Table D.18: Results of data processing hemoglobin datasets with 10 fs and 3 fs pulse duration.

<b>Runs</b>	73 - 76	99 - 101	102 - 110	180 - 182	<b>187</b> - <b>195</b>	210 - 216	227 - 228	<b>230</b> - <b>246</b>	270 - 271	283 - 284
<b>Sample name</b>	HS4	HS3	HS5	HS8	<b>HF2</b>	HF2	HF2	<b>HF2</b>	HF2	HF2
<b>Pulse duration</b>	10fs	10fs	10fs	10fs	<b>10fs</b>	3fs	3fs	<b>3fs</b>	3fs	3fs
<b>N. patterns/ hits</b>	111905/93438 (51513/49209)				130296/127365 (75133/73503)					
<b>N. indexed patterns/ crystals</b>	90484/120329 (47571/61182)				108851/129124 (64014/75958)					
<b>Resolution</b>	54.52 - 2.40				54.52 - 2.40					

Table D.18: Results of data processing hemoglobin datasets with 10 fs and 3 fs pulse duration.

<b>Runs</b>	73 - 76	99 - 101	102 - 110	180 - 182	<b>187</b> - <b>195</b>	210 - 216	227 - 228	<b>230</b> - <b>246</b>	270 - 271	283 - 284
<b>CC<sub>1/2</sub></b>	0.996 (0.994)				0.992 (0.988)					
<b>R<sub>split</sub></b> (%)	5.31 (6.35)				7.33 (9.22)					
<b>CC*</b>	0.998 (0.998)				0.998 (0.997)					
<b>SNR</b>	21.08 (17.14)				17.13 (13.44)					
<b>Completeness</b> (%)	97.62 (96.75)				97.01 (96.57)					
<b>Multiplicity</b>	435.00 (224.07)				466.37 (272.46)					
<b>Total Measurements</b>	9687964 (4945804)				10315340 (5998581)					
<b>Unique Reflections</b>	22271 (22073)				22118 (22016)					
<b>Wilson B-factor</b>	45.12 (45.32)				34.72 (35.77)					

### D.1.2.1 Crystal preparation

Equine hemoglobin (Sigma, CAS-9047090) stock solution of 8-10  $mgml^{-1}$  was prepared in 50 mM HEPES, pH 7.5 and precipitated using the stirred batch method [316] by mixing into 24- 26% v PEG3350 in a 1:1 ratio. The solution was kept at room temperature while continuously stirring. The crystals that appeared in about 2 h were filtered using 2  $\mu m$  stainless steel filters (Upchurch) and quenched with 25% PEG3350 for immediate use. Although hemoglobin usually crystallises in a monoclinic c-centred form, the crystals used here exhibited orthorhombic symmetry ( $P2_12_12_1$ ).

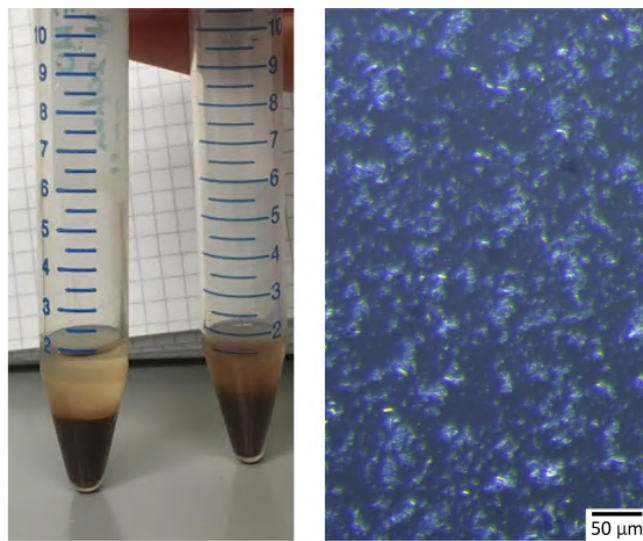


Figure D.1: The photographs displayed in this section are sourced from [266] and visually represent the hemoglobin crystallization process. The left image depicts the final outcome of the process, showing the hemoglobin sample after successful crystallization. In this image, the crystals are settled at the bottom of the falcon tubes, indicating the successful formation of well-defined crystals. The right image presents a micrograph of the crystalline slurry, revealing the presence of crystals with sizes as small as one micrometre. This micrograph provides a closer look at the individual crystals within the slurry, highlighting their size and structure. These photographs visually depict the hemoglobin crystallization process and the resulting crystal samples.



# Compression and data reduction in serial crystallography

## E.1 Existing lossless compressions and its evaluation

Table E.1: The evaluation of available lossless compressions on different datasets.

Lossless algorithm	AGIPD			Eiger 16M		
	CR, float	CR, int	CR, int; incr. step, N = 9 (512), with layers, lim_up = 32k	CR, int	CR, int; incr. step, N=0 (1), w/o layers	CR, int; incr. step, N=0 (1), with layers
<b>Bitshuffle</b>	1	1	1	1	1	1
<b>Bitshuffle + lz4</b>	1.103	3.345	8.25	3.581	9.939	7.805
<b>Blosc, blosclz, lev. 6 + bitshuffle</b>	1.118	2.837	6.361	3.571	10.476	7.897
<b>Blosc, blosclz, lev. 6 + shuffle</b>	1.095	1.88	3.453	2.194	8.388	5.325
<b>Blosc, blosclz, lev. 6</b>	1	2.335	2.091	1.986	3.781	2.48
<b>Blosc, blosclz, lev. 9 + bitshuffle</b>	1.118	2.837	6.361	3.571	10.476	7.897
<b>Blosc, blosclz, lev. 9 + shuffle</b>	1.095	1.927	3.453	2.267	8.388	5.325
<b>Blosc, blosclz, lev. 9</b>	1	2.335	2.091	1.986	3.781	2.48
<b>Blosc, lz4hc, lev. 6 + bitshuffle</b>	1.138	3.038	7.295	3.784	11.502	8.29

Table E.1: The evaluation of available lossless compressions on different datasets.

Lossless algorithm	AGIPD			Eiger 16M		
	CR, float	CR, int	CR, int; incr. step, N = 9 (512), with layers, lim_up = 32k	CR, int	CR, int; incr. step, N=0 (1), w/o layers	CR, int; incr. step, N=0 (1), with layers
<b>Blosc, lz4hc, lev. 6 + shuffle</b>	1.156	2.379	5.181	2.671	11.453	6.464
<b>Blosc, lz4hc, lev. 6</b>	1	3.06	5.298	2.275	7.235	4.147
<b>Blosc, lz4hc, lev. 9 + bitshuffle</b>	1.139	3.044	7.34	3.788	11.573	8.303
<b>Blosc, lz4hc, lev. 9 + shuffle</b>	1.162	2.471	5.744	2.672	12.636	6.479
<b>Blosc, lz4hc, lev. 9</b>	1	3.285	6.232	2.371	8.477	4.707
<b>Blosc, lz4, lev. 6 + bitshuffle</b>	1.119	2.907	6.424	3.654	10.453	8.019
<b>Blosc, lz4, lev. 6 + shuffle</b>	1.094	1.889	3.185	2.313	7.143	5.243
<b>Blosc, lz4, lev. 6</b>	1	1.639	2.374	1.508	2.872	1.726
<b>Blosc, lz4, lev. 9 + bitshuffle</b>	1.122	2.977	6.683	3.664	10.572	8.05
<b>Blosc, lz4, lev. 9 + shuffle</b>	1.096	1.9	3.196	2.411	7.16	5.508
<b>Blosc, lz4, lev. 9</b>	1	1.638	2.374	1.512	2.872	1.726
<b>Blosc, zlib, lev. 6 + bitshuffle</b>	1.155	3.096	8	3.785	12.524	8.561
<b>Blosc, zlib, lev. 6 + shuffle</b>	1.281	3.461	7.918	3.673	17.007	8.952
<b>Blosc, zlib, lev. 6</b>	1.14	4.339	8.461	3.094	12.958	6.809
<b>Blosc, zlib, lev. 9 + bitshuffle</b>	1.156	3.109	8.144	3.826	12.732	8.613
<b>Blosc, zlib, lev. 9 + shuffle</b>	1.284	3.504	8.416	3.676	17.613	8.954

Table E.1: The evaluation of available lossless compressions on different datasets.

Lossless algorithm	AGIPD			Eiger 16M		
	CR, float	CR, int	CR, int; incr. step, N = 9 (512), with layers, lim_up = 32k	CR, int	CR, int; incr. step, N=0 (1), w/o layers	CR, int; incr. step, N=0 (1), with layers
<b>Blosc, zlib, lev. 9</b>	1.14	4.244	9.353	3.126	13.978	6.989
<b>Blosc, zstd, lev. 6 + bitshuffle</b>	1.155	3.608	11.201	3.887	13.253	8.753
<b>Blosc, zstd, lev. 6 + shuffle</b>	1.279	3.453	8.235	3.841	17.568	9.409
<b>Blosc, zstd, lev. 6</b>	1.121	4.412	8.346	3.291	12.729	7.004
<b>Blosc, zstd, lev. 9 + bitshuffle</b>	1.161	3.63	11.449	3.943	13.547	8.855
<b>Blosc, zstd, lev. 9 + shuffle</b>	1.283	3.871	9.439	3.935	20.276	9.929
<b>Blosc, zstd, lev. 9</b>	1.193	5.022	10.972	3.619	16.87	8.547
<b>bzip2, lev. 6 + shuffle</b>	1.266	3.579	9.368	4.233	20.355	10.47
<b>bzip2, lev. 6</b>	1.227	5.918	12.006	4.209	19.7	10.35
<b>bzip2, lev. 9 + shuffle</b>	1.265	3.643	9.355	4.236	20.405	10.49
<b>bzip2, lev. 9</b>	1.235	5.923	12.029	4.216	19.752	10.38
<b>gzip, lev. 6 + shuffle</b>	1.28	3.448	7.887	3.726	17.214	9.233
<b>gzip, lev. 6</b>	1.14	4.506	8.465	3.182	13.077	6.865
<b>gzip, lev. 9 + shuffle</b>	1.283	3.504	8.439	3.729	18.264	9.243
<b>gzip, lev. 9</b>	1.14	4.408	9.542	3.229	14.427	7.169
<b>lz4, nbytes=0</b>	1	1	1	1	1	1
<b>lz4, nbytes=16384</b>	1	1	1	1	2.867	1.726
<b>lz4, nbytes=2048</b>	0.998	1.621	2.335	1.474	2.824	1.721
<b>lzf + shuffle</b>	1.106	2.028	3.747	2.509	8.051	5.662

Table E.1: The evaluation of available lossless compressions on different datasets.

Lossless algorithm	AGIPD			Eiger 16M		
	CR, float	CR, int	CR, int; incr. step, N = 9 (512), with layers, lim_up = 32k	CR, int	CR, int; incr. step, N=0 (1), w/o layers	CR, int; incr. step, N=0 (1), with layers
<b>lzf</b>	1	2.429	3.162	2.078	3.951	2.541
<b>zfp, reversible</b>	1	1	1	1	1	1
<b>zstd</b>	1.12	4.41	6.893	3.276	10.546	5.97

## E.2 Lossy compression

### E.2.1 Binning to lower the number of detector pixels

As mentioned before, it is necessary to use efficient compression before storing raw datasets in file systems to save disk space and, if possible, to improve the time spent on data processing by rejecting useless data (non-hits rejection). Moreover, if the distance between Bragg peaks is  $> 10$  pixels, it is possible to use a binning approach to reduce the amount of data, where each square of the pattern with a size of 2 by 2 pixels is summed, considering a bad pixel mask. We tested non-hit rejection and binning on data collected in November 2020 at the P11 beamline of Petra III with an Eiger 2X 16M detector (see Table E.2). Using only non-hits rejection without loss of data quality helped us to reduce the total data volume (89 TB) approximately three times (30TB), where binning could compress data 7 times. The raw data were deleted after running a non-hits rejection pipeline. While non-hits rejection has already been used during various SX experiments, binning can only be performed offline now. Binning requires well-optimised parameters for peak finding (for non-hits rejection). These parameters affect the resulting volume and the quality of the reduced data.

The binning procedure is part of modern Cheetah and OM software [317]. Advanced users could use this feature in two modes graphically introduced in Fig. E.2. Applying binning after peak finding lets us fully automatise this procedure by picking up the necessary parameters for the Cheetah template to perform binning.

### E.2.2 Quantization of detector output

One of the possible lossy approaches for data reduction can be quantisation in a constant step. The idea of such compression lies in applying simple arithmetic operations on the data, which will lead to discarding the data every equal step.  $R_{\text{free}}/R_{\text{work}}$  can be used as parameters for estimating data loss, and the compression rate is used to assess how much we can reduce such data without losing data quality. In this case, compression rate = reduced data volume / initial data volume. Several tests with different constant steps were performed on the AGIPD lysozyme data set. For reliability, estimates of quality change after applying such lossy compression were also carried out on a small part of the initial dataset, and  $R_{\text{free}}/R_{\text{work}}$  parameters were also calculated (see the results in Table 7.3). From Table 7.3, it can be wrongly concluded that the results for  $(\text{int div } 2e12) * 2e12$  data representation could bring the highest compression rate, but according to  $R_{\text{free}}/R_{\text{work}}$  one can

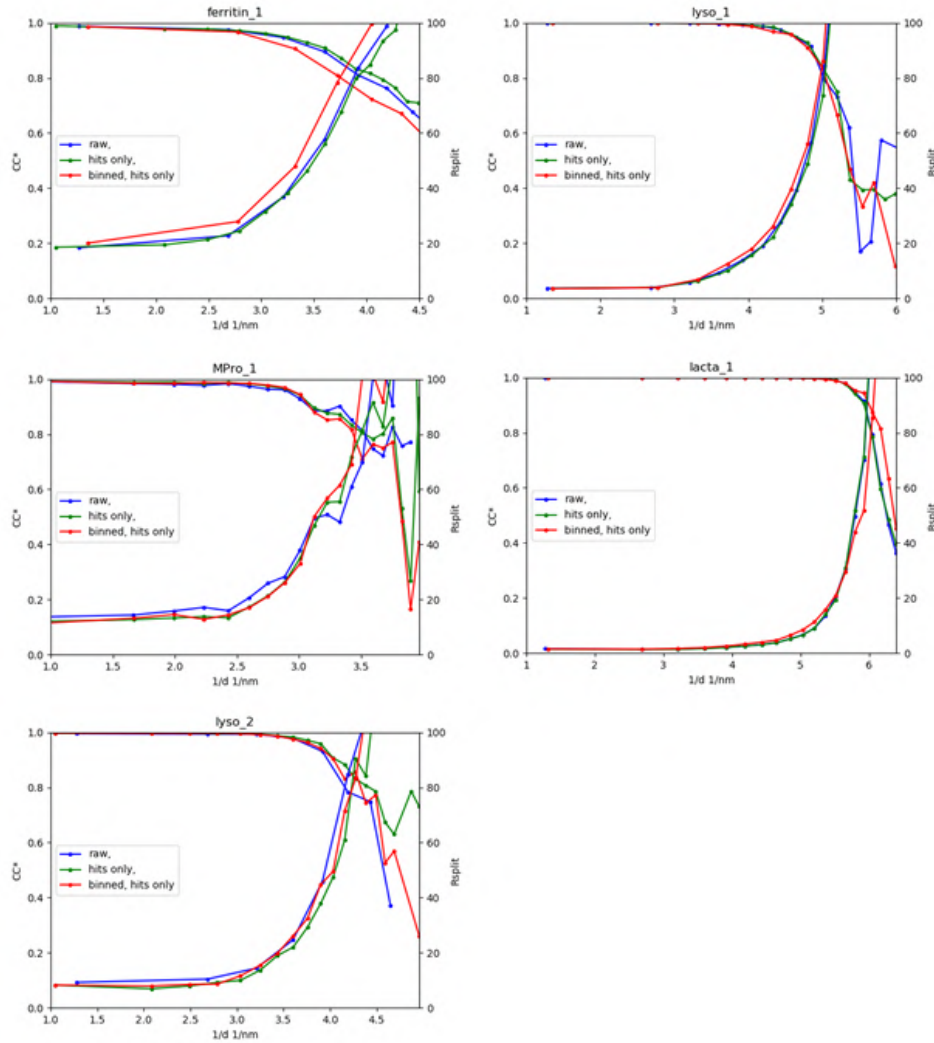


Figure E.1: The visual comparison of such metrics as  $CC^*$  and  $R_{split}$  for the data for four different samples: ferritin, lysozyme, SARS-CoV-2 Main protease (MPro), and lactamase described in Section 7.4.2.4. Blue curves correspond to the processing of only raw data, green – only patterns with determined crystal diffraction (hits), and the red curve – binned hits. The results are presented in Table E.2

see the degradation in data quality. Therefore, it is necessary every time to look at the following characteristics:  $CC^*/R_{split}$  metrics (see Fig. E.3),  $R_{free}/R_{work}$ , and visually look at the electron density. We also repeat the test of the dependence of the data quality on the number of diffraction patterns. The two approaches were compared: the reduction in the number of patterns versus the reduction in intensity precision (Table 7.4 and Fig. 7.5).

### E.2.3 Non-uniform quantisation

Revealing the detailed protein structure is crucial for understanding life processes at the molecular and atomic levels. The phases of each reflection are important for electron-density reconstruction. However, the collected diffraction patterns contain only the magnitude of the diffracted X-rays, and a consequence of phase information is lost. It is a so-called known phase problem. Also, some undesired artefacts can appear in the final electron maps because of intensity, phase errors, and/or incomplete datasets. Thus, the most difficult part of the structure

Table E.2: The results of applied non-hits rejection approaches (with/without binning) on data collected with TapeDrive 2.0 in November 2020 at P11, Petra III with Eiger 2X 16M detector

Subset sample		ferritin_1	lyso_1	MPro_1	lyso_2	lacta_1
<b>Raw</b>	<b>Vol.</b>	8.66T	3.7T	3.7T	260G	1.9T
	<b>Num. patterns /hits</b>	102.6k/6.8k	374k/62.7k	400k/8.9k	40k/23k	200k/198k
	<b>Ind. patterns /crystals</b>	6.5k/7k	42.8k/61.8k	5.4k/5.6k	5.8k/6.5k	187.7k/50.7k
<b>Raw, only hits</b>	<b>Vol.</b>	134.2G	721G	192G	193G	1.8T
	<b>Num. patterns /hits</b>	15k/6.6k	75.8k/62k	20.9k/13.7k	30.6k/27.8k	199.6k/198k
	<b>Ind. patterns /crystals</b>	6k/6k	42.5k/61k	8k/8.9k	7.9k/9k	18.8k/50.5k
	<b>CR</b>	66.08	5.25	19.73	1.35	1.06
<b>Binned, only hits</b>	<b>Vol.</b>	84.2G	310G	42G	85G	711G
	<b>Num. patterns /hits</b>	31.9k/10.5k	97.8k/60.6k	11.8k/11.8k	33.8k/29.1k	199.7k/199.7k
	<b>Ind. patterns /crystals</b>	7k/7k	42.7k/62k	8.6k/9k	9k/10.8k	188.9k/55.5k
	<b>CR</b>	105.32	12.22	90.21	3.06	2.74

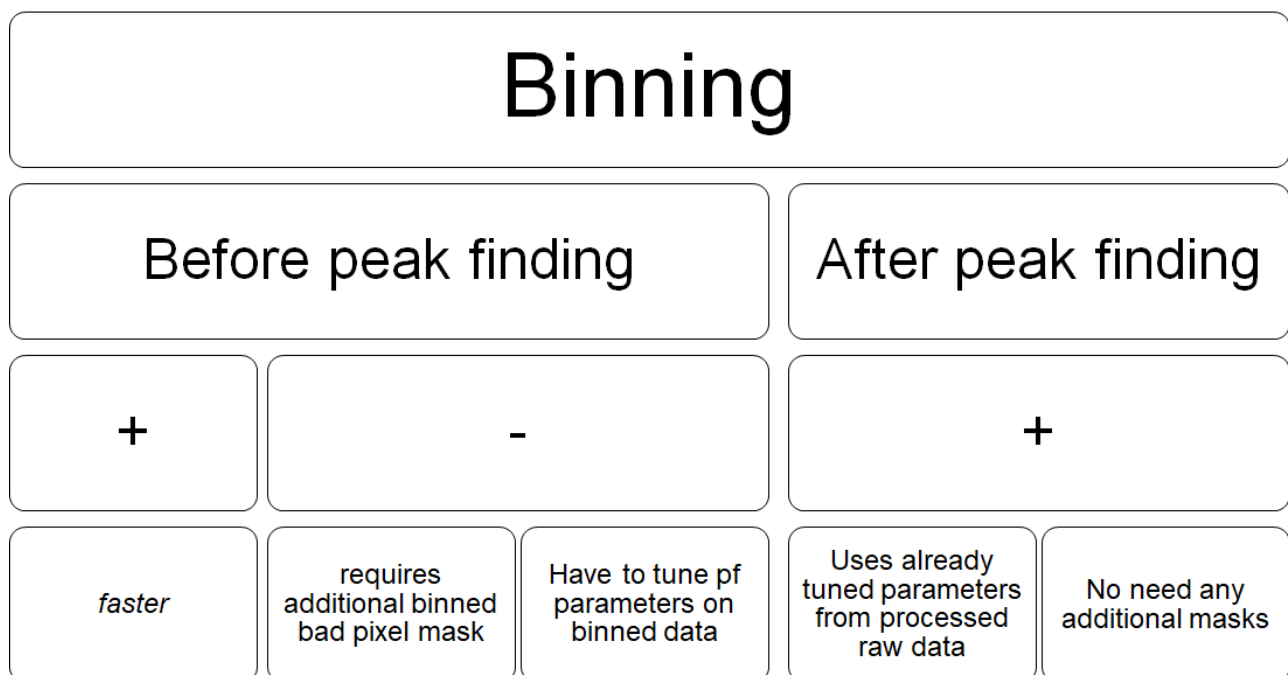


Figure E.2: Two modes of binning procedure described briefly in Section 7.4.2.4 and being a part of modern Cheetah and OM software [317]

determination step is obtaining the corrected missing phases. Several ways can help deal with such issues:

1. molecular replacement (when we have a previously available structurally similar model for further

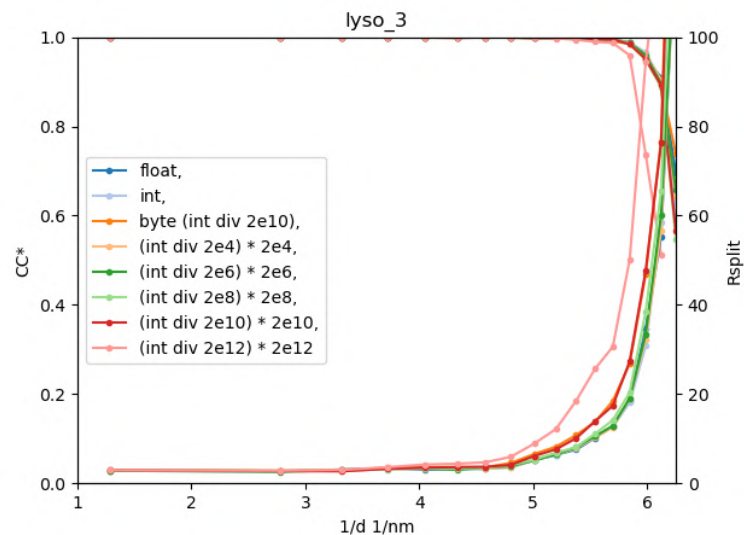


Figure E.3: Data quality ( $CC^*$  and  $R_{split}$ ) for the datasets rounded to the different power of two.

calculation initial phases

2. experimental phasing (in single-wavelength anomalous diffraction (SAD) case, it is possible to exploit intensity differences between corresponding Friedel pairs)

Although lossless compression does not spoil the data, applying lossy compression to the dataset requires checking the final data quality. Data loss should not result in the inability to reconstruct the protein structure and should not give much worse characteristics of data sets. Nevertheless, the question of reliable data quality metrics for biological data sets is still open in scientific society. Therefore, when working with biological samples, one should avoid losing important information for the entire structure refinement pipeline during the phasing steps. Thus, all newly developed techniques that can lead to a reduction in the amount of data should be carefully tested throughout the whole pipeline, which will end with some known metrics and will be able to overcome the phase problem. Data from different experiments at synchrotrons and FELs with different detectors and samples were used to investigate the influence of various data reduction approaches on data quality.

Moreover, we tested compression techniques on the SAD dataset [249] (thaumatin collected at 4.57 keV measured at SwissFEL with the JUNGFRÄU 16M detector), which is much more sensitive to data quality. It allowed us to determine the application limits for most compression algorithms. We have also tested different data-saving precision, leaving 1, 2 and 3 the most significant bits. The resulting volume can be found in Table 7.6 in Section 7.4.2.6, and overall statistics can be found in Table E.3. For the substructure detection and initial model building, we used the CRANK2 pipeline. According to Table 7.6, sequentially applied lossy compressions, where a lot of information is lost, can lead to the fact that it will be impossible to solve the phase problem, which in turn means failure in protein reconstruction.

Table E.3: Overall statistics for SAD dataset of thaumatin (measured at SwissFEL with JUNGFRAU 16M detector): original and reduced in different ways.

	<b>raw</b>	<b>binned</b>	<b>binned, 3 sign bits</b>	<b>binned, 2 sign bits</b>	<b>binned, 1 sign bit</b>
<b>Num. patterns/ hits</b>	52207/ 52207	52207/ 51906	52207/ 51784	52207/ 51597	52207/ 52184
<b>Ind. patterns/ crystals</b>	50844/ 59004	47929/ 53635	47499/ 53040	46221/ 51264	26965/ 28171
<b>Resolution, Å</b>	25.78 - 2.42	25.78 - 2.42	25.78 - 2.42	25.78 - 2.42	25.78 - 2.42
<b>R<sub>split</sub> (%)</b>	5.97	6.35	6.65	6.81	7.94
<b>CC<sub>1/2</sub></b>	0.993	0.994	0.993	0.992	0.991
<b>CC*</b>	0.998	0.998	0.998	0.998	0.998
<b>CC<sub>ano</sub></b>	0.327	0.320	0.247	0.271	0.251
<b>SNR</b>	15.993	14.713	13.520	13.252	10.528
<b>Completeness (%)</b>	89.794	88.142	88.591	88.257	88.773
<b>Multiplicity</b>	287.504	251.600	276.584	264.276	172.362
<b>Total Measurements</b>	4952834	4254810	4701093	4474988	2935670
<b>Unique Reflections</b>	17227	16911	16997	16933	17032
<b>Wilson B-factor</b>	132.16	194.1	157.77	111.7	133.63

### E.2.3.1 Chunk summation of diffraction patterns

Dimensional reduction is a well-known method for reducing the number of variables or dimensions in a data set while retaining the maximum amount of information possible. This can be done by removing redundant, irrelevant, or noisy data. Generally, such an approach is commonly used in machine learning and data science to make data easier to work with and analyse. It can also be used to visualise high-dimensional data in a lower-dimensional space. Typical examples of dimensional reduction are feature selection and feature extraction. Binning, discussed in Section 7.4.2.4, is another dimensional reduction technique. However, this approach is rapidly developing, and an overview of existing methods can be found in this work [318]. The benefits of dimensional reduction include reducing the complexity of the data, improving computational efficiency, reducing



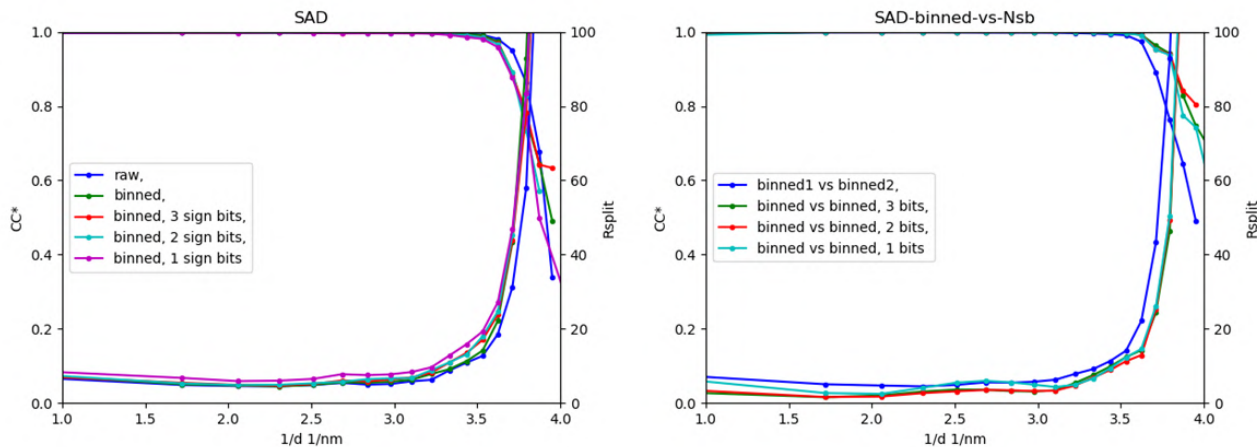


Figure E.4: The metrics  $R_{split}$  and  $CC^*$  for a) raw, binned and rounded to 1,2,3 the most significant bits; b) between binned data and rounded. The blue curve corresponds to the correlation of 2 halves of the binned dataset while all other curves - between the whole binned dataset and whole rounded datasets.

the risk of over-fitting, and improving the interpretability and visualisation of the data. However, it is important to note that dimensional reduction can also lead to information loss, so it should be used cautiously and evaluated carefully. The typical example of dimensional reduction widely used in data sets collected with X-rays is a calculation of the radial curve.

Data reduction typically involves processing diffraction patterns obtained from X-ray or electron diffraction experiments in crystallography. These diffraction patterns contain a large amount of data, and the goal of data reduction is to extract the necessary information about the crystal structure while reducing noise and eliminating redundancies in the data.

One possible approach we could refer to as a strategy for reducing dimensional data in crystallography is to identify the symmetry of the crystal lattice. This involves analysing the diffraction pattern to identify the symmetry operations that leave the pattern unchanged. The symmetry operations can then reduce the number of independent measurements required to determine the crystal structure. This principle is used in conventional crystallography.

Another technique for reducing the dimensions of data is principal component analysis (PCA). PCA is a statistical technique that can be used to identify the most significant features, in the case of crystallography, the features of the diffraction pattern. When analysing the principal components of the data, it is possible to reduce the dimensionality of the data while retaining the essential information necessary for structure determination. Generally, PCA has an application in Machine Learning (ML) to play the role of a peak finder to label non-hits. Other ML algorithms developed based on feature extraction used for the classification of diffraction patterns [319–321].

The `indeximajig` has a feature called `--multi`, which enables multi-lattice indexing. This method involves removing the blemishes associated with a successful indexing solution before attempting indexing again to find a second lattice. Using this feature, we came up with the idea of summing several patterns of datasets from SFX experiments to test a simple strategy for reducing high-dimensional data. We expected such a strategy could reduce the total data volume with negligible loss. Although the approach aims at combining lists of already found Bragg peaks from several patterns into one, a simple addition leads to a deterioration in the signal-to-noise

ratio, which significantly affects the indexing and further merging stages; see Tables E.4 and Fig. E.5. For the test, we have chosen the lysozyme data set described in detail in [107] and evaluated the hypothesis of the influence of the number of chunks of interest and the influence of several peaks in the newly defined pattern on the final quality of the data. Based on the results shown in Table E.4, we can conclude that this approach can give us an additional compression ratio. Still, the data evaluation showed significant distortion at low and high resolution compared to the original data. The reason could also be that tested data were preliminary binned, and during multi-indexing, many patterns could be rejected. To examine the latter assumption, we compared the quality metrics of the dimensionally reduced data with a chunk size of 3 with a randomly selected every third pattern from the total dataset. Based on the results shown in Fig. E.6, we can conclude that multi-indexing dramatically helps improve the final data quality results. Still, a high signal-to-noise ratio leads to degraded data quality. Therefore, dimensional data reduction cannot be used for SX data in the current state. However, such methods can be applied to other multidimensional datasets collected at synchrotron and FEL facilities.

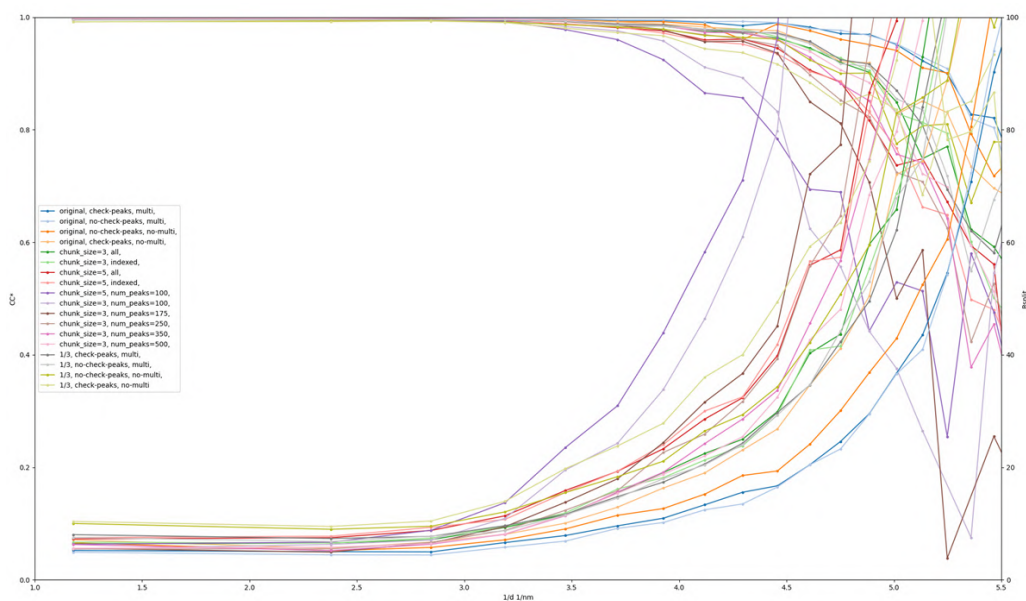


Figure E.5: Different tests of dimensional reduction on a dataset collected with TapeDrive [107].

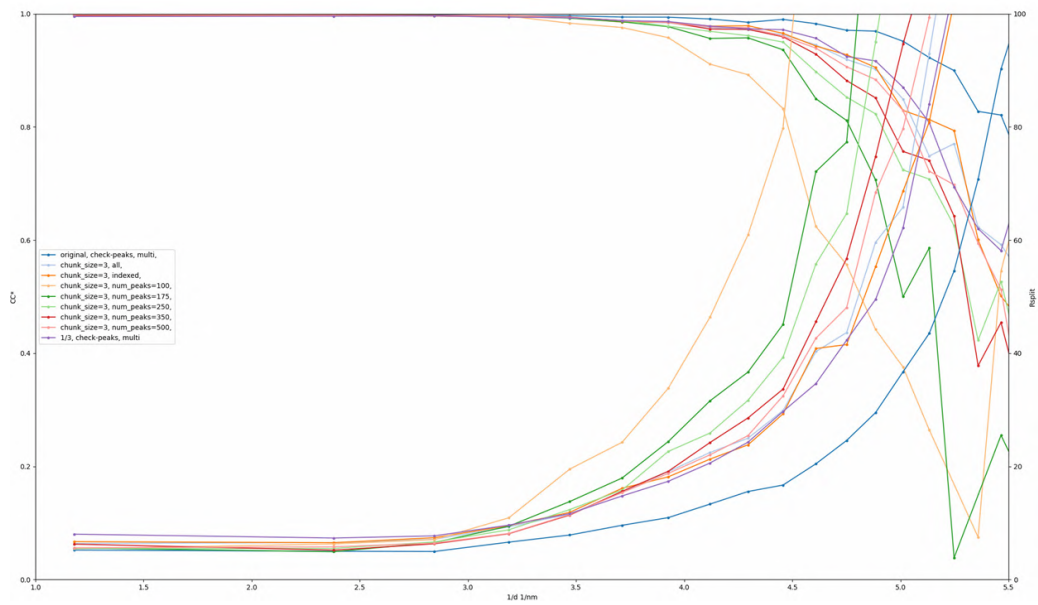


Figure E.6: Comparison of  $CC^*/R_{split}$  metrics of dimensionally reduced data with a chunk size of 3 with a randomly selected every third pattern from the total dataset.

Table E.4: Dimensional reduction

	Num. patterns/ hits	Ind. patterns/ crystals	Resol. Resol.	R <sub>split</sub> (%)	CC <sub>1/2</sub>	CC*	SNR	Comple. (%)	Multipl.	Total Measur.	Unique Reflect.	R <sub>free</sub> / R <sub>work</sub>	Wilson B-factor	CR
<b>1/3, no-check-peaks, no-multi</b>	12351/ 12351	8267/ 8267	39.60 - 1.75	16.54	0.974	0.993	6.481	90.035	59.580	683142	11466	0.1551/ 0.1955	16.98	3
<b>1/3, no-check-peaks, multi</b>	12351/ 12351	8346/ 13264	39.60 - 1.75	14.61	0.984	0.996	7.319	95.100	92.443	1119583	12111	0.1492/ 0.1935	16.33	3
<b>1/3, check-peaks, no-multi</b>	12351/ 12351	5904/ 5904	39.60 - 1.75	17.64	0.969	0.992	5.880	80.801	46.356	477003	10290	0.1645/ 0.2094	14.07	3
<b>1/3, check-peaks, multi</b>	12351/ 12351	7770/ 10944	39.60 - 1.75	14.62	0.983	0.996	6.801	94.707	79.143	954540	12061	0.1479/ 0.1899	18.11	3
<b>original, check-peaks, multi</b>	37053/ 37053	23257/ 32728	39.60 - 1.75	11.12	0.991	0.998	10.674	99.552	224.438	2845422	12678	0.1406/ 0.1700	15.77	1
<b>original, no-check-peaks, multi</b>	37053/ 37053	24877/ 39542	39.60 - 1.75	10.76	0.993	0.998	11.280	99.662	258.888	3285810	12692	0.1426/ 0.1726	15.49	1

Table E.4: Dimensional reduction

	Num. patterns/ hits	Ind. patterns/ crystals	Resol. Resol.	R <sub>split</sub> (%)	CC <sub>1/2</sub>	CC*	SNR	Comple. (%)	Multipl.	Total Measur.	Unique Reflect.	R <sub>free</sub> / R <sub>work</sub>	Wilson B-factor	CR
<b>original, no-check-peaks, no-multi</b>	37053/ 37053	24877/ 24877	39.60 - 1.75	12.45	0.987	0.997	9.346	98.508	162.815	2042513	12545	0.1490/ 0.1802	15.32	1
<b>original, check-peaks, no-multi</b>	37053/ 37053	17796/ 17796	39.60 - 1.75	13.88	0.985	0.996	8.046	94.896	118.466	1431657	12085	0.1501/ 0.1828	14.98	1
<b>chunk_size=3, num_peaks=250</b>	11419/ 11419	9634/ 20909	39.60 - 1.75	20.11	0.978	0.994	6.220	98.861	159.823	2012166	12590	0.1644/ 0.2164	22.2	2.84
<b>chunk_size=3, all</b>	12358/ 12358	9962/ 22161	39.60 - 1.75	19.68	0.979	0.995	7.032	99.741	187.557	2382344	12702	0.1594/ 0.1976	14.54	2.32
<b>chunk_size=3, indexed</b>	7830/ 7830	7518/ 19449	39.60 - 1.75	18.81	0.980	0.995	6.965	99.702	175.438	2227532	12697	0.1592/ 0.1935	14.62	4
<b>chunk_size=3, num_peaks=100</b>	9816/ 9816	8148/ 15780	39.60 - 1.75	23.6	0.966	0.991	4.756	93.663	105.247	1255391	11928	0.1774/ 0.2326	35.04	3.26

Table E.4: Dimensional reduction

	<b>Num.</b> patterns/ hits	<b>Ind.</b> patterns/ crystals	<b>Resol.</b>	$R_{\text{split}}$ (%)	$CC_{1/2}$	$CC^*$	<b>SNR</b>	<b>Comple.</b> (%)	<b>Multipl.</b>	<b>Total Measur.</b>	<b>Unique Reflect.</b>	$R_{\text{free}}/$ $R_{\text{work}}$	<b>Wilson B-factor</b>	<b>CR</b>
<b>chunk_size=3,</b> <b>num_peaks=175</b>	10974/ 10974	9280/ 19154	39.60 - 1.75	21.04	0.942	0.985	5.829	97.856	138.085	1720820	12462	0.1638/ 0.2193	22.78	2.93
<b>chunk_size=3,</b> <b>num_peaks=350</b>	11715/ 11715	9846/ 21676	39.60 - 1.75	18.61	0.982	0.995	6.634	99.262	176.436	2230332	12641	0.1623/ 0.2133	17.4	2.75
<b>chunk_size=3,</b> <b>num_peaks=500</b>	11965/ 11965	9941/ 21997	39.60 - 1.75	18.34	0.984	0.996	6.876	99.403	184.684	2337919	12659	0.1622/ 0.2059	19.4	2.67
<b>chunk_size=5,</b> <b>all</b>	7420/ 7420	5742/ 14872	39.60 - 1.75	22.32	0.979	0.995	5.392	99.945	154.701	1969032	12728	0.1676/ 0.2188	19.67	3.38
<b>chunk_size=5,</b> <b>indexed</b>	4701/ 4701	4319/ 12996	39.60 - 1.75	23.52	0.976	0.994	5.267	99.906	140.972	1793590	12723	0.1605/ 0.2113	20.46	6.29
<b>chunk_size=5,</b> <b>num_peaks=100</b>	5895/ 5895	4955/ 11581	39.60 - 1.75	30.76	0.971	0.993	3.771	99.380	89.608	1134073	12656	0.1881/ 0.2518	32.83	4.89

### E.3 Different samples used for the tests

The information about the samples (unit cell parameters and space group) used for different tests is presented in Table E.5.

Table E.5: Information about samples

<b>Sample</b>	<b>Unit cell parameters</b>	<b>Space group</b>
Lysozyme (Lysozyme (lyso))	79.2 79.2 38 90 90 90	P 43 21 2
Lactamase (Lactamase (lacta))	41.84 41.84 233.28 90 90 120	P 32 2 1
Ferritin	180.98 180.98 180.98 90 90 90	P 2 3
Granulovirus polyhedrin (Granulovirus polyhedrin (gv))	103.4 103.4 103.4 90 90 90	I 2 3
SARS-CoV-2 Main protease (MPro)	114.67 53.84 45.12 90 101.86 90	C 1 2 1
Thaumatococcus (Thaumatococcus (thau))	58.5 58.5 151.25 90 90 90	P 41 21 2





---

# Acronyms

<b>ADE</b>	a droplet-based injection
<b>lyso</b>	Lysozyme
<b>lacta</b>	Lactamase
<b>gv</b>	Granulovirus polyhedrin
<b>MPro</b>	SARS-CoV-2 Main protease
<b>thau</b>	Thaumatococcus
<b>GDVN</b>	dynamic virtual nozzles
<b>DFFN</b>	double-flow focusing nozzle
<b>MESH</b>	the microfluidic electrokinetic sample holder
<b>Mpix</b>	mega-pixel
<b>TDN</b>	the TapeDrive nozzle
<b>UC</b>	unit cell
<b>CR</b>	compression ratio
<b>HVE</b>	high-viscosity extrusion
<b>UV</b>	ultra-violet
<b>BITS</b>	a combined inject-and-transfer system
<b>EuXFEL</b>	European X-ray free electron laser
<b>XFEL</b>	X-ray free-electron laser
<b>FEL</b>	Free electron laser
<b>GUI</b>	graphical user interface
<b>SONICC</b>	second-order nonlinear imaging of chiral crystals
<b>PYP</b>	photo-active yellow protein
<b>PEG</b>	polyethylene glycol
<b>LCP</b>	lipidic cubic phase
<b>SX</b>	serial crystallography

**SSX** synchrotron serial crystallography  
**SFX** serial femtosecond crystallography  
**TR-SFX** time-resolved serial femtosecond crystallography  
**MX** macromolecular crystallography  
**MR** molecular replacement  
**MIR** multiple isomorphous replacement  
**SAD** single anomalous dispersion  
**MAD** multiple anomalous dispersion  
**HiPhaX** High-Throughput Pharmaceutical X-ray screening instrument  
**ESRF** European Synchrotron Radiation Facility  
**LCLS** Linac Coherent Light Source  
**SVD** singular value decomposition  
**CPSC** constrained pixel sum compression  
**HPC** high-performance computing  
**SwissFEL** Switzerland's X-ray free-electron laser at the Paul Scherrer Institute





---

# List of Publications

1. Christian Betzel, Andreas Prester, Markus Perbandt, Marina Galchenkova, Dominik Oberthuer, Nadine Werner, Alessandra Henkel, Julia Maracke, Oleksandr Yefanov, Johanna Hakanpää, et al. Time-resolved crystallography of boric acid binding to the active site serine of the  $\beta$ -lactamase ctx-m-14 and subsequent 1, 2-diol esterification. 2023.
2. Nina-Eleni Christou, Virginia Apostolopoulou, Diogo VM Melo, Matthias Ruppert, Alisia Fadini, Alessandra Henkel, Janina Sprenger, Dominik Oberthuer, Sebastian Günther, Anastasios Pateras, et al. Time-resolved crystallography captures light-driven dna repair. *Science*, 382(6674):1015–1020, 2023.
3. M Galchenkova, A Tolstikova, O Yefanov, H Chapman, et al. Data reduction in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 78:a266–a266, 2023.
4. Marina Galchenkova and Aleksei Korzhenkov. Modern tools for annotation of small genomes of non-model eukaryotes. *arXiv preprint arXiv:2102.04058*, 2021.
5. Marina Galchenkova, Aida Rahmani Mashhour, Patrick YA Reinke, Sebastian Günther, Jan Meyer, Henry N Chapman, and Oleksandr M Yefanov. An optimized approach for serial crystallography using chips. *Crystals*, 13(8):1225, 2023.
6. Marina Galchenkova, Alexandra Tolstikova, Bjarne Klopprogge, Janina Sprenger, Dominik Oberthuer, Wolfgang Brehm, Thomas A White, Anton Barty, Henry N Chapman, and Oleksandr Yefanov. Data reduction in protein serial crystallography. *IUCrJ*, 11(2), 2024.
7. Yaroslav Gevorkov, Marina Galchenkova, Valerio Mariani, Anton Barty, Thomas A White, Henry N Chapman, and Oleksandr Yefanov. Fdip—a fast diffraction image processing library for x-ray crystallography experiments. *Crystals*, 14(2):164, 2024.
8. Sebastian Guenther, Patrick YA Reinke, Yaiza Fernandez-Garcia, Julia Lieske, Thomas J Lane, Helen Ginn, Faisal Koua, Christiane Ehrt, Wiebke Ewert, Dominik Oberthuer, et al. Massive x-ray screening reveals two allosteric drug binding sites of sars-cov-2 main protease (preprint). *Biorxiv*, 2020.
9. Sebastian Günther, Patrick YA Reinke, Yaiza Fernández-García, Julia Lieske, Thomas J Lane, Helen Ginn, Faisal HM Koua, Christiane Ehrt, Wiebke Ewert, Dominik Oberthuer, et al. Massive x-ray screening reveals two allosteric drug binding sites of sars-cov-2 main protease. *bioRxiv*, pages 2020–11, 2020.
10. Sebastian Günther, Patrick YA Reinke, Yaiza Fernández-García, Julia Lieske, Thomas J Lane, Helen M Ginn, Faisal HM Koua, Christiane Ehrt, Wiebke Ewert, Dominik Oberthuer, et al. Inhibition of sars-cov-2 main protease by allosteric drug-binding. Technical report, 2020.

11. Sebastian Günther, Patrick YA Reinke, Yaiza Fernández-García, Julia Lieske, Thomas J Lane, Helen M Ginn, Faisal HM Koua, Christiane Ehrt, Wiebke Ewert, Dominik Oberthuer, et al. X-ray screening identifies active site and allosteric inhibitors of sars-cov-2 main protease. *Science*, 372(6542):642–646, 2021.
12. Sebastian Günther, Patrick YA Reinke, Dominik Oberthuer, Oleksandr Yefanov, Helen Ginn, Susanne Meier, Thomas J Lane, Kristina Lorenzen, Luca Gelisio, Wolfgang Brehm, et al. Catalytic cleavage of heat and subsequent covalent binding of the tetralone moiety by the sars-cov-2 main protease. *bioRxiv*, pages 2020–05, 2020.
13. Marjan Hadian-Jazi, Alireza Sadri, Anton Barty, Oleksandr Yefanov, Marina Galchenkova, Dominik Oberthuer, Dana Komadina, Wolfgang Brehm, Henry Kirkwood, Grant Mills, et al. Data reduction for serial crystallography using a robust peak finder. *Journal of Applied Crystallography*, 54(5):1360–1378, 2021.
14. A Henkel, J Maracke, A Munke, M Galchenkova, A Rahmani Mashhour, P Reinke, M Domaracky, H Fleckenstein, J Hakanpää, J Meyer, et al. Cfel tapedrive 2.0: a conveyor belt-based sample-delivery system for multi-dimensional serial crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 78:a560–a560, 2023.
15. Alessandra Henkel, Marina Galchenkova, Julia Maracke, Oleksandr Yefanov, Johanna Hakanpää, Jeroen R Mesters, Henry N Chapman, and Dominik Oberthür. Jinxed: Just in time crystallization for easy structure determination of biological macromolecules. *bioRxiv*, pages 2022–10, 2022.
16. Alessandra Henkel, Marina Galchenkova, Julia Maracke, Oleksandr Yefanov, Bjarne Klopprogge, Johanna Hakanpää, Jeroen R Mesters, Henry N Chapman, and Dominik Oberthuer. Jinxed: just in time crystallization for easy structure determination of biological macromolecules. *IUCrJ*, 10(3), 2023.
17. Alireza Sadri, Marjan Hadian-Jazi, Oleksandr Yefanov, Marina Galchenkova, Henry Kirkwood, Grant Mills, Marcin Sikorski, Romain Letrun, Raphael de Wijn, Mohammad Vakili, et al. Automatic bad-pixel mask maker for x-ray pixel detectors with application to serial crystallography. *Journal of Applied Crystallography*, 55(6), 2022.
18. Vasundara Srinivasan, Hévila Brognaro, Prince R Prabhu, Edmarcia Elisa de Souza, Sebastian Günther, Patrick YA Reinke, Thomas J Lane, Helen Ginn, Huijong Han, Wiebke Ewert, et al. Sars-cov-2 papain-like protease plpro in complex with natural compounds reveal allosteric sites for antiviral drug design. *Biorxiv*, pages 2021–11, 2021.
19. Vasundara Srinivasan, Hévila Brognaro, Prince R Prabhu, Edmarcia Elisa de Souza, Sebastian Günther, Patrick YA Reinke, Thomas J Lane, Helen Ginn, Huijong Han, Wiebke Ewert, et al. Antiviral activity of natural phenolic compounds in complex at an allosteric site of sars-cov-2 papain-like protease. *Communications Biology*, 5(1):805, 2022.
20. Lainey J Williamson, Marina Galchenkova, Hannah L Best, Richard J Bean, Anna Munke, Salah Awel, Gisel Pena, Juraj Knoska, Robin Schubert, Katerina Doerner, et al. Structure of the lysinibacillus sphaericus tpp49aa1 pesticidal protein elucidated from natural crystals using mhz-sfx. *bioRxiv*, pages 2022–01, 2022.
21. Lainey J Williamson, Marina Galchenkova, Hannah L Best, Richard J Bean, Anna Munke, Salah Awel, Gisel Pena, Juraj Knoska, Robin Schubert, Katerina Doerner, et al. Structure of the lysinibacillus sphaericus tpp49aa1 pesticidal protein elucidated from natural crystals using mhz-sfx. *Proceedings of the National Academy of Sciences*, 120(49):e2203241120, 2023.

---

# Bibliography

- [1] Tom L Blundell. Protein crystallography and drug discovery: recollections of knowledge exchange between academia and industry. *IUCrJ*, 4(4):308–321, 2017.
- [2] Anna Pomés, Maksymilian Chruszcz, Alla Gustchina, Wladek Minor, Geoffrey A Mueller, Lars C Pedersen, Alexander Wlodawer, and Martin D Chapman. 100 years later: Celebrating the contributions of x-ray crystallography to allergy and clinical immunology. *Journal of Allergy and Clinical Immunology*, 136(1):29–37, 2015.
- [3] Vasundara Srinivasan, Hévila Brognaro, Prince R Prabhu, Edmarcia Elisa de Souza, Sebastian Günther, Patrick YA Reinke, Thomas J Lane, Helen Ginn, Huijong Han, Wiebke Ewert, et al. Antiviral activity of natural phenolic compounds in complex at an allosteric site of sars-cov-2 papain-like protease. *Communications Biology*, 5(1):805, 2022.
- [4] Eric Ennifar. X-ray crystallography as a tool for mechanism-of-action studies and drug discovery. *Current pharmaceutical biotechnology*, 14(5):537–550, 2013.
- [5] Shuke Wu, Radka Snajdrova, Jeffrey C Moore, Kai Baldenius, and Uwe T Bornscheuer. Biocatalysis: enzymatic synthesis for industrial applications. *Angewandte Chemie International Edition*, 60(1):88–119, 2021.
- [6] Elspeth F Garman. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D: Biological Crystallography*, 66(4):339–351, 2010.
- [7] Karol Nass. Radiation damage in protein crystallography at x-ray free-electron lasers. *Acta Crystallographica Section D: Structural Biology*, 75(2):211–218, 2019.
- [8] Gisela Brändén and Richard Neutze. Advances and challenges in time-resolved macromolecular crystallography. *Science*, 373(6558):eaba0954, 2021.
- [9] Henry N Chapman, Petra Fromme, Anton Barty, Thomas A White, Richard A Kirian, Andrew Aquila, Mark S Hunter, Joachim Schulz, Daniel P DePonte, Uwe Weierstall, et al. Femtosecond x-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.
- [10] Francesco Stellato, Dominik Oberthür, Mengning Liang, Richard Bean, Cornelius Gati, Oleksandr Yefanov, Anton Barty, Anja Burkhardt, Pontus Fischer, Lorenzo Galli, et al. Room-temperature macromolecular serial crystallography using synchrotron radiation. *IUCrJ*, 1(4):204–212, 2014.

- [11] Cornelius Gati, Gleb Bourenkov, Marco Klinge, Dirk Rehders, Francesco Stellato, Dominik Oberthür, Oleksandr Yefanov, Benjamin P Sommer, Stefan Mogk, Michael Duszhenko, et al. Serial crystallography on in vivo grown microcrystals using synchrotron radiation. *IUCrJ*, 1(2):87–94, 2014.
- [12] Sébastien Boutet, Lukas Lomb, Garth J Williams, Thomas RM Barends, Andrew Aquila, R Bruce Doak, Uwe Weierstall, Daniel P DePonte, Jan Steinbrener, Robert L Shoeman, et al. High-resolution protein structure determination by serial femtosecond crystallography. *Science*, 337(6092):362–364, 2012.
- [13] Richard Neutze, Remco Wouts, David Van der Spoel, Edgar Weckert, and Janos Hajdu. Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406(6797):752–757, 2000.
- [14] Jörg Standfuss and John Spence. Serial crystallography at synchrotrons and x-ray lasers. *IUCrJ*, 4(Pt 2):100, 2017.
- [15] Jose M Martin-Garcia, Chelsie E Conrad, Jesse Coe, Shatabdi Roy-Chowdhury, and Petra Fromme. Serial femtosecond crystallography: A revolution in structural biology. *Archives of biochemistry and biophysics*, 602:32–47, 2016.
- [16] Marius Schmidt. Time-resolved macromolecular crystallography at pulsed x-ray sources. *International Journal of Molecular Sciences*, 20(6):1401, 2019.
- [17] Marius Schmidt. Reaction initiation in enzyme crystals by diffusion of substrate. *Crystals*, 10(2):116, 2020.
- [18] A Tolstikova, M Levantino, O Yefanov, V Hennicke, P Fischer, J Meyer, A Mozzanica, S Redford, E Crosas, NL Opara, et al. 1 khz fixed-target serial crystallography using a multilayer monochromator and an integrating pixel detector. *IUCrJ*, 6(5):927–937, 2019.
- [19] Marie Luise Grünbein and Gabriela Nass Kovacs. Sample delivery for serial crystallography at free-electron lasers and synchrotrons. *Acta Crystallographica Section D: Structural Biology*, 75(2):178–191, 2019.
- [20] Feng-Zhu Zhao, Bin Zhang, Er-Kai Yan, Bo Sun, Zhi-Jun Wang, Jian-Hua He, and Da-Chuan Yin. A guide to sample delivery systems for serial crystallography. *The FEBS Journal*, 286(22):4402–4417, 2019.
- [21] Julia Lieske, Maximilian Cerv, Stefan Kreida, Dana Komadina, Janine Fischer, Miriam Barthelmess, Pontus Fischer, Tim Pakendorf, Oleksandr Yefanov, Valerio Mariani, et al. On-chip crystallization for serial crystallography experiments and on-chip ligand-binding studies. *IUCrJ*, 6(4):714–728, 2019.
- [22] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Jordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.
- [23] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex Bridgland, et al. Alphafold 2. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction; DeepMind: London, UK*, 2020.



- [24] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- [25] Thomas A White, Valerio Mariani, Wolfgang Brehm, Oleksandr Yefanov, Anton Barty, Kenneth R Beyerlein, Fedor Chervinskii, Lorenzo Galli, Cornelius Gati, Takanori Nakane, et al. Recent developments in crystfel. *Journal of applied crystallography*, 49(2):680–689, 2016.
- [26] Oleksandr Yefanov, Valerio Mariani, Cornelius Gati, Thomas A White, Henry N Chapman, and Anton Barty. Accurate determination of segmented x-ray detector geometry. *Optics express*, 23(22):28459–28470, 2015.
- [27] William Henry Bragg, Max von Laue, and Carl Hermann. *Internationale tabellen zur bestimmung von kristallstrukturen: bd. Gruppentheoretische tafeln*, volume 1. JW Edwards, 1935.
- [28] Kathleen Lonsdale et al. *International tables for X-ray crystallography*. Kynock Press, 1952.
- [29] Dame Kathleen Lonsdale. Simplified structure factor and electron density formulae for the 230 space groups of mathematical crystallography. Royal institution, 1936.
- [30] Th Hahn. International tables for crystallography. vol. a. space-group symmetry.-5th rev. 2005.
- [31] Kathleen Lonsdale. The structure of the benzene ring. *Nature*, 122(3082):810–810, 1928.
- [32] JD Bernal. Trans. faraday. soc. 1933.
- [33] J Monteath Robertson. 136. an x-ray study of the structure of the phthalocyanines. part i. the metal-free, nickel, copper, and platinum compounds. *Journal of the Chemical Society (Resumed)*, pages 615–621, 1935.
- [34] JM Bijvoet. Phase determination in direct fourier-synthesis of crystal structures. In *Proc. K. Ned. Akad. Wet*, volume 52, pages 313–314, 1949.
- [35] JM Bijvoet, AF Peerdeman, and AJ Van Bommel. Determination of the absolute configuration of optically active compounds by means of x-rays. *Nature*, 168:271–272, 1951.
- [36] D Crowfoot. X-ray crystallographic studies of compounds of biochemical interest. *Annual review of biochemistry*, 17(1):115–146, 1948.
- [37] D Crowfoot, CW Bunn, BW Rogers-Low, and A Turner-Jones. The chemistry of penicillin. *HT Clarke, JR Johnson & R. Robinson (eds.)*, pages 310–367, 1949.
- [38] Dorothy Mary Crowfoot Hodgkin, Jennifer Kamper, June Lindsey, Maureen MacKay, Jenny Pickworth, JH Robertson, Clara Brink-Shoemaker, John Graham White, RJ Prosen, and KN Trueblood. The structure of vitamin b12. i. an outline of the crystallographic investigation of vitamin b12. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 242(1229):228–263, 1957.
- [39] Jens Als-Nielsen and Des McMorrow. *Elements of modern X-ray physics*. John Wiley & Sons, 2011.
- [40] Philip Willmott. *An introduction to synchrotron radiation: techniques and applications*. John Wiley & Sons, 2019.

- [41] Sebastian Günther, Patrick YA Reinke, Yaiza Fernández-García, Julia Lieske, Thomas J Lane, Helen M Ginn, Faisal HM Koua, Christiane Ehrh, Wiebke Ewert, Dominik Oberthuer, et al. X-ray screening identifies active site and allosteric inhibitors of sars-cov-2 main protease. *Science*, 372(6542):642–646, 2021.
- [42] Ya S Derbenev, AM Kondratenko, and EL Saldin. On the possibility of using a free electron laser for polarization of electrons in storage rings. *Nuclear Instruments and Methods in Physics Research*, 193(3):415–421, 1982.
- [43] Alke Meents, MO Wiedorn, V Srajer, R Henning, I Sarrou, J Bergtholdt, M Barthelmeß, PYA Reinke, D Dierksmeyer, A Tolstikova, et al. Pink-beam serial crystallography. *Nature communications*, 8(1):1–12, 2017.
- [44] Jose M Martin-Garcia, Lan Zhu, Derek Mendez, M-Y Lee, Eugene Chun, Chufeng Li, Hao Hu, Ganesh Subramanian, David Kissick, Craig Ogata, et al. High-viscosity injector-based pink-beam serial crystallography of microcrystals at a synchrotron radiation source. *IUCrJ*, 6(3):412–425, 2019.
- [45] Yaroslav Gevorkov, Anton Barty, Wolfgang Brehm, Thomas A White, Aleksandra Tolstikova, Max O Wiedorn, Alke Meents, R-R Grigat, Henry N Chapman, and Oleksandr Yefanov. pinkindexer—a universal indexer for pink-beam x-ray and electron diffraction snapshots. *Acta Crystallographica Section A: Foundations and Advances*, 76(2):121–131, 2020.
- [46] Karol Nass, Camila Bacellar, Claudio Cirelli, Florian Dworkowski, Yaroslav Gevorkov, Daniel James, Philip JM Johnson, Demet Kekilli, Gregor Knopp, Isabelle Martiel, et al. Pink-beam serial femtosecond crystallography for accurate structure-factor determination at an x-ray free-electron laser. *IUCrJ*, 8(6), 2021.
- [47] Aleksandra Tolstikova. *Development of diffraction analysis methods for serial crystallography*. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2020.
- [48] Michele Cianci, Gleb Bourenkov, Guillaume Pompidor, Ivars Karpics, Johanna Kallio, Isabel Bento, Manfred Roessle, Florent Cipriani, Stefan Fiedler, and Thomas R Schneider. P13, the embl macromolecular crystallography beamline at the low-emittance petra iii ring for high-and low-energy phasing with variable beam focusing. *Journal of synchrotron radiation*, 24(1):323–332, 2017.
- [49] I Margiolaki and JP Wright. Powder crystallography on macromolecules. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):169–180, 2008.
- [50] Isabelle Martiel, Henrike M Müller-Werkmeister, and Aina E Cohen. Strategies for sample delivery for femtosecond crystallography. *Acta Crystallographica Section D: Structural Biology*, 75(2):160–177, 2019.
- [51] Allison Doerr. Diffraction before destruction. *Nature Methods*, 8(4):283–283, 2011.
- [52] Henry N Chapman, Carl Caleman, and Nicusor Timneanu. Diffraction before destruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647):20130313, 2014.

- [53] DP DePonte, Uwe Weierstall, Kevin Schmidt, J Warner, D Starodub, JCH Spence, and RB Doak. Gas dynamic virtual nozzle for generation of microscopic droplet streams. *Journal of Physics D: Applied Physics*, 41(19):195505, 2008.
- [54] Raymond G Sierra, Hartawan Laksmono, Jan Kern, Rosalie Tran, Johan Hattne, Roberto Alonso-Mori, Benedikt Lassalle-Kaiser, Carina Glöckner, Julia Hellmich, Donald W Schafer, et al. Nanoflow electrospinning serial femtosecond crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 68(11):1584–1587, 2012.
- [55] Uwe Weierstall, JCH Spence, and RB Doak. Injector for scattering measurements on fully solvated biospecies. *Review of Scientific Instruments*, 83(3):035108, 2012.
- [56] Michihiro Sugahara, Eiichi Mizohata, Eriko Nango, Mamoru Suzuki, Tomoyuki Tanaka, Tetsuya Masuda, Rie Tanaka, Tatsuro Shimamura, Yoshiki Tanaka, Chiyo Suno, et al. Grease matrix as a versatile carrier of proteins for serial crystallography. *Nature methods*, 12(1):61–63, 2015.
- [57] Sabine Botha, Karol Nass, Thomas RM Barends, Wolfgang Kabsch, Beatrice Latz, Florian Dworkowski, Lutz Foucar, Ezequiel Panepucci, Meitian Wang, Robert L Shoeman, et al. Room-temperature serial crystallography at synchrotron x-ray sources using slowly flowing free-standing high-viscosity microstreams. *Acta Crystallographica Section D: Biological Crystallography*, 71(2):387–397, 2015.
- [58] Uwe Weierstall, Daniel James, Chong Wang, Thomas A White, Dingjie Wang, Wei Liu, John CH Spence, R Bruce Doak, Garrett Nelson, Petra Fromme, et al. Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nature communications*, 5(1):1–6, 2014.
- [59] Franklin D Fuller, Sheraz Gul, Ruchira Chatterjee, E Sethe Burgie, Iris D Young, Hugo Lebrette, Vivek Srinivas, Aaron S Brewster, Tara Michels-Clark, Jonathan A Clinger, et al. Drop-on-demand sample delivery for studying biocatalysts in action at x-ray free-electron lasers. *Nature methods*, 14(4):443–449, 2017.
- [60] D Bourgeois and M Weik. Kinetic protein crystallography: a tool to watch proteins in action. *Crystallography reviews*, 15(2):87–118, 2009.
- [61] Dominique Bourgeois, Friedrich Schotte, Maurizio Brunori, and Beatrice Vallone. Time-resolved methods in biophysics. 6. time-resolved laue crystallography as a tool to investigate photo-activated protein dynamics. *Photochemical & Photobiological Sciences*, 6(10):1047–1056, 2007.
- [62] Vukica Šrajer and Marius Schmidt. Watching proteins function with time-resolved x-ray crystallography. *Journal of physics D: Applied physics*, 50(37):373001, 2017.
- [63] Ilme Schlichting. Serial femtosecond crystallography: the first five years. *IUCrJ*, 2(2):246–255, 2015.
- [64] Thomas RM Barends, Lutz Foucar, Albert Ardevol, Karol Nass, Andrew Aquila, Sabine Botha, R Bruce Doak, Konstantin Falahati, Elisabeth Hartmann, Mario Hilpert, et al. Direct observation of ultrafast collective motions in co myoglobin upon ligand dissociation. *Science*, 350(6259):445–450, 2015.

- [65] Kanupriya Pande, Christopher DM Hutchison, Gerrit Groenhof, Andy Aquila, Josef S Robinson, Jason Tenboer, Shibom Basu, Sébastien Boutet, Daniel P DePonte, Mengning Liang, et al. Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science*, 352(6286):725–729, 2016.
- [66] Mark A Wilson. Mapping enzyme landscapes by time-resolved crystallography with synchrotron and x-ray free electron laser light. *Annual Review of Biophysics*, 51:79–98, 2022.
- [67] Henry N Chapman. X-ray free-electron lasers for the structure and dynamics of macromolecules. *Annual review of biochemistry*, 88:35–58, 2019.
- [68] Allen M Orville. Recent results in time resolved serial femtosecond crystallography at xfels. *Current Opinion in Structural Biology*, 65:193–208, 2020.
- [69] Christopher Kupitz, Jose L Olmos Jr, Mark Holl, Lee Tremblay, Kanupriya Pande, Suraj Pandey, Dominik Oberthür, Mark Hunter, Mengning Liang, Andrew Aquila, et al. Structural enzymology using x-ray free electron lasers. *Structural Dynamics*, 4(4):044003, 2017.
- [70] Suraj Pandey, Richard Bean, Tokushi Sato, Ishwor Poudyal, Johan Bielecki, Jorvani Cruz Villarreal, Oleksandr Yefanov, Valerio Mariani, Thomas A White, Christopher Kupitz, et al. Time-resolved serial femtosecond crystallography at the european xfel. *Nature methods*, 17(1):73–78, 2020.
- [71] Andrew Aquila, Mark S Hunter, R Bruce Doak, Richard A Kirian, Petra Fromme, Thomas A White, Jakob Andreasson, David Arnlund, Saša Bajt, Thomas RM Barends, et al. Time-resolved protein nanocrystallography using an x-ray free-electron laser. *Optics express*, 20(3):2706–2716, 2012.
- [72] Jacques-Philippe Colletier, Giorgio Schirò, and Martin Weik. Time-resolved serial femtosecond crystallography, towards molecular movies of biomolecules in action. *X-Ray Free Electron Lasers: A Revolution in Structural Biology*, pages 331–356, 2018.
- [73] Andrew C Kruse, Aashish Manglik, Brian K Kobilka, and William I Weis. Applications of molecular replacement to g protein-coupled receptors. *Acta Crystallographica Section D: Biological Crystallography*, 69(11):2287–2292, 2013.
- [74] Marius Schmidt. Mix and inject: Reaction initiation by diffusion for time-resolved macromolecular crystallography. *Advances in Condensed Matter Physics*, 2013, 2013.
- [75] JR Stagno, Y Liu, YR Bhandari, CE Conrad, S Panja, Mamata Swain, L Fan, Gerald Nelson, C Li, DR Wendel, et al. Structures of riboswitch rna reaction states by mix-and-inject xfel serial crystallography. *Nature*, 541(7636):242–246, 2017.
- [76] Harshwardhan Poddar, Derren J Heyes, Giorgio Schirò, Martin Weik, David Leys, and Nigel S Scrutton. A guide to time-resolved structural analysis of light-activated proteins. *The FEBS Journal*, 289(3):576–595, 2022.
- [77] Jan Kubelka. Time-resolved methods in biophysics. 9. laser temperature-jump methods for investigating biomolecular dynamics. *Photochemical & Photobiological Sciences*, 8(4):499–512, 2009.

- [78] Ilme Schlichting, Joel Berendzen, Kelvin Chu, Ann M Stock, Shelley A Maves, David E Benson, Robert M Sweet, Dagmar Ringe, Gregory A Petsko, and Stephen G Sligar. The catalytic pathway of cytochrome p450cam at atomic resolution. *Science*, 287(5458):1615–1622, 2000.
- [79] Jacques-Philippe Colletier, Dominique Bourgeois, Benoît Sanson, Didier Fournier, Joel L Sussman, Israel Silman, and Martin Weik. Shoot-and-trap: Use of specific x-ray damage to study structural protein dynamics by temperature-controlled cryo-crystallography. *Proceedings of the National Academy of Sciences*, 105(33):11742–11747, 2008.
- [80] Doeke R Hekstra, K Ian White, Michael A Socolich, Robert W Henning, Vukica Šrajer, and Rama Ranganathan. Electric-field-stimulated protein mechanics. *Nature*, 540(7633):400–405, 2016.
- [81] Wolfgang Kabsch. Processing of x-ray snapshots from crystals in random orientations. *Acta Crystallographica Section D: Biological Crystallography*, 70(8):2204–2216, 2014.
- [82] Thomas A White. Post-refinement method for snapshot serial crystallography. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647):20130330, 2014.
- [83] Nicholas K Sauter. Xfel diffraction: developing processing methods to optimize data quality. *Journal of synchrotron radiation*, 22(2):239–248, 2015.
- [84] Monarin Uervirojnangkoorn, Oliver B Zeldin, Artem Y Lyubimov, Johan Hattne, Aaron S Brewster, Nicholas K Sauter, Axel T Brunger, and William I Weis. Enabling x-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *Elife*, 4:e05421, 2015.
- [85] Helen Mary Ginn, Aaron S Brewster, Johan Hattne, Gwyndaf Evans, Armin Wagner, Jonathan M Grimes, Nicholas K Sauter, Geoff Sutton, and David Ian Stuart. A revised partiality model and post-refinement algorithm for x-ray free-electron laser data. *Acta Crystallographica Section D: Biological Crystallography*, 71(6):1400–1410, 2015.
- [86] Loes MJ Kroon-Batenburg, Antoine MM Schreurs, Raimond BG Ravelli, and Piet Gros. Accounting for partiality in serial crystallography using ray-tracing principles. *Acta Crystallographica Section D: Biological Crystallography*, 71(9):1799–1811, 2015.
- [87] Thomas A White, Anton Barty, Francesco Stellato, James M Holton, Richard A Kirian, Nadia A Zatsepin, and Henry N Chapman. Crystallographic data processing for free-electron laser sources. *Acta Crystallographica Section D: Biological Crystallography*, 69(7):1231–1240, 2013.
- [88] Dima Chirgadze. Protein crystallisation in action. *University of Cambridge*, 2001.
- [89] R Bosch, P Lautenschlager, L Potthast, and J Stapelmann. Experiment equipment for protein crystallization in  $\mu\text{g}$  facilities. *Journal of crystal growth*, 122(1-4):310–316, 1992.
- [90] David Jeruzalmi and Thomas A Steitz. Use of organic cosmotropic solutes to crystallize flexible proteins: application to t7 rna polymerase and its complex with the inhibitor t7 lysozyme. *Journal of molecular biology*, 274(5):748–756, 1997.
- [91] Stéphane B Richard, Dominique Madern, Elsa Garcin, and Giuseppe Zaccai. Halophilic adaptation: novel solvent protein interactions observed in the 2.9 and 2.6 Å resolution structures of the wild type and a mutant of malate dehydrogenase from haloarcula marismortui. *Biochemistry*, 39(5):992–1000, 2000.

- [92] Robert S Snyder, Klaus Fuhrmann, and Hannes U Walter. Protein crystallization facilities for microgravity experiments. *Journal of crystal growth*, 110(1-2):333–338, 1991.
- [93] Pamela J Bjorkman, MA Saper, B\_ Samraoui, William S Bennett, JL t Strominger, and DC Wiley. Structure of the human class i histocompatibility antigen, hla-a2. *Nature*, 329(6139):506–512, 1987.
- [94] Ali A Kermani. A guide to membrane protein x-ray crystallography. *The FEBS journal*, 288(20):5788–5804, 2021.
- [95] Joscha Breibeck and Annette Rompel. Successful amphiphiles as the key to crystallization of membrane proteins: Bridging theory and practice. *Biochimica Et Biophysica Acta (BBA)-General Subjects*, 1863(2):437–455, 2019.
- [96] Dominik Oberthuer, Juraj Knoška, Max O Wiedorn, Kenneth R Beyerlein, David A Bushnell, Elena G Kovaleva, Michael Heymann, Lars Gumprecht, Richard A Kirian, Anton Barty, et al. Double-flow focused liquid injector for efficient serial femtosecond crystallography. *Scientific reports*, 7(1):1–12, 2017.
- [97] Austin Echelmeier, Daihyun Kim, Jorvani Cruz Villarreal, Jesse Coe, Sebastian Quintana, Gerrit Brehm, Ana Egatz-Gomez, Reza Nazari, Raymond G Sierra, Jason E Koglin, et al. 3d printed droplet generation devices for serial femtosecond crystallography enabled by surface coating. *Journal of applied crystallography*, 52(5):997–1008, 2019.
- [98] A Echelmeier, G Nelson, BG Abdallah, D James, S Roy-Chowdhury, A Tolstikova, V Mariani, Richard Kirian, D Oberthür, K Dörner, et al. Biphasic droplet-based sample delivery of protein crystals for serial femtosecond crystallography with an x-ray free electron laser. In *19th International Conference on Miniaturized Systems for Chemistry and Life Sciences, MicroTAS 2015*, pages 1374–1376. Chemical and Biological Microsystems Society, 2015.
- [99] Daihyun Kim, Austin Echelmeier, Jorvani Cruz Villarreal, Sahir Gandhi, Sebastian Quintana, Ana Egatz-Gomez, and Alexandra Ros. Electric triggering for enhanced control of droplet generation. *Analytical chemistry*, 91(15):9792–9799, 2019.
- [100] Fumitaka Mafuné, Ken Miyajima, Kensuke Tono, Yoshihiro Takeda, Jun-ya Kohno, Naoya Miyauchi, Jun Kobayashi, Yasumasa Joti, Eriko Nango, So Iwata, et al. Microcrystal delivery by pulsed liquid droplet for serial femtosecond crystallography. *Acta Crystallographica Section D: Structural Biology*, 72(4):520–523, 2016.
- [101] Christian G Roessler, Rakhi Agarwal, Marc Allaire, Roberto Alonso-Mori, Babak Andi, José FR Bachega, Martin Bommer, Aaron S Brewster, Michael C Browne, Ruchira Chatterjee, et al. Acoustic injectors for drop-on-demand serial femtosecond crystallography. *Structure*, 24(4):631–640, 2016.
- [102] Christian G Roessler, Anthony Kuczewski, Richard Stearns, Richard Ellson, Joseph Olechno, Allen M Orville, Marc Allaire, Alexei S Soares, and Annie Héroux. Acoustic methods for high-throughput protein crystal mounting at next-generation macromolecular crystallographic beamlines. *Journal of synchrotron radiation*, 20(5):805–808, 2013.
- [103] Soichiro Tsujino, Akira Shinoda, and Takashi Tomizaki. On-demand droplet loading of ultrasonic acoustic levitator and its application for protein crystallography experiments. *Applied Physics Letters*, 114(21):213702, 2019.

- [104] Christopher DM Hutchison, Violeta Cordon-Preciado, Rhodri ML Morgan, Takanori Nakane, Josie Ferreira, Gabriel Dorlhiac, Alvaro Sanchez-Gonzalez, Allan S Johnson, Ann Fitzpatrick, Clyde Fare, et al. X-ray free electron laser determination of crystal structures of dark and light states of a reversibly photoswitching fluorescent protein at room temperature. *International Journal of Molecular Sciences*, 18(9):1918, 2017.
- [105] Kenneth R Beyerlein, Dennis Dierksmeyer, Valerio Mariani, Manuela Kuhn, Iosifina Sarrou, Angelica Ottaviano, Salah Awel, Juraj Knoska, Silje Fuglerud, Olof Jönsson, et al. Mix-and-diffuse serial synchrotron crystallography. *IUCrJ*, 4(6):769–777, 2017.
- [106] Kara A Zielinski, Andreas Prester, Hina Andaleeb, Soi Bui, Oleksandr Yefanov, Lucrezia Catapano, Alessandra Henkel, Max O Wiedorn, Olga Lorbeer, Eva Crosas, et al. Rapid and efficient room-temperature serial synchrotron crystallography using the cfel tapedrive. *IUCrJ*, 9(6):778–791, 2022.
- [107] Alessandra Henkel, Marina Galchenkova, Julia Maracke, Oleksandr Yefanov, Johanna Hakanpää, Jeroen R Mesters, Henry N Chapman, and Dominik Oberthür. Jinxed: Just in time crystallization for easy structure determination of biological macromolecules. *bioRxiv*, pages 2022–10, 2022.
- [108] A Henkel, J Maracke, A Munke, M Galchenkova, A Rahmani Mashhour, P Reinke, M Domaracky, H Fleckenstein, J Hakanpää, J Meyer, et al. Cfel tapedrive 2.0: a conveyor belt-based sample-delivery system for multi-dimensional serial crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 78:a560–a560, 2023.
- [109] Jan Kern, Roberto Alonso-Mori, Julia Hellmich, Rosalie Tran, Johan Hattne, Hartawan Laksmono, Carina Glöckner, Nathaniel Echols, Raymond G Sierra, Jonas Sellberg, et al. Room temperature femtosecond x-ray diffraction of photosystem ii microcrystals. *Proceedings of the National Academy of Sciences*, 109(25):9721–9726, 2012.
- [110] Raymond G Sierra, Cornelius Gati, Hartawan Laksmono, E Han Dao, Sheraz Gul, Franklin Fuller, Jan Kern, Ruchira Chatterjee, Mohamed Ibrahim, Aaron S Brewster, et al. Concentric-flow electrokinetic injector enables serial crystallography of ribosome and photosystem ii. *Nature methods*, 13(1):59–62, 2016.
- [111] C Mueller, Alexander Marx, Sascha W Epp, Y Zhong, A Kuo, AR Balo, J Soman, F Schotte, HT Lemke, RL Owen, et al. Fixed target matrix for femtosecond time-resolved and in situ serial micro-crystallography. *Structural Dynamics*, 2(5):054302, 2015.
- [112] P Roedig, I Vartiainen, R Duman, S Panneerselvam, N Stübe, O Lorbeer, M Warmer, G Sutton, DI Stuart, E Weckert, et al. A micro-patterned silicon chip as sample holder for macromolecular crystallography experiments with minimal background scattering. *Scientific reports*, 5(1):1–11, 2015.
- [113] Saeed Oghbaey, Antoine Sarracini, Helen M Ginn, Olivier Pare-Labrosse, Anling Kuo, Alexander Marx, Sascha W Epp, Darren A Sherrell, Bryan T Eger, Yinpeng Zhong, et al. Fixed target combined with spectral mapping: approaching 100% hit rates for serial crystallography. *Acta Crystallographica Section D: Structural Biology*, 72(8):944–955, 2016.

- [114] Robin L Owen, Danny Axford, Darren A Sherrell, Anling Kuo, Oliver P Ernst, Eike C Schulz, RJ Dwayne Miller, and Henrike M Mueller-Werkmeister. Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallographica Section D: Structural Biology*, 73(4):373–378, 2017.
- [115] Gera Kisselman, Wei Qiu, Vladimir Romanov, Christine M Thompson, Robert Lam, Kevin P Battaile, Emil F Pai, and Nickolay Y Chirgadze. X-chip: an integrated platform for high-throughput protein crystallization and on-the-chip x-ray diffraction data collection. *Acta Crystallographica Section D: Biological Crystallography*, 67(6):533–539, 2011.
- [116] Philip Roedig, Helen M Ginn, Tim Pakendorf, Geoff Sutton, Karl Harlos, Thomas S Walter, Jan Meyer, Pontus Fischer, Ramona Duman, Ismo Vartiainen, et al. High-speed fixed-target serial virus crystallography. *Nature methods*, 14(8):805–810, 2017.
- [117] Suk-Youl Park, Hyeongju Choi, Cheolsoo Eo, Yunje Cho, and Ki Hyun Nam. Fixed-target serial synchrotron crystallography using nylon mesh and enclosed film-based sample holder. *Crystals*, 10(9):803, 2020.
- [118] Deepshika Gilbile, Megan L Shelby, Artem Y Lyubimov, Jennifer L Wierman, Diana CF Monteiro, Aina E Cohen, Silvia Russi, Matthew A Coleman, Matthias Frank, and Tonya L Kuhl. Plug-and-play polymer microfluidic chips for hydrated, room temperature, fixed-target serial crystallography. *Lab on a Chip*, 21(24):4831–4845, 2021.
- [119] KY Lee, N LaBianca, SA Rishton, S Zolgharnain, JD Gelorme, J Shaw, and TH-P Chang. Micromachining applications of a high resolution ultrathick photoresist. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 13(6):3012–3016, 1995.
- [120] Mark S Hunter, Brent Segelke, Marc Messerschmidt, Garth J Williams, Nadia A Zatsepin, Anton Barty, W Henry Benner, David B Carlson, Matthew Coleman, Alexander Graf, et al. Fixed-target protein serial microcrystallography with an x-ray free electron laser. *Scientific reports*, 4(1):1–5, 2014.
- [121] Jennifer L Wierman, Olivier Paré-Labrosse, Antoine Sarracini, Jessica E Besaw, Michael J Cook, Saeed Oghbaey, Hazem Daoud, Pedram Mehrabi, Irina Kriksunov, Anling Kuo, et al. Fixed-target serial oscillation crystallography at room temperature. *IUCrJ*, 6(2):305–316, 2019.
- [122] Alexander N Asanov, Heather M McDonald, Philip B Oldham, Mark J Jedrzejas, and W William Wilson. Intrinsic fluorescence as a potential rapid scoring tool for protein crystals. *Journal of crystal growth*, 232(1-4):603–609, 2001.
- [123] Leonard MG Chavas, Yusuke Yamada, Masahiko Hiraki, Noriyuki Igarashi, Naohiro Matsugaki, and Soichi Wakatsuki. Uv led lighting for automated crystal centring. *Journal of Synchrotron Radiation*, 18(1):11–15, 2011.
- [124] Xavier Vernede, Bernard Lavault, Jeremy Ohana, Didier Nurizzo, Jacques Joly, Lilian Jacquamet, Franck Felisaz, Florent Cipriani, and Dominique Bourgeois. Uv laser-excited fluorescence as a tool for the visualization of protein crystals mounted in loops. *Acta Crystallographica Section D: Biological Crystallography*, 62(3):253–261, 2006.



- [125] Dominique Bourgeois, Xavier Vernede, Virgile Adam, Emanuela Fioravanti, and Thomas Ursby. A microspectrophotometer for uv–visible absorption and fluorescence studies of protein crystals. *Journal of applied crystallography*, 35(3):319–326, 2002.
- [126] David J Kissick, Christopher M Dettmar, Michael Becker, Anne M Mulichak, Vadim Cherezov, Stephan L Ginell, Kevin P Battaile, Lisa J Keefe, Robert F Fischetti, and Garth J Simpson. Towards protein-crystal centering using second-harmonic generation (shg) microscopy. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):843–851, 2013.
- [127] Thomas D Murray, Artem Y Lyubimov, Craig M Ogata, Huy Vo, Monarin Uervirojnangkoorn, Axel T Brunger, and James M Berger. A high-transparency, micro-patternable chip for x-ray diffraction analysis of microcrystals under native growth conditions. *Acta Crystallographica Section D: Biological Crystallography*, 71(10):1987–1997, 2015.
- [128] Keondo Lee, Jihan Kim, Sangwon Baek, Jaehyun Park, Sehan Park, Jong-Lam Lee, Wan Kyun Chung, Yunje Cho, and Ki Hyun Nam. Combination of an inject-and-transfer system for serial femtosecond crystallography. *bioRxiv*, 2022.
- [129] Kunio Hirata, Kyoko Shinzawa-Itoh, Naomine Yano, Shuhei Takemura, Koji Kato, Miki Hatanaka, Kazumasa Muramoto, Takako Kawahara, Tomitake Tsukihara, Eiki Yamashita, et al. Determination of damage-free crystal structure of an x-ray–sensitive protein using an xfel. *Nature methods*, 11(7):734–736, 2014.
- [130] Irimpan I Mathews, Kim Allison, Thomas Robbins, Artem Y Lyubimov, Monarin Uervirojnangkoorn, Axel T Brunger, Chaitan Khosla, Hasan DeMirci, Scott E McPhillips, Michael Hollenbeck, et al. The conformational flexibility of the acyltransferase from the disorazole polyketide synthase is revealed by an x-ray free-electron laser using a room-temperature sample delivery method for serial crystallography. *Biochemistry*, 56(36):4751–4756, 2017.
- [131] Eike C Schulz, Johannes Kaub, Frederik Busse, Pedram Mehrabi, Henrike M Müller-Werkmeister, Emil F Pai, Wesley D Robertson, and RJ Dwayne Miller. Protein crystals ir laser ablated from aqueous solution at high speed retain their diffractive properties: Applications in high-speed serial crystallography. *Journal of Applied Crystallography*, 50(6):1773–1781, 2017.
- [132] Kresimir Franjic and RJ Dwayne Miller. Vibrationally excited ultrafast thermodynamic phase transitions at the water/air interface. *Physical chemistry chemical physics*, 12(20):5225–5239, 2010.
- [133] Anton Barty, Richard A Kirian, Filipe RNC Maia, Max Hantke, Chun Hong Yoon, Thomas A White, and Henry Chapman. Cheetah: software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data. *Journal of applied crystallography*, 47(3):1118–1131, 2014.
- [134] A Allahgholi, J Becker, L Bianco, A Delfs, R Dinapoli, G Ariño-Estrada, P Goettlicher, Heinz Graafsma, D Greiffenberg, H Hirsemann, et al. Front end ASIC for agipd, a high dynamic range fast detector for the european xfel. *Journal of Instrumentation*, 11(01):C01057, 2016.
- [135] D Mezza, A Allahgholi, A Delfs, R Dinapoli, P Goettlicher, Heinz Graafsma, D Greiffenberg, H Hirsemann, A Klyuev, T Laurus, et al. New calibration circuitry and concept for agipd. *Journal of Instrumentation*, 11(11):C11019, 2016.

- [136] I Steller, R Bolotovskiy, and MG Rossmann. An algorithm for automatic indexing of oscillation images using fourier analysis. *Journal of Applied Crystallography*, 30(6):1036–1040, 1997.
- [137] Wolfgang Kabsch. xds. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):125–132, 2010.
- [138] Michael Krug, Manfred S Weiss, Udo Heinemann, and Uwe Mueller. Xdsapp: a graphical user interface for the convenient processing of diffraction data using xds. *Journal of Applied Crystallography*, 45(3):568–572, 2012.
- [139] Graeme Winter, David G Waterman, James M Parkhurst, Aaron S Brewster, Richard J Gildea, Markus Gerstel, Luis Fuentes-Montero, Melanie Vollmar, Tara Michels-Clark, Iris D Young, et al. Dials: implementation and evaluation of a new integration package. *Acta Crystallographica Section D*, 74(2):85–97, 2018.
- [140] David G Waterman, Graeme Winter, James M Parkhurst, Luis Fuentes-Montero, Johan Hattne, Aaron Brewster, Nicholas K Sauter, Gwyndaf Evans, and P Rosenstrom. The dials framework for integration software. *CCP4 Newslett. Protein Crystallogr*, 49:13–15, 2013.
- [141] Thomas A White, Richard A Kirian, Andrew V Martin, Andrew Aquila, Karol Nass, Anton Barty, and Henry N Chapman. Crystfel: a software suite for snapshot serial crystallography. *Journal of applied crystallography*, 45(2):335–341, 2012.
- [142] Stefan Zaefferer. New developments of computer-aided crystallographic analysis in transmission electron microscopy. *Journal of Applied Crystallography*, 33(1):10–25, 2000.
- [143] Harold R Powell. The rossmann fourier autoindexing algorithm in mosfilm. *Acta Crystallographica Section D: Biological Crystallography*, 55(10):1690–1695, 1999.
- [144] Albert J M Duisenberg. Indexing in single-crystal diffractometry with an obstinate list of reflections. *Journal of applied crystallography*, 25(2):92–96, 1992.
- [145] Helen Mary Ginn, Philip Roedig, Anling Kuo, Gwyndaf Evans, Nicholas K Sauter, Oliver P Ernst, Alke Meents, Henrike Mueller-Werkmeister, RJ Dwayne Miller, and David Ian Stuart. Taketwo: an indexing algorithm suited to still images with known crystal parameters. *Acta Crystallographica Section D: Structural Biology*, 72(8):956–965, 2016.
- [146] Kenneth R Beyerlein, Thomas A White, Oleksandr Yefanov, Cornelius Gati, Ivan G Kazantsev, NF-G Nielsen, Peter M Larsen, Henry N Chapman, and Søren Schmidt. Felix: an algorithm for indexing multiple crystallites in x-ray free-electron laser snapshot diffraction images. *Journal of applied crystallography*, 50(4):1075–1083, 2017.
- [147] Yaroslav Gevorkov, Oleksandr Yefanov, Anton Barty, Thomas A White, Valerio Mariani, Wolfgang Brehm, Aleksandra Tolstikova, R-R Grigat, and Henry N Chapman. Xgandalf—extended gradient descent algorithm for lattice finding. *Acta Crystallographica Section A: Foundations and Advances*, 75(5):694–704, 2019.

- [148] Robert Bücker, Pascal Hogan-Lamarre, Pedram Mehrabi, Eike C Schulz, Lindsey A Bultema, Yaroslav Gevorkov, Wolfgang Brehm, Oleksandr Yefanov, Dominik Oberthür, Günther H Kassier, et al. Serial protein crystallography in an electron microscope. *Nature communications*, 11(1):996, 2020.
- [149] Stef Smeets, Xiaodong Zou, and Wei Wan. Serial electron crystallography for structure determination and phase analysis of nanocrystalline materials. *Journal of applied crystallography*, 51(5):1262–1273, 2018.
- [150] Johan Hattne, Nathaniel Echols, Rosalie Tran, Jan Kern, Richard J Gildea, Aaron S Brewster, Roberto Alonso-Mori, Carina Glöckner, Julia Hellmich, Hartawan Laksmono, et al. Accurate macromolecular structures using minimal measurements from x-ray free-electron lasers. *Nature methods*, 11(5):545–548, 2014.
- [151] Wolfgang Brehm and Kay Diederichs. Breaking the indexing ambiguity in serial crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 70(1):101–109, 2014.
- [152] Richard A Kirian, Xiaoyu Wang, Uwe Weierstall, Kevin E Schmidt, John CH Spence, Mark Hunter, Petra Fromme, Thomas White, Henry N Chapman, and James Holton. Femtosecond protein nanocrystallography—data analysis methods. *Optics express*, 18(6):5713–5723, 2010.
- [153] Richard A Kirian, Thomas A White, James M Holton, Henry N Chapman, Petra Fromme, Anton Barty, Lukas Lomb, Andrew Aquila, Filipe RNC Maia, Andrew V Martin, et al. Structure-factor analysis of femtosecond microdiffraction patterns from protein nanocrystals. *Acta Crystallographica Section A: Foundations of Crystallography*, 67(2):131–140, 2011.
- [154] MG Rossmann and David M Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta crystallographica*, 15(1):24–31, 1962.
- [155] Michael G Rossmann. The molecular replacement method. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(2):73–82, 1990.
- [156] DW Green, Vernon Martin Ingram, and Max Ferdinand Perutz. The structure of haemoglobin-iv. sign determination by the isomorphous replacement method. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 225(1162):287–307, 1954.
- [157] MF Perutz. Isomorphous replacement and phase determination in non-centrosymmetric space groups. *Acta Crystallographica*, 9(11):867–873, 1956.
- [158] DM Blow and MG Rossmann. The single isomorphous replacement method. *Acta crystallographica*, 14(11):1195–1202, 1961.
- [159] David Harker. The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. *Acta Crystallographica*, 9(1):1–9, 1956.
- [160] Jerome Karle. Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology. In *Chemical Crystallography with Pulsed Neutrons and Synchrotron X-rays*, pages 387–397. Springer, 1980.
- [161] Wayne A Hendrickson, Janet L Smith, and Steven Sheriff. Direct phase determination based on anomalous scattering. In *Methods in enzymology*, volume 115, pages 41–55. Elsevier, 1985.

- [162] Janet L Smith. Multiwavelength anomalous diffraction in macromolecular crystallography. In *Direct methods for solving macromolecular structures*, pages 211–225. Springer, 1998.
- [163] Carl Caleman, Francisco Jares Junior, Oscar Grånäs, and Andrew V Martin. A perspective on molecular structure and bond-breaking in radiation damage in serial femtosecond crystallography. *Crystals*, 10(7):585, 2020.
- [164] Oliver B Zeldin, Markus Gerstel, and Elspeth F Garman. Raddose-3d: time-and space-resolved modelling of dose in macromolecular crystallography. *Journal of applied crystallography*, 46(4):1225–1230, 2013.
- [165] CCF Blake and DC Phillips. Effects of x-irradiation on single crystals of myoglobin. *Biological effects of ionizing radiation at the molecular level*, page 183, 1962.
- [166] Martin Weik, Raimond BG Ravelli, Gitay Kryger, Sean McSweeney, Maria L Raves, Michal Harel, Piet Gros, Israel Silman, Jan Kroon, and Joel L Sussman. Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences*, 97(2):623–628, 2000.
- [167] Richard Henderson. Cryo-protection of protein crystals against radiation damage in electron and x-ray diffraction. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 241(1300):6–8, 1990.
- [168] Robin Leslie Owen, Enrique Rudiño-Piñera, and Elspeth F Garman. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proceedings of the National Academy of Sciences*, 103(13):4912–4917, 2006.
- [169] Malcolm R Howells, Tobias Beetz, Henry N Chapman, C Cui, JM Holton, CJ Jacobsen, J Kirz, Enju Lima, Stefano Marchesini, Huijie Miao, et al. An assessment of the resolution limitation due to radiation-damage in x-ray diffraction microscopy. *Journal of electron spectroscopy and related phenomena*, 170(1-3):4–12, 2009.
- [170] Philip Roedig, Ramona Duman, Juan Sanchez-Weatherby, Ismo Vartiainen, Anja Burkhardt, Martin Warmer, Christian David, Armin Wagner, and Alke Meents. Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering. *Journal of applied crystallography*, 49(3):968–975, 2016.
- [171] Anton Barty, Carl Caleman, Andrew Aquila, Nicusor Timneanu, Lukas Lomb, Thomas A White, Jakob Andreasson, David Arnlund, Saša Bajt, Thomas RM Barends, et al. Self-terminating diffraction gates femtosecond x-ray nanocrystallography measurements. *Nature photonics*, 6(1):35–40, 2012.
- [172] Carl Caleman and Andrew V Martin. When diffraction stops and destruction begins. *X-ray Free Electron Lasers: A Revolution in Structural Biology*, pages 185–207, 2018.
- [173] Carl Caleman, Nicușor Timneanu, Andrew V Martin, H Olof Jönsson, Andrew Aquila, Anton Barty, Howard A Scott, Thomas A White, and Henry N Chapman. Ultrafast self-gating bragg diffraction of exploding nanocrystals in an x-ray laser. *Optics express*, 23(2):1213–1231, 2015.

- [174] Lorenzo Galli, S-K Son, M Klinge, S Bajt, A Barty, R Bean, C Betzel, KR Beyerlein, Carl Caleman, RB Doak, et al. Electronic damage in s atoms in a native protein crystal induced by an intense x-ray free-electron laser pulse. *Structural dynamics*, 2(4), 2015.
- [175] Karol Nass, Sébastien Boutet, Andrew Aquila, Thomas RM Barends, R Bruce Doak, Lutz Foucar, and Mario Hilpert. Radiation damage in protein crystallography at x-ray free-electron lasers. In *ACTA CRYSTALLOGRAPHICA A-FOUNDATION AND ADVANCES*, volume 73, pages C1025–C1025. INTERNATIONAL UNION OF PURE AND APPLIED PHYSICS, CHESTER, ENGLAND, 2017.
- [176] Ali Ebrahim, Tadeo Moreno-Chicano, Martin V Appleby, Amanda K Chaplin, John H Beale, Darren A Sherrell, Helen ME Duyvesteyn, Shigeki Owada, Kensuke Tono, Hiroshi Sugimoto, et al. Dose-resolved serial synchrotron and xfel structures of radiation-sensitive metalloproteins. *IUCrJ*, 6(4):543–551, 2019.
- [177] Solange Delagenière, Patrice Brechereau, Ludovic Launer, Alun W Ashton, Ricardo Leal, Stéphanie Veyrier, José Gabadinho, Elspeth J Gordon, Samuel D Jones, Karl Erik Levik, et al. Ispyb: an information management system for synchrotron macromolecular crystallography. *Bioinformatics*, 27(22):3186–3192, 2011.
- [178] SJ Fisher, KE Levik, MA Williams, AW Ashton, and KE McAuley. Synchweb: a modern interface for ispyb. *Journal of Applied Crystallography*, 48(3):927–932, 2015.
- [179] Justyna Aleksandra Wojdyla, Jakub W Kaminski, Ezequiel Panepucci, Simon Ebner, Xiaoqiang Wang, Jose Gabadinho, and Meitian Wang. Da+ data acquisition and analysis software at the swiss light source macromolecular crystallography beamlines. *Journal of Synchrotron Radiation*, 25(1):293–303, 2018.
- [180] Timothy M McPhillips, Scott E McPhillips, H-J Chiu, Aina E Cohen, Ashley M Deacon, Paul J Ellis, Elspeth Garman, Ana Gonzalez, Nicholas K Sauter, R Paul Phizackerley, et al. Blu-ice and the distributed control system: software for data acquisition and instrument control at macromolecular crystallography beamlines. *Journal of synchrotron radiation*, 9(6):401–406, 2002.
- [181] Go Ueno, Hiroyuki Kanda, Takashi Kumasaka, and Masaki Yamamoto. Beamline scheduling software: administration software for automatic operation of the riken structural genomics beamlines at spring-8. *Journal of synchrotron radiation*, 12(3):380–384, 2005.
- [182] John M Skinner, Matt Cowan, Rick Buono, William Nolan, Heinz Bosshard, Howard H Robinson, Annie Héroux, Alexei S Soares, Dieter K Schneider, and Robert M Sweet. Integrated software for macromolecular crystallography synchrotron beamlines ii: revision, robots and a database. *Acta Crystallographica Section D: Biological Crystallography*, 62(11):1340–1347, 2006.
- [183] Yusuke Yamada, Nobuo Phonda, Naohiro Matsugaki, Noriyuki Igarashi, Masahiko Hiraki, and Soichi Wakatsuki. Implementation of remote monitoring and diffraction evaluation systems at the photon factory macromolecular crystallography beamlines. *Journal of Synchrotron Radiation*, 15(3):296–299, 2008.
- [184] José Gabadinho, Antonia Beteva, Matias Guijarro, Vicente Rey-Bakaikoa, Darren Spruce, Matthew W Bowler, Sandor Brockhauser, David Flot, Elspeth J Gordon, David R Hall, et al. Mxcube: a synchrotron beamline control environment customized for macromolecular crystallography experiments. *Journal of synchrotron radiation*, 17(5):700–707, 2010.

- [185] Sergey Stepanov, Oleg Makarov, Mark Hilgart, Sudhir Babu Pothineni, Alex Urakhchin, Satish Devarapalli, Derek Yoder, Michael Becker, Craig Ogata, Ruslan Sanishvili, et al. Jbluice–epics control system for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 67(3):176–188, 2011.
- [186] Graeme Winter and Katherine E McAuley. Automated data collection for macromolecular crystallography. *Methods*, 55(1):81–93, 2011.
- [187] Andrzej Joachimiak. High-throughput crystallography for structural genomics. *Current opinion in structural biology*, 19(5):573–584, 2009.
- [188] Matthew W Bowler, Didier Nurizzo, Ray Barrett, Antonia Beteva, Marjolaine Bodin, Hugo Caserotto, Solange Delagenière, Fabian Dobias, David Flot, Thierry Giraud, et al. Massif-1: a beamline dedicated to the fully automatic characterization and data collection from crystals of biological macromolecules. *Journal of synchrotron radiation*, 22(6):1540–1547, 2015.
- [189] LMG Chavas, N Matsugaki, Y Yamada, M Hiraki, N Igarashi, M Suzuki, and S Wakatsuki. Beamline ar-nw12a: high-throughput beamline for macromolecular crystallography at the photon factory. *Journal of synchrotron radiation*, 19(3):450–454, 2012.
- [190] Michel Fodje, Pawel Grochulski, Kathryn Janzen, Shaunivan Labiuk, James Gorin, and Russ Berg. 08b1-1: an automated beamline for macromolecular crystallography experiments at the canadian light source. *Journal of synchrotron radiation*, 21(3):633–637, 2014.
- [191] Annie Héroux, Marc Allaire, Richard Buono, Matthew L Cowan, Joseph Dvorak, Leon Flaks, Steven LaMarra, Stuart F Myers, Allen M Orville, Howard H Robinson, et al. Macromolecular crystallography beamline x25 at the nsls. *Journal of synchrotron radiation*, 21(3):627–632, 2014.
- [192] Seán McSweeney. Searching for needles in haystacks: Automation and the task of crystal structure determination. In *Advancing Methods for Biomolecular Crystallography*, pages 47–57. Springer, 2013.
- [193] Robin L Owen, Jordi Juanhuix, and Martin Fuchs. Current advances in synchrotron radiation instrumentation for mx experiments. *Archives of Biochemistry and Biophysics*, 602:21–31, 2016.
- [194] Didier Nurizzo, Matthew W Bowler, Hugo Caserotto, Fabien Dobias, Thierry Giraud, John Surr, Nicolas Guichard, Gergely Papp, Matias Guijarro, Christoph Mueller-Dieckmann, et al. Robodiff: combining a sample changer and goniometer for highly automated macromolecular crystallography experiments. *Acta Crystallographica Section D: Structural Biology*, 72(8):966–975, 2016.
- [195] Olof Svensson, Stéphanie Malbet-Monaco, Alexander Popov, Didier Nurizzo, and Matthew W Bowler. Fully automatic characterization and data collection from crystals of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography*, 71(8):1757–1767, 2015.
- [196] Uwe Mueller, Marjolein Thunnissen, Jie Nan, Mikel Eguiraun, Fredrick Bolmsten, Antonio Milàn-Otero, Mathias Guijarro, Markus Oscarsson, Daniele de Sanctis, and Gordon Leonard. Mxcube3: A new era of mx-beamline control begins. *Synchrotron Radiation News*, 30(1):22–27, 2017.
- [197] Vincent Mariaule, Florine Dupeux, and José A Márquez. Estimation of crystallization likelihood through a fluorimetric thermal stability assay. *Structural Genomics: General Applications*, pages 189–195, 2014.

- [198] Florine Dupeux, Martin Röwer, Gael Seroul, Delphine Blot, and José A Márquez. A thermal stability assay can help to estimate the crystallization likelihood of biological samples. *Acta Crystallographica Section D: Biological Crystallography*, 67(11):915–919, 2011.
- [199] Florent Cipriani, Martin Röwer, Christophe Landret, Ulrich Zander, Franck Felisaz, and José Antonio Márquez. Crystaldirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallographica Section D: Biological Crystallography*, 68(10):1393–1399, 2012.
- [200] José A Márquez and Florent Cipriani. Crystaldirect™: a novel approach for automated crystal harvesting based on photoablation of thin films. *Structural Genomics: General Applications*, pages 197–203, 2014.
- [201] Ulrich Zander, Guillaume Hoffmann, Irina Cornaciu, J-P Marquette, Gergely Papp, Christophe Landret, Gaël Seroul, Jérémy Sinoir, Martin Röwer, Frank Felisaz, et al. Automated harvesting and processing of protein crystals through laser photoablation. *Acta Crystallographica Section D: Structural Biology*, 72(4):454–466, 2016.
- [202] Irina Cornaciu, Raphael Bourgeas, Guillaume Hoffmann, Florine Dupeux, Anne-Sophie Humm, Vincent Mariaule, Andrea Pica, Damien Clavel, Gael Seroul, Peter Murphy, et al. The automated crystallography pipelines at the embl htx facility in grenoble. *JoVE (Journal of Visualized Experiments)*, (172):e62491, 2021.
- [203] Thomas Ursby, Karl Åhnberg, Roberto Appio, Oskar Aurelius, Artur Barczyk, Antonio Bartalesi, Monika Bjelčić, Fredrik Bolmsten, Yngve Cerenius, R Bruce Doak, et al. Biomax—the first macromolecular crystallography beamline at max iv laboratory. *Journal of Synchrotron Radiation*, 27(5):1415–1429, 2020.
- [204] Anastasya Shilova, Hugo Lebrette, Oskar Aurelius, Jie Nan, Martin Welin, Rebeka Kovacic, Swagatha Ghosh, Cecilia Safari, Ross J Friel, Mirko Milas, et al. Current status and future opportunities for serial crystallography at max iv laboratory. *Journal of Synchrotron Radiation*, 27(5):1095–1102, 2020.
- [205] Stéphanie Monaco, Elspeth Gordon, Matthew W Bowler, Solange Delageniere, Matias Guijarro, Darren Spruce, Olof Svensson, Sean M McSweeney, Andrew A McCarthy, Gordon Leonard, et al. Automatic processing of macromolecular crystallography x-ray diffraction data at the esrf. *Journal of applied crystallography*, 46(3):804–810, 2013.
- [206] Daniil Prigozhin, Jeff Dickert, John Taylor, Randall Cayford, Kevin Royal, Anthony Rozales, Stacey Ortega, Adrian Spucces, Jay Nix, and Marc Allaire<sup>10</sup>. The berkeley center for structural biology: A suite of macromolecular crystallography beamlines at the advanced light source. *Foundations of Crystallography*, 78:a289, 2022.
- [207] Pawel Grochulski, Michel Fodje, Shaun Labiuk, Tomasz W Wysokinski, George Belev, Malgorzata Korbas, and Scott M Rosendahl. Review of canadian light source facilities for biological applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 411:17–21, 2017.
- [208] Leo R Dalesio, Jeffrey O Hill, Martin Kraimer, Stephen Lewis, Douglas Murray, Stephan Hunt, William Watson, Matthias Clausen, and John Dalesio. The experimental physics and industrial control system

architecture: past, present, and future. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 352(1-2):179–184, 1994.

- [209] Michel Fodje, Kathryn Janzen, Russ Berg, Gillian Black, Shaunivan Labiuk, James Gorin, and Pawel Grochulski. Mxdc and mxlive: software for data acquisition, information management and remote access to macromolecular crystallography beamlines. *Journal of Synchrotron Radiation*, 19(2):274–280, 2012.
- [210] Wolfgang Kabsch. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of applied crystallography*, 26(6):795–800, 1993.
- [211] Philip Evans. Scaling and assessment of data quality. *Acta Crystallographica Section D: Biological Crystallography*, 62(1):72–82, 2006.
- [212] Gleb P Bourenkov and Alexander N Popov. A quantitative approach to data-collection strategies. *Acta Crystallographica Section D: Biological Crystallography*, 62(1):58–64, 2006.
- [213] Kay Diederichs. Some aspects of quantitative analysis and correction of radiation damage. *Acta Crystallographica Section D: Biological Crystallography*, 62(1):96–101, 2006.
- [214] Martyn D Winn, Charles C Ballard, Kevin D Cowtan, Eleanor J Dodson, Paul Emsley, Phil R Evans, Ronan M Keegan, Eugene B Krissinel, Andrew GW Leslie, Airlie McCoy, et al. Overview of the ccp4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):235–242, 2011.
- [215] Paul D Adams, Pavel V Afonine, Gábor Bunkóczi, Vincent B Chen, Ian W Davis, Nathaniel Echols, Jeffrey J Headd, L-W Hung, Gary J Kapral, Ralf W Grosse-Kunstleve, et al. Phenix: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):213–221, 2010.
- [216] Qi-Sheng Wang, Kun-Hao Zhang, Yin Cui, Zhi-Jun Wang, Qiang-Yan Pan, Ke Liu, Bo Sun, Huan Zhou, Min-Jun Li, Qin Xu, et al. Upgrade of macromolecular crystallography beamline bl17u1 at ssrf. *Nuclear Science and Techniques*, 29:1–7, 2018.
- [217] Clemens Vornrhein, Claus Flensburg, Peter Keller, Andrew Sharff, Oliver Smart, Wlodek Paciorek, Thomas Womack, and Gérard Bricogne. Data processing and analysis with the autoproc toolbox. *Acta Crystallographica Section D: Biological Crystallography*, 67(4):293–302, 2011.
- [218] Ralf W Grosse-Kunstleve, Nicholas K Sauter, Nigel W Moriarty, and Paul D Adams. The computational crystallography toolbox: crystallographic algorithms in a reusable software framework. *Journal of Applied Crystallography*, 35(1):126–136, 2002.
- [219] Isabel Usón and George M Sheldrick. An introduction to experimental phasing of macromolecules illustrated by shelx; new autotracing features. *Acta Crystallographica Section D: Structural Biology*, 74(2):106–116, 2018.
- [220] Keitaro Yamashita, Kunio Hirata, and Masaki Yamamoto. Kamo: towards automated data processing for microcrystals. *Acta Crystallographica Section D: Structural Biology*, 74(5):441–449, 2018.



- [221] Valerio Mariani, Andrew Morgan, Chun Hong Yoon, Thomas J Lane, Thomas A White, Christopher O'Grady, Manuela Kuhn, Steve Aplin, Jason Koglin, Anton Barty, et al. Onda: online data analysis and feedback for serial x-ray imaging. *Journal of applied crystallography*, 49(3):1073–1080, 2016.
- [222] Kunio Hirata, Keitaro Yamashita, Go Ueno, Yoshiaki Kawano, Kazuya Hasegawa, Takashi Kumasaka, and Masaki Yamamoto. Zoo: an automatic data-collection system for high-throughput structure analysis in protein microcrystallography. *Acta Crystallographica Section D: Structural Biology*, 75(2):138–150, 2019.
- [223] Karthik S Paithankar and Elspeth F Garman. Know your dose: Raddose. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):381–388, 2010.
- [224] Takanori Nakane, Yasumasa Joti, Kensuke Tono, Makina Yabashi, Eriko Nango, So Iwata, Ryuichiro Ishitani, and Osamu Nureki. Data processing pipeline for serial femtosecond crystallography at sacla. *Journal of applied crystallography*, 49(3):1035–1041, 2016.
- [225] Shibom Basu, Jakub W Kaminski, Ezequiel Panepucci, C-Y Huang, Rangana Warshamanage, Meitian Wang, and Justyna Aleksandra Wojdyla. Automated data collection and real-time data analysis suite for serial synchrotron crystallography. *Journal of synchrotron radiation*, 26(1):244–252, 2019.
- [226] Dieter K Schneider, Alexei S Soares, Edwin O Lazo, Dale F Kreidler, Kun Qian, Martin R Fuchs, Dileep K Bhogadi, Steve Antonelli, Stuart S Myers, Bruno S Martins, et al. Amx—the highly automated macromolecular crystallography (17-id-1) beamline at the nsls-ii. *Journal of Synchrotron Radiation*, 29(6), 2022.
- [227] Dieter K Schneider, Wuxian Shi, Babak Andi, Jean Jakoncic, Yuan Gao, Dileep K Bhogadi, Stuart F Myers, Bruno Martins, John M Skinner, Jun Aishima, et al. Fmx—the frontier microfocusing macromolecular crystallography beamline at the national synchrotron light source ii. *Journal of Synchrotron Radiation*, 28(2):650–665, 2021.
- [228] DK Schneider, LE Berman, O Chubar, WA Hendrickson, SL Hulbert, M Lucas, RM Sweet, and L Yang. Three biomedical beamlines at nsls-ii for macromolecular crystallography and small-angle scattering. In *Journal of Physics: Conference Series*, volume 425, page 012003. IOP Publishing, 2013.
- [229] Martin R Fuchs, Dileep K Bhogadi, Jean Jakoncic, Stuart Myers, Robert M Sweet, Lonny E Berman, John Skinner, Mourad Idir, Oleg Chubar, Sean McSweeney, et al. Nsls-ii biomedical beamlines for microcrystallography, fmx, and for highly automated crystallography, amx: New opportunities for advanced data collection. In *AIP Conference Proceedings*, volume 1741, page 030006. AIP Publishing LLC, 2016.
- [230] MR Fuchs, RM Sweet, LE Berman, WA Hendrickson, O Chubar, N Canestrari, M Idir, L Yang, and DK Schneider. Nsls-ii biomedical beamlines for macromolecular crystallography, fmx and amx, and for x-ray scattering, lix: current developments. In *Journal of Physics: Conference Series*, volume 493, page 012021. IOP Publishing, 2014.
- [231] George M Sheldrick. Crystal structure refinement with shelxl. *Acta Crystallographica Section C: Structural Chemistry*, 71(1):3–8, 2015.

- [232] Gongrui Guo, Ping Zhu, Martin R Fuchs, Wuxian Shi, Babak Andi, Yuan Gao, Wayne A Hendrickson, Sean McSweeney, and Qun Liu. Synchrotron microcrystal native-sad phasing at a low energy. *IUCrJ*, 6(4):532–542, 2019.
- [233] Lina Takemaru, Gongrui Guo, Ping Zhu, Wayne A Hendrickson, Sean McSweeney, and Qun Liu. Pymda: microcrystal data assembly using python. *Journal of Applied Crystallography*, 53(1):277–281, 2020.
- [234] Philip R Evans and Garib N Murshudov. How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography*, 69(7):1204–1214, 2013.
- [235] Yuan Gao, Weihe Xu, Wuxian Shi, Alexei Soares, Jean Jakoncic, Stuart Myers, Bruno Martins, John Skinner, Qun Liu, Herbert Bernstein, et al. High-speed raster-scanning synchrotron serial microcrystallography with a high-precision piezo-scanner. *Journal of synchrotron radiation*, 25(5):1362–1370, 2018.
- [236] Ulrich Zander, Gleb Bourenkov, Alexander N Popov, Daniele De Sanctis, Olof Svensson, Andrew A McCarthy, Ekaterina Round, Valentin Gordeliy, Christoph Mueller-Dieckmann, and Gordon A Leonard. Meshandcollect: an automated multi-crystal data-collection workflow for synchrotron macromolecular crystallography beamlines. *Acta Crystallographica Section D: Biological Crystallography*, 71(11):2328–2343, 2015.
- [237] Nicholas K Sauter, Johan Hattne, Ralf W Grosse-Kunstleve, and Nathaniel Echols. New python-based methods for data processing. *Acta Crystallographica Section D: Biological Crystallography*, 69(7):1274–1282, 2013.
- [238] Zbyszek Otwinowski and Wladek Minor. [20] processing of x-ray diffraction data collected in oscillation mode. In *Methods in enzymology*, volume 276, pages 307–326. Elsevier, 1997.
- [239] Thomas Pape and Thomas R Schneider. Hkl2map: a graphical user interface for macromolecular phasing with shelx programs. *Journal of applied crystallography*, 37(5):843–844, 2004.
- [240] P Andrew Karplus and Kay Diederichs. Assessing and maximizing data quality in macromolecular crystallography. *Current opinion in structural biology*, 34:60–68, 2015.
- [241] P Andrew Karplus and Kay Diederichs. Linking crystallographic model and data quality. *Science*, 336(6084):1030–1033, 2012.
- [242] Greta Assmann, Wolfgang Brehm, and Kay Diederichs. Identification of rogue datasets in serial crystallography. *Journal of applied crystallography*, 49(3):1021–1028, 2016.
- [243] Axel T Brünger. Free r value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, 1992.
- [244] Thomas C Terwilliger, Gábor Bunkóczi, L-W Hung, Peter H Zwart, Janet L Smith, David L Akey, and Paul D Adams. Can i solve my structure by sad phasing? planning an experiment, scaling data and evaluating the useful anomalous correlation and anomalous signal. *Acta Crystallographica Section D: Structural Biology*, 72(3):359–374, 2016.

- [245] Takanori Nakane, Shinya Hanashima, Mamoru Suzuki, Haruka Saiki, Taichi Hayashi, Keisuke Kakinouchi, Shigeru Sugiyama, Satoshi Kawatake, Shigeru Matsuoka, Nobuaki Matsumori, et al. Membrane protein structure determination by sad, sir, or siras phasing in serial femtosecond crystallography using an iododetergent. *Proceedings of the National Academy of Sciences*, 113(46):13039–13044, 2016.
- [246] Keitaro Yamashita, Naoyuki Kuwabara, Takanori Nakane, Tomohiro Murai, Eiichi Mizohata, Michihiro Sugahara, Dongqing Pan, Tetsuya Masuda, Mamoru Suzuki, Tomomi Sato, et al. Experimental phase determination with selenomethionine or mercury-derivatization in serial femtosecond crystallography. *IUCrJ*, 4(5):639–647, 2017.
- [247] Kay Diederichs and P Andrew Karplus. Improved r-factors for diffraction data analysis in macromolecular crystallography. *Nature structural biology*, 4(4):269–275, 1997.
- [248] Robbie P Joosten, Fei Long, Garib N Murshudov, and Anastassis Perrakis. The pdb\_redo server for macromolecular structure model optimization. *IUCrJ*, 1(4):213–220, 2014.
- [249] Karol Nass, Robert Cheng, Laura Vera, Aldo Mozzanica, Sophie Redford, Dmitry Ozerov, Shibom Basu, Daniel James, Gregor Knopp, Claudio Cirelli, et al. Advances in long-wavelength native phasing at x-ray free-electron lasers. *IUCrJ*, 7(6):965–975, 2020.
- [250] Oleksandr Yefanov, Cornelius Gati, Gleb Bourenkov, Richard A Kirian, Thomas A White, John CH Spence, Henry N Chapman, and Anton Barty. Mapping the continuous reciprocal space intensity distribution of x-ray serial crystallography. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647):20130333, 2014.
- [251] Wea Kabsch. Automatic indexing of rotation diffraction patterns. *Journal of Applied Crystallography*, 21(1):67–72, 1988.
- [252] Anja Burkhardt, Tim Pakendorf, Bernd Reime, Jan Meyer, Pontus Fischer, Nicolas Stübe, Saravanan Panneerselvam, Olga Lorbeer, Karolina Stachnik, Martin Warmer, et al. Status of the crystallography beamlines at petra iii. *The European Physical Journal Plus*, 131:1–9, 2016.
- [253] Eugenio de La Mora, Nicolas Coquelle, Charles S Bury, Martin Rosenthal, James M Holton, Ian Carmichael, Elspeth F Garman, Manfred Burghammer, Jacques-Philippe Colletier, and Martin Weik. Radiation damage and dose limits in serial synchrotron crystallography at cryo-and room temperatures. *Proceedings of the National Academy of Sciences*, 117(8):4142–4151, 2020.
- [254] Clemens Vornrhein, Ian J Tickle, Claus Flensburg, Peter Keller, Wlodek Paciorek, Andrew Sharff, and Gerard Bricogne. 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, 2018.
- [255] Marek Grabowski, Marcin Cymborowski, Przemyslaw J Porebski, Tomasz Osinski, Ivan G Shabalin, David R Cooper, and Wlodek Minor. The integrated resource for reproducibility in macromolecular crystallography: Experiences of the first four years. *Structural Dynamics*, 6(6), 2019.
- [256] Monica Duke, Michael Day, Rachel Heery, Leslie A Carr, and Simon J Coles. Enhancing access to research data: the challenge of crystallography. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital libraries*, pages 46–55, 2005.

- [257] Jürgen Bernard, Tobias Ruppert, Maximilian Scherer, Jörn Kohlhammer, and Tobias Schreck. Content-based layouts for exploratory metadata search in scientific research data. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 139–148, 2012.
- [258] John R Helliwell. Fact and fair with big data allows objectivity in science: The view of crystallography. *Structural Dynamics*, 6(5), 2019.
- [259] Herbert J Bernstein, Lawrence C Andrews, Jorge A Diaz, Jean Jakoncic, Thu Nguyen, Nicholas K Sauter, Alexei S Soares, Justin Y Wei, Maciej R Wlodek, and Mario A Xerri. Best practices for high data-rate macromolecular crystallography (hdrmx). *Structural Dynamics*, 7(1), 2020.
- [260] A Vondrous, T Jejkal, D Ressmann, W Mexner, and R Stotzka. Beamline data management at the synchrotron anka.
- [261] Matias Guijarro, A Beteva, G Berruyer, T Coutinho, L Claustre, S Debionne, MC Dominguez, P Guillou, C Guilloud, A Homs, et al. Beamline experiments at esrf with bliss. In *17th International Conference on Accelerator and Large Experimental Physics Control Systems*, page MOCPL03, 2019.
- [262] Brian H Toby, Yu Huang, Don Dohan, David Carroll, Xuesong Jiao, Lynn Ribaud, Jennifer A Doebbler, Matthew R Suchomel, Jun Wang, Curt Preissner, et al. Management of metadata and automation for mail-in measurements with the aps 11-bm high-throughput, high-resolution synchrotron powder diffractometer. *Journal of Applied Crystallography*, 42(6):990–993, 2009.
- [263] Sebastian Guenther, Patrick YA Reinke, Yaiza Fernandez-Garcia, Julia Lieske, Thomas J Lane, Helen Ginn, Faisal Koua, Christiane Ehrt, Wiebke Ewert, Dominik Oberthuer, et al. Massive x-ray screening reveals two allosteric drug binding sites of sars-cov-2 main protease (preprint). *Biorxiv*, 2020.
- [264] Linda C Johansson, David Arnlund, Gergely Katona, Thomas A White, Anton Barty, Daniel P DePonte, Robert L Shoeman, Cecilia Wickstrand, Amit Sharma, Garth J Williams, et al. Structure of a photosynthetic reaction centre determined by serial femtosecond crystallography. *Nature communications*, 4(1):2911, 2013.
- [265] Lars Redecke, Karol Nass, Daniel P DePonte, Thomas A White, Dirk Rehders, Anton Barty, Francesco Stellato, Mengning Liang, Thomas RM Barends, Sébastien Boutet, et al. Natively inhibited trypanosoma brucei cathepsin b structure determined by using an x-ray laser. *Science*, 339(6116):227–230, 2013.
- [266] Juraj Knoška, Luigi Adriano, Salah Awel, Kenneth R Beyerlein, Oleksandr Yefanov, Dominik Oberthuer, Gisel E Peña Murillo, Nils Roth, Iosifina Sarrou, Pablo Villanueva-Perez, et al. Ultracompact 3d microfluidics for time-resolved structural biology. *Nature communications*, 11(1):657, 2020.
- [267] B Henrich, A Bergamaschi, C Broennimann, R Dinapoli, EF Eikenberry, I Johnson, M Kobas, P Kraft, A Mozzanica, and B Schmitt. Pilatus: A single photon counting pixel detector for x-ray applications. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 607(1):247–249, 2009.
- [268] Pavol Skubák and Navraj S Pannu. Automatic protein structure solution from weak x-ray data. *Nature communications*, 4(1):2777, 2013.

- [269] Philip Hart, Sébastien Boutet, Gabriella Carini, Mikhail Dubrovin, Brian Duda, David Fritz, Gunther Haller, Ryan Herbst, Sven Herrmann, Chris Kenney, et al. The cspad megapixel x-ray camera at lcls. In *X-Ray free-electron lasers: beam diagnostics, beamline instrumentation, and applications*, volume 8504, pages 51–61. SPIE, 2012.
- [270] Hugh T Philipp, Marianne Hromalik, Mark Tate, Lucas Koerner, and Sol M Gruner. Pixel array detector for x-ray free electron laser experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 649(1):67–69, 2011.
- [271] Max O Wiedorn, Dominik Oberthür, Richard Bean, Robin Schubert, Nadine Werner, Brian Abbey, Martin Aepfelbacher, Luigi Adriano, Aschkan Allahgholi, Nasser Al-Qudami, et al. Megahertz serial crystallography. *Nature communications*, 9(1):4025, 2018.
- [272] Jeanette Held and Sander van Smaalen. The active site of hen egg-white lysozyme: flexibility and chemical bonding. *Acta Crystallographica Section D: Biological Crystallography*, 70(4):1136–1146, 2014.
- [273] Kartik Ayyer, Hugh T Philipp, Mark W Tate, Jennifer L Wierman, Veit Elser, and Sol M Gruner. Determination of crystallographic intensities from sparse data. *IUCrJ*, 2(1):29–34, 2015.
- [274] Kartik Ayyer, Oleksandr M Yefanov, Dominik Oberthür, Shatabdi Roy-Chowdhury, Lorenzo Galli, Valerio Mariani, Shibom Basu, Jesse Coe, Chelsie E Conrad, Raimund Fromme, et al. Macromolecular diffractive imaging using imperfect crystals. *Nature*, 530(7589):202–206, 2016.
- [275] Airlie J McCoy, Ralf W Grosse-Kunstleve, Paul D Adams, Martyn D Winn, Laurent C Storoni, and Randy J Read. Phaser crystallographic software. *Journal of applied crystallography*, 40(4):658–674, 2007.
- [276] Paul Emsley, Bernhard Lohkamp, William G Scott, and Kevin Cowtan. Features and development of coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):486–501, 2010.
- [277] Roberto Alonso-Mori, Dimosthenis Sokaras, Marco Cammarata, Yuantao Ding, Yiping Feng, David Fritz, Kelly J Gaffney, Jerome Hastings, Chi-Chang Kao, Henrik T Lemke, et al. Femtosecond electronic structure response to high intensity xfel pulses probed by iron x-ray emission spectroscopy. *Scientific reports*, 10(1):16837, 2020.
- [278] Jason R Stagno, Juraj Knoska, Henry N Chapman, and Yun-Xing Wang. Mix-and-inject serial femtosecond crystallography to capture rna riboswitch intermediates. In *RNA Structure and Dynamics*, pages 243–249. Springer, 2022.
- [279] Thomas RM Barends, Benjamin Stauch, Vadim Cherezov, and Ilme Schlichting. Serial femtosecond crystallography. *Nature Reviews Methods Primers*, 2(1):59, 2022.
- [280] Roberto Dinapoli, Anna Bergamaschi, Beat Henrich, Roland Horisberger, Ian Johnson, Aldo Mozzanica, Elmar Schmid, Bernd Schmitt, Akos Schreiber, Xintian Shi, et al. Eiger: Next generation single photon counting detector for x-ray applications. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 650(1):79–83, 2011.

- [281] A Mozzanica, M Andrä, R Barten, A Bergamaschi, S Chiriotti, M Brückner, R Dinapoli, E Fröjd, D Greiffenberg, F Leonarski, et al. The jungfrau detector for applications at synchrotron light sources and xfels. *Synchrotron Radiation News*, 31(6):16–20, 2018.
- [282] A Mozzanica, A Bergamaschi, M Brueckner, S Cartier, R Dinapoli, D Greiffenberg, J Jungmann-Smith, D Maliakal, D Mezza, M Ramilli, et al. Characterization results of the jungfrau full scale readout asic. *Journal of Instrumentation*, 11(02):C02047, 2016.
- [283] Filip Leonarski, Aldo Mozzanica, Martin Brückner, Carlos Lopez-Cuenca, Sophie Redford, Leonardo Sala, Andrej Babic, Heinrich Billich, Oliver Bunk, Bernd Schmitt, et al. Jungfrau detector for brighter x-ray sources: Solutions for it and data science challenges in macromolecular crystallography. *Structural Dynamics*, 7(1):014305, 2020.
- [284] D Pennicard, S Lange, S Smoljanin, H Hirsemann, H Graafsma, M Epple, M Zuvic, MO Lampert, T Fritsch, and M Rothermund. The lambda photon-counting pixel detector. In *Journal of Physics: Conference Series*, volume 425, page 062010. IOP Publishing, 2013.
- [285] A Dragone, P Caragiulo, B Markovic, R Herbst, K Nishimura, B Reese, S Herrmann, P Hart, G Blaj, J Segal, et al. epix: a class of front-end asics for second generation lcls integrating hybrid pixel detectors. In *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, pages 1–5. IEEE, 2013.
- [286] B Henrich, J Becker, R Dinapoli, P Goettlicher, H Graafsma, H Hirsemann, R Klanner, H Krueger, R Mazzocco, A Mozzanica, et al. The adaptive gain integrating pixel detector agipd a detector for the european xfel. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 633:S11–S14, 2011.
- [287] Mathew Hart, C Angelsen, S Burge, J Coughlan, R Halsall, A Koch, M Kuster, T Nicholls, M Prydderch, P Seller, et al. Development of the lpd, a high dynamic range pixel detector for the european xfel. In *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, pages 534–537. IEEE, 2012.
- [288] Andreas Koch, Matthew Hart, Tim Nicholls, Christian Angelsen, John Coughlan, Marcus French, Steffen Hauf, Markus Kuster, Jolanta Sztuk-Dambietz, Monica Turcato, et al. Performance of an lpd prototype detector at mhz frame rates under synchrotron and fel radiation. *Journal of Instrumentation*, 8(11):C11001, 2013.
- [289] M Porro, L Andricek, Luca Bombelli, G De Vita, C Fiorini, P Fischer, K Hansen, P Lechner, G Lutz, L Strüder, et al. Expected performance of the depfet sensor with signal compression: A large format x-ray imager with mega-frame readout capability for the european xfel. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 624(2):509–519, 2010.
- [290] Matteo Porro, Ladislav Andricek, Andrea Castoldi, Carlo Fiorini, Peter Fischer, Heinz Graafsma, Karsten Hansen, Andreas Kugel, Gerhard Lutz, Ullrich Pietsch, et al. Large format x-ray imager with mega-frame readout capability for xfel, based on the depfet active pixel sensor. In *2008 IEEE Nuclear Science Symposium Conference Record*, pages 1578–1586. IEEE, 2008.

- [291] Markus Kuster, Djelloul Boukhelef, Mattia Donato, J-S Dambietz, Steffen Hauf, Luis Maia, Natascha Raab, Janusz Szuba, Monica Turcato, Krzysztof Wrona, et al. Detectors and calibration concept for the european xfel. *Synchrotron radiation news*, 27(4):35–38, 2014.
- [292] Jolanta Sztuk-Dambietz, Steffen Hauf, Andreas Koch, Markus Kuster, and Monica Turcato. Status of detector development for the european xfel. In *Advances in X-ray Free-Electron Lasers II: Instrumentation*, volume 8778, pages 112–124. SPIE, 2013.
- [293] Dionisio Doering, Maciej Kwiatkowski, Umanath R Kamath, Camillo Tamma, Lorenzo Rota, Larry Ruckman, Ryan T Herbst, Benjamin A Reese, Pietro Caragiulo, Gabriel Blaj, et al. Readout system for epxhr x-ray detectors: A framework and case study. In *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–4. IEEE, 2020.
- [294] Xavier Delaunay, Aurélie Courtois, and Flavien Gouillon. Evaluation of lossless and lossy algorithms for the compression of scientific datasets in netcdf-4 or hdf5 files. *Geoscientific Model Development*, 12(9):4099–4113, 2019.
- [295] Kira Duwe, Jakob Lüttgau, Georgiana Mania, Jannek Squar, Anna Fuchs, Michael Kuhn, Eugen Betke, and Thomas Ludwig. State of the art and future trends in data reduction for high-performance computing. *Supercomputing Frontiers and Innovations*, 7(1):4–36, 2020.
- [296] Dingwen Tao, Sheng Di, Zizhong Chen, and Franck Cappello. Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 1129–1139. IEEE, 2017.
- [297] J Kieffer, S Petitedemange, and T Vincent. Real-time diffraction computed tomography data reduction. *Journal of Synchrotron Radiation*, 25(2):612–617, 2018.
- [298] Abhik Datta, Kian Fong Ng, Deepan Balakrishnan, Melissa Ding, See Wee Chee, Yvonne Ban, Jian Shi, and N Duane Loh. A data reduction and compression description for high throughput time-resolved electron microscopy. *Nature communications*, 12(1):1–15, 2021.
- [299] Bálint Balázs, Joran Deschamps, Marvin Albert, Jonas Ries, and Lars Hufnagel. A real-time compression library for microscopy images. *bioRxiv*, page 164624, 2017.
- [300] Lars Loetgering, Max Rose, David Treffer, Ivan A Vartanyants, Axel Rosenhahn, and Thomas Wilhein. Data compression strategies for ptychographic diffraction imaging. *Advanced Optical Technologies*, 6(6):475–483, 2017.
- [301] Klaus Wakonig, H-C Stadler, Michal Odstrčil, Esther HR Tsai, Ana Diaz, Mirko Holler, Ivan Usov, Jörg Raabe, Andreas Menzel, and Manuel Guizar-Sicairos. Ptychoshelves, a versatile high-level framework for high-performance analysis of ptychographic data. *Journal of applied crystallography*, 53(2):574–586, 2020.
- [302] Panpan Huang, Ming Du, Mike Hammer, Antonino Miceli, and Chris Jacobsen. Fast digital lossy compression for x-ray ptychographic data. *Journal of synchrotron radiation*, 28(1):292–300, 2021.

- [303] Yair Wiseman and Erick Fredj. Contour extraction of compressed jpeg images. *Journal of Graphics Tools*, 6(3):37–43, 2001.
- [304] Amhamed Saffor, Abdul Rahman Ramli, and Kwan-Hoong Ng. A comparative study of image compression between jpeg and wavelet. *Malaysian Journal of computer science*, 14(1):39–45, 2001.
- [305] Federica Marone, Jakob Vogel, and Marco Stampanoni. Impact of lossy compression of x-ray projections onto reconstructed tomographic slices. *Journal of Synchrotron Radiation*, 27(5):1326–1338, 2020.
- [306] Majid Rabbani and Rajan Joshi. An overview of the jpeg 2000 still image compression standard. *Signal processing: Image communication*, 17(1):3–48, 2002.
- [307] Kiyoshi Masui, Mandana Amiri, Liam Connor, Meiling Deng, Mateus Fandino, Carolin Höfer, Mark Halpern, David Hanna, Adam D Hincks, Gary Hinshaw, et al. A compression scheme for radio data in high performance computing. *Astronomy and Computing*, 12:181–190, 2015.
- [308] Xin Liang, Sheng Di, Dingwen Tao, Zizhong Chen, and Franck Cappello. An efficient transformation scheme for lossy data compression with point-wise relative error bound. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 179–189. IEEE, 2018.
- [309] Ludmila Urzhumtseva, Pavel V Afonine, Paul D Adams, and Alexandre Urzhumtsev. Crystallographic model quality at a glance. *Acta Crystallographica Section D: Biological Crystallography*, 65(3):297–300, 2009.
- [310] Ian W Davis, Andrew Leaver-Fay, Vincent B Chen, Jeremy N Block, Gary J Kapral, Xueyi Wang, Laura W Murray, W Bryan Arendall III, Jack Snoeyink, Jane S Richardson, et al. Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research*, 35(suppl\_2):W375–W383, 2007.
- [311] Cornelius Gati, Dominik Oberthuer, Oleksandr Yefanov, Richard D Bunker, Francesco Stellato, Elaine Chiu, Shin-Mei Yeh, Andrew Aquila, Shibom Basu, Richard Bean, et al. Atomic structure of granulin determined from native nanocrystalline granulovirus using an x-ray free-electron laser. *Proceedings of the National Academy of Sciences*, 114(9):2247–2252, 2017.
- [312] Robert Underwood, Chun Yoon, Ali Gok, Sheng Di, and Franck Cappello. Roibin-sz: Fast and science-preserving compression for serial crystallography. *arXiv preprint arXiv:2206.11297*, 2022.
- [313] John CH Spence. Serial crystallography: Preface. *Crystals*, 10(2):135, 2020.
- [314] Oleksandr Yefanov, Dominik Oberthür, Richard Bean, Max O Wiedorn, Juraj Knoska, Gisel Pena, Salah Awel, Lars Gumprecht, Martin Domaracky, Iosifina Sarrou, et al. Evaluation of serial crystallographic structure determination within megahertz pulse trains. *Structural Dynamics*, 6(6):064702, 2019.
- [315] Chris Gisriel, Jesse Coe, Romain Letrun, Oleksandr M Yefanov, Cesar Luna-Chavez, Natasha E Stander, Stella Lisova, Valerio Mariani, Manuela Kuhn, Steve Aplin, et al. Membrane protein megahertz crystallography at the european xfel. *Nature communications*, 10(1):5021, 2019.



- [316] Jason Tenboer, Shibom Basu, Nadia Zatsepin, Kanupriya Pande, Despina Milathianaki, Matthias Frank, Mark Hunter, Sébastien Boutet, Garth J Williams, Jason E Koglin, et al. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science*, 346(6214):1242–1246, 2014.
- [317] A Tolstikova, V Mariani, TD Grant, A Barty, et al. Om and cheetah: a common framework for online and offline analysis in serial crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 78:a408–a408, 2023.
- [318] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58, 2020.
- [319] Artur Souza, Leonardo B Oliveira, Sabine Hollatz, Matt Feldman, Kunle Olukotun, James M Holton, Aina E Cohen, and Luigi Nardi. Deepfreak: Learning crystallography diffraction patterns with automated machine learning. *arXiv preprint arXiv:1904.11834*, 2019.
- [320] Vahid Rahmani, Shah Nawaz, David Pennicard, Shabarish Pala Ramakantha Setty, and Heinz Graafsma. Data reduction for x-ray serial crystallography using machine learning. *Journal of Applied Crystallography*, 56(1), 2023.
- [321] T-W Ke, Aaron S Brewster, Stella X Yu, Daniela Ushizima, Chao Yang, and Nicholas K Sauter. A convolutional neural network-based screening tool for x-ray serial crystallography. *Journal of synchrotron radiation*, 25(3):655–670, 2018.
- [322] PP Ewald. Introduction to the dynamical theory of x-ray diffraction. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 25(1):103–108, 1969.
- [323] Mukul Sonker, Diandra Doppler, Ana Egatz-Gomez, Sahba Zaare, Mohammad T Rabbani, Abhik Manna, Jorvani Cruz Villarreal, Garrett Nelson, Gihan K Ketawala, Konstantinos Karpos, et al. Electrically stimulated droplet injector for reduced sample consumption in serial crystallography. *Biophysical Reports*, 2(4):100081, 2022.
- [324] MW Parker. Protein structure from x-ray diffraction. *Journal of biological physics*, 29(4):341–362, 2003.
- [325] J Marquez. Online crystallography: Fully automated, remote controlled protein-to-structure pipelines. In *ACTA CRYSTALLOGRAPHICA A-FOUNDATION AND ADVANCES*, volume 76, pages A214–A214. INT UNION CRYSTALLOGRAPHY 2 ABBEY SQ, CHESTER, CH1 2HU, ENGLAND, 2020.
- [326] José A Márquez. Online crystallography: Automated, remote controlled protein-to-structure pipelines for drug design. In *ACTA CRYSTALLOGRAPHICA A-FOUNDATION AND ADVANCES*, volume 77, pages C144–C144. INT UNION CRYSTALLOGRAPHY 2 ABBEY SQ, CHESTER, CH1 2HU, ENGLAND, 2021.
- [327] Raphael Ponsard, Nicolas Janvier, Jerome Kieffer, Dominique Houzet, and Vincent Fristot. Rdma data transfer and gpu acceleration methods for high-throughput online processing of serial crystallography images. *Journal of Synchrotron Radiation*, 27(5):1297–1306, 2020.

- [328] Eike C Schulz, Briony A Yorke, Arwen R Pearson, and Pedram Mehrabi. Best practices for time-resolved serial synchrotron crystallography. *Acta Crystallographica Section D: Structural Biology*, 78(1):14–29, 2022.
- [329] Sébastien Boutet, Aina E Cohen, and Soichi Wakatsuki. The new macromolecular femtosecond crystallography (mfx) instrument at lcls. *Synchrotron radiation news*, 29(1):23–28, 2016.
- [330] Lukas Lomb, Thomas RM Barends, Stephan Kassemeyer, Andrew Aquila, Sascha W Epp, Benjamin Erk, Lutz Foucar, Robert Hartmann, Benedikt Rudek, Daniel Rolles, et al. Radiation damage in protein serial femtosecond crystallography using an x-ray free-electron laser. *Physical Review B*, 84(21):214111, 2011.

---

# Acknowledgements

First, I must thank my research supervisors, Henry N. Chapman and Oleksandr Yefanov. This thesis would have never been accomplished without their assistance and dedicated involvement in every step of the process. I want to thank you very much for your support, encouragement, patience, and understanding over these past three years.

Working with all the former and current members of the Coherent Imaging Division at CFEL has been a great pleasure. In particular, I would like to express my gratitude and appreciation for Anton Barty for the fruitful discussions, support, and great sense of humour that diversified lunch breaks. I would also like to acknowledge Luca Gelisio for allowing me to work with his team and present my work in internal group meetings. Furthermore, I value Valerio Mariani for his interest in the work regarding data reduction and for giving me a chance to give a talk in front of his wonderful team at SLAC, where I received good feedback.

Getting through my dissertation required more than academic support, and I have many, many people to thank for listening to me and, at times, having to tolerate me over the past three years. I cannot begin to express my gratitude and appreciation for their friendship. Aleksandra Tolstikova, Tang Li, Janina Sprenger, Aida Rahmani Mashhour, Chufeng Li and Margarita Zakharova have been unwavering in their personal and professional support during my PhD. I must thank everyone above Yaroslav Gevorkov, Alessandra Henkel and Aram Kalaydzhyan for many memorable evenings out, exciting experiments and trips. I thank Dominik Oberthür for his guidance throughout my research, numerous helpful discussions and help with structure refinements. I would also like to thank Ellen Petersen and Irmtraud Kleine, who opened their hearts to me when I arrived in the city.

My sincere thanks also go to Johanna Hakanpää, Alke Meents, Sebastian Günther and Patrick Y. A. Reinke, who provided me with an opportunity to work with them and who gave me access to their beamlines where I could test my ideas and developed software. Without their precious support, conducting this research would not be possible.

Most importantly, none of this could have happened without my family and my partner. Despite my limited devotion to small talk, my parents encouraged me through phone calls and chats every week. With his own brand of humour, my brother has been kind and supportive of me over the last several years. To my family and Alberto Prades Ibáñez – it would be an understatement to say that, as a team, we have experienced some ups and downs in the past three years. Every time I was ready to quit, you did not let me, and I am forever grateful. This dissertation stands as a testament to your unconditional love and encouragement.



---

# **Eidesstattliche Versicherung/Declaration on oath**

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Die Dissertation wurde in der vorgelegten oder einer ähnlichen Form nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

---

Hamburg, den 11.12.2023

---

Galchenkova Marina