Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Development of diffraction analysis methods for serial crystallography

Dissertation

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

vorgelegt von

Aleksandra Tolstikova

Hamburg
2020

Author e-mail: aotolstikova@gmail.com

Gutachter der Dissertation:	Prof. Dr. Henry N. Chapman Dr. Thomas A. White
Zusammensetzung der Prüfungskommission:	Prof. Dr. Henry N. Chapman Dr. Thomas A. White Prof. Dr. Daniela Pfannkuche Prof. Dr. Arwen Pearson Prof. Dr. Adrian Mancuso
Vorsitzende/r der Prüfungskommission:	Prof. Dr. Daniela Pfannkuche
Datum der Disputation:	11.05.2020
Vorsitzender Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Günter Hans Walter Sigl
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Wolfgang Hansen
Dekan der Fakultät MIN:	Prof. Dr. Heinrich Graener

Abstract

Serial crystallography, initially developed for use at X-ray free-electron lasers, has opened new opportunities to investigate structure and dynamics of biomolecules at physiologically relevant temperatures. It has since spread out to 3rd generation synchrotron sources where it allows us to measure protein microcrystals at room temperature, perform time-resolved experiments on biological crystals and obtain structures of radiation-sensitive proteins. Lately, extending the method of serial synchrotron crystallography to polychromatic X-ray beams has become of particular interest. Polychromatic beams provide two orders of magnitude higher photon flux, allowing significantly reduced exposure times compared to synchrotron experiments with monochromatic X-rays. This, in turn, allows accessing much shorter timescales in time-resolved diffraction experiments at synchrotrons.

Serial crystallography is based on merging data from still diffraction patterns collected from small randomly-oriented crystals only once exposed by X-rays, which differs significantly from conventional crystallography where one crystal is measured in different orientations while being rotated in the X-ray beam. Therefore, serial crystallography requires specific data analysis techniques capable of assembling a complete three-dimensional dataset of structure factor moduli from large numbers of individual still diffraction patterns. Analysis of serial crystallographic data has proven to be a complex problem, and despite the huge progress made in the field in the last decade, there is still a lot of room for improvement.

The aim of this dissertation is the development of new approaches to the processing and analysis of serial crystallographic data. Several experiments at both FELs and synchrotrons are presented to illustrate different analysis techniques. In particular, the major topic of the dissertation is extending the existing analysis software to serial crystallography with polychromatic beams. Following the first proof-of-principle study with the bandwidth of 2.5% of a 15 keV X-ray beam, a full data analysis pipeline for pink-beam serial crystallography is developed. The pipeline is then applied to three different datasets collected with the full undulator bandwidth of 5%, which demonstrates its feasibility even for the particularly difficult cases. The advantages of the analysis pipeline include the possibility of automated processing of large amounts of data, and analysis of polychromatic diffraction data from small crystals below 10 micron in size. This opens up new possibilities for time-resolved studies of irreversible biological reactions at sub-nanosecond timescales.

Zusammenfassung

Serielle Kristallographie, ursprünglich für den Einsatz an Freie-Elektronen-Röntgenlasern entwickelt, hat neue Möglichkeiten eröffnet, Strukturen und Dynamik von Biomolekülen zu untersuchen - und zwar bei physiologisch relevanten Temperaturen. Inzwischen findet sie auch in Synchrotronen der dritten Generation Anwendung, wo sie erlaubt, Protein-Mikrokristalle bei Raumtemperatur zu untersuchen, zeitaufgelöste Experimente an biologischen Kristallen durchzuführen und Strukturen strahlungsempfindlicher Proteine zu erlangen. In letzter Zeit ist die Erweiterung der Methode der seriellen Synchrotron-Kristallographie auf einen polychromatischen Röntgenstrahl ins besondere Interesse gerückt. Ein polychromatischer Strahl liefert einen um zwei Größenordnungen höheren Photonenfluss und ermöglicht damit eine signifikante Reduzierung der Belichtungszeiten im Vergleich zu Synchrotron Experimenten mit monochromatischer Röntgenstrahlung. Das wiederum ermöglicht viel kürzere Zeitskalen bei zeitaufgelösten Beugungsexperimenten an Synchrotronen.

Die serielle Kristallographie basiert auf der Zusammenführung von Daten aus Beugungsmustern von unbewegten, zufällig orientierten Kristallen, die nur einmalig Röntgenstrahlen ausgesetzt wurden. Das unterscheidet sich signifikant von der konventionellen Kristallographie, bei der ein Kristall in verschiedenen Ausrichtungen gemessen wird, während er im Röntgenstrahl rotiert wird. Daher erfordert die serielle Kristallographie eine spezifische Datenanalysetechnik, die in der Lage ist, einen vollständigen dreidimensionalen Datensatz von Strukturfaktoren aus einer großen Menge einzelner unbewegter Beugungsmuster zusammenzustellen. Die Analyse von Daten aus der seriellen Kristallographie hat sich als ein komplexes Problem erwiesen. Trotz der enormen Fortschritte, die in diesem Bereich im letzten Jahrzehnt erzielt wurden, gibt es noch viel Raum für Verbesserungen.

Das Ziel dieser Dissertation ist die Entwicklung neuer Ansätze zur Verarbeitung und Analyse von Daten aus der seriellen Kristallographie. Mehrere Experimente sowohl an FELs als auch an Synchrotronen werden vorgestellt, um verschiedene Analysetechniken zu veranschaulichen. Insbesondere ist das Hauptthema dieser Dissertation die Erweiterung der bestehenden Analysesoftware auf die serielle Kristallographie mit einem polychromatischen Strahl. Nach der ersten Machbarkeitsstudie mit 2.5% Bandbreite und einem 15 keV Röntgenstrahl, wird eine vollständige Datenanalyse-Pipeline für die serielle Pink-Beam-Kristallographie entwickelt. Danach wird die Pipeline auf drei weitere Datensätze angewendet, die mit der vollständigen Undulator Bandbreite von 5% aufgenommen wurden, was die Machbarkeit selbst für die besonders schwierigen Fälle belegt. Die Vorteile der entwickelten Analyse-Pipeline einschließlich der Möglichkeit der automatischen Verarbeitung großer Datenmengen und Analyse von polychromatischen Beugungsdaten aus kleinen Kristallen unter 10 Mikrometer Größe eröffnen neue Möglichkeiten für zeitaufgelöste Studien von irreversiblen biologischen Reaktionen im Sub-Nanosekundenbereich an Synchrotronen.

Contents

CHAPTER 1

Motivation

A vast majority of functions within living organisms are performed by proteins. Proteins are large macromolecules consisting of one or more linear chains of amino acid residues. The sequence of amino acids in a protein is defined in the genetic code but the function of a protein and its ability to interact with other molecules is determined by its three-dimensional structure produced in the process of protein folding. Knowing the three-dimensional structure of a protein is essential to understanding its function and properties.

Three main experimental techniques used to determine protein structure include crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy. Crystallography is responsible for the overwhelming majority, almost 90%, of all experimentally determined protein structures to date [1]. Crystallography relies on the process of diffraction by a crystal - a constructive interference between the waves scattered from atoms in a crystal in the directions defined by the Bragg's law [2]. Diffraction experiments can be performed with X-rays, electrons or neutrons. For a variety of reasons, including high availability, high throughput and usually higher achievable resolution, X-rays are by far the most widely used type of radiation in crystallography.

To perform crystallography experiment the protein first has to be crystallized, i.e. protein molecules must be organized into a three-dimensional lattice. The crystal is then put into the X-ray beam with the wavelength of around 1 Å, which is slightly smaller than the typical interatomic distance in a protein. The diffraction recorded from the crystal rotated in the X-ray beam is then used to determine crystal structure to near-atomic resolution. With the use of powerful third-generation synchrotrons and development of highly reliable diffraction analysis and crystal structure determination software, macromolecular crystallography became a routine technique in the last 20-30 years. However, there are two major limitations of this technique. First, it requires the protein to form crystals of sufficient size and quality: it may take years of research to obtain suitable crystals of the protein under investigation. Second, crystallographic data collection is significantly complicated by the radiation damage induced on the sample by the X-rays. Cooling the crystal down to cryogenic temperatures partially solves the problem as it increases the tolerable dose limit by two orders of magnitude. However, even cryo-cooled crystals often do not survive long enough in the X-ray beam at a modern synchrotron and valuable high-resolution diffraction data gets lost due to radiation damage.

A new experimental technique called serial crystallography was introduced 10 years ago when the first hard X-ray free electron laser (FEL) started operation. FELs produce X-ray beam with unique properties: pulses of only several tens of femtoseconds which are more than 10 orders of magnitudes brighter than

the most powerful synchrotrons (Eqn. 3.2). When a crystal is put into the FEL beam it gets destroyed by a single pulse, but since the pulse is so short diffraction occurs before the radiation damage affects the crystal. In the method of serial crystallography, diffraction signal is collected from thousands of protein crystals in random orientations and merged together to produce a full diffraction data set. In addition to effectively overcoming radiation damage, serial crystallography opens several new opportunities. Due to the extreme brightness of the FEL beam, it allows collecting data from very small protein crystals below micrometer in size. As it doesn't require cryo-cooling of the crystals it provides valuable information about protein structure at room temperature, i.e. in biologically relevant conditions. Finally, it allows to study protein dynamics in time-resolved fashion. Irreversible biological processes can be triggered externally and probed with sub-picosecond time resolution using extremely short pulses of an FEL.

While offering all these advantages, serial crystallography posed several new challenges. First, the main experimental challenge is sample delivery: a fresh crystal has to be delivered into the X-ray beam for each X-ray pulse. Several sample delivery methods have been developed for serial crystallography trying to minimize data collection time and the total amount of sample needed to solve the structure while keeping up with the repetition rate of FELs. The second major challenge is data analysis. Serial crystallographic data is substantially different from conventional crystallography and requires development of novel analysis methods.

The main focus of this thesis is the analysis of serial crystallographic data. In particular, it is meant to extend available analysis methods to serial crystallography at synchrotrons using polychromatic X-ray beams. Using polychromatic X-rays should not only reduce data collection time and amount of sample required to solve crystal structure using serial synchrotron crystallography and make it viable compared to the conventional data collection, but also allow to perform time-resolved experiments at synchrotrons with sub-nanosecond resolution. This would make time-resolved serial crystallography at synchrotrons an alternative to FELs and make it more accessible to broader community.

Chapters 2 and 3 of this thesis give a general introduction to crystallography covering theoretical background and experimental aspects of X-ray crystallography, respectively. Chapter 4 gives an overview of the main data analysis methods in serial crystallography. Chapter 5 describes the analysis of two serial crystallography experiments at the FEL highlighting specifically implemented data processing steps. Chapter 6 introduces pink-beam serial crystallography and presents the first experiment and data analysis pipeline for serial crystallography with 2.5% X-ray bandwidth. In Chapter 7 the pipeline is extended to the full undulator bandwidth and applied to three different pink-beam datasets to investigate how the resulting data quality depends on the number of diffraction patterns and strength of the diffraction data. Chapter 8 summarizes the results of the thesis and gives an outlook on future research.

CHAPTER 2

Introduction to X-ray crystallography

2.1 Scattering of X-rays

There are three major processes happening with some probability when X-rays pass through matter. The first is absorption, when a photon loses all its energy to eject an electron from an atom (photoelectric effect). The ionized atom then emits a photon (fluorescence) or an Auger electron. The second is coherent or Thomson scattering, when photons change their direction but conserve their energy. The last is incoherent or Compton scattering, when the photon uses part of its energy to ionize an electron and scatters in a different direction.

Let us consider coherent X-ray scattering. As described by classical electrodynamics, a free charged particle placed in the periodic electromagnetic field of the incident wave undergoes oscillatory motion with the same frequency as electric field and becomes itself a source of electromagnetic radiation of the same frequency. The intensity of resulting radiation is

$$I_{Th} = I_0\Big(\frac{e^2}{mrc^2}\Big)^2(sin^2\mu + cos^2\mu\ cos^2 2\Theta) \tag{2.1}$$

where I_0 is the intensity of incident wave, e and m are charge and mass of the particle respectively, r is the distance from the particle, μ is the angle between the polarization direction of the incident wave and the scattering plane and 2Θ is the angle between the scattered and incident beam (scattering angle).

If the incident beam is non-polarized Eqn. 2.1 becomes

$$I_{Th} = I_0\Big(\frac{e^2}{mrc^2}\Big)^2\Big(\frac{1 + cos^2 2\Theta}{2}\Big) \tag{2.2}$$

where $P = \frac{1+cos^2 2\Theta}{2}$ is called polarization factor and describes the dependence of the scattered intensity on the scattering angle.

Since neutrons do not have electric charge and therefore don't contribute to coherent X-ray scattering and protons are about 1837 times heavier than electrons which makes their contribution negligible, from now on when talking about X-ray scattering we consider only scattering on electrons.

The scattering is coherent because there is a defined phase relation between incident and scattered beam: $\Delta\phi = \pi$ for electrons, therefore the scattered waves will interfere. If a plane wave with the wavelength λ going in the direction $\boldsymbol{s_0}$ scatters from two scattering centers O and O' (Fig. 2.1) the phase difference between the wave scattered from the point O' in position $\boldsymbol{r}$ and O in position $\boldsymbol{r} = 0$ in the

direction $\boldsymbol{s}$ is equal to $\frac{2\pi}{\lambda}(\boldsymbol{s}-\boldsymbol{s_0})\cdot\boldsymbol{r}$. Therefore, the wave scattered from O' is described by

$$f\exp\frac{2\pi i}{\lambda}(\boldsymbol{s}-\boldsymbol{s_0})\cdot\boldsymbol{r}=f\exp 2\pi(\boldsymbol{k}-\boldsymbol{k_0})\cdot\boldsymbol{r}=f\exp 2\pi i\boldsymbol{r^*}\cdot\boldsymbol{r} \tag{2.3}$$

where $\boldsymbol{k_0}=\frac{\boldsymbol{s_0}}{\lambda}$ and $\boldsymbol{k}=\frac{\boldsymbol{s}}{\lambda}$ are the wave vectors of the incident and scattered waves respectively, and $\boldsymbol{r^*}=\boldsymbol{k}-\boldsymbol{k_0}$ (Fig. 2.1).

Expanding this formula to an object consisting of n scatterers with a scattering amplitude f_j we get

$$F(\boldsymbol{r^*})=\sum_{j=1}^{n}f_j\exp 2\pi i\boldsymbol{r^*}\cdot\boldsymbol{r}. \tag{2.4}$$

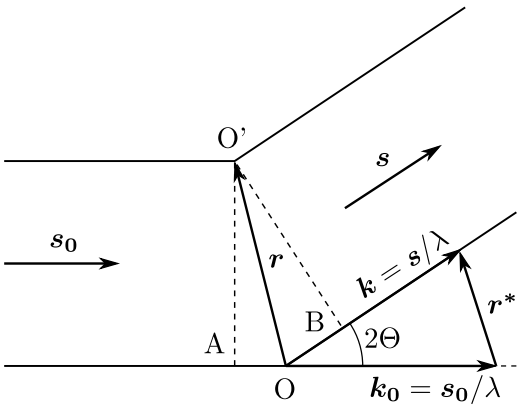


Figure 2.1: The path difference between the waves scattered from O and O' is δx = AO + OB = $-\boldsymbol{s}\cdot\boldsymbol{r}+\boldsymbol{s_0}\cdot\boldsymbol{r}$. The incident and scattered waves have the same wavelength λ $\Rightarrow$ the phase difference is $\delta\phi=\frac{2\pi}{\lambda}\delta x=2\pi(\boldsymbol{k}-\boldsymbol{k_0})\cdot\boldsymbol{r}$.

When X-rays interact with an object the scatterers are electrons. Their scattering amplitude is derived from Thomson formula (Eqn. 2.2) as $f_e=f_0\sqrt{I_{Th}/I_0}$, where f_0 is the amplitude of the incident wave. It is convenient to omit f_e from the subsequent calculations and use so called scattering factor which is defined as scattering amplitude of an object divided by f_e.

Then, in the case of an object with a continuous electron density $\rho(\boldsymbol{r})$ the sum from Eqn. 2.4 should be replaced with an integral and the scattering factor becomes

$$f(\boldsymbol{r^*})=\int_V\rho(\boldsymbol{r})\exp 2\pi i\boldsymbol{r^*}\cdot\boldsymbol{r}d\boldsymbol{r}=\mathcal{F}(\rho(\boldsymbol{r})) \tag{2.5}$$

where $\mathcal{F}(\boldsymbol{r})$ represents Fourier transform operator. The space of $\boldsymbol{r^*}$ vectors is called reciprocal space.

2.1.1 Scattering by atoms and molecules

When X-rays pass through an object they interact with atomic electrons which occupy different energy states. If the electron conserves its energy state after interaction with an X-ray then the scattering is elastic. The scattering factor of an atom with electron density $\rho(\boldsymbol{r})$ is defined by Eqn. 2.5.

The wave scattered from a molecule is described as a sum of waves scattered by each atom. If a molecule consists of n atoms at positions r_j its scattering factor is

$$f_M(\boldsymbol{r^*}) = \sum_{j=1}^{n} f_{aj} \exp 2\pi i \boldsymbol{r^*} \cdot \boldsymbol{r_j} \tag{2.6}$$

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

2.1.2 Temperature factor

Atoms in a crystal oscillate around their mean positions due to thermal energy. Since the timescale of the scattering experiment is much longer than the period of thermal motion, the electron density of an atom, which defines the scattering, is the average over time electron density.

In case of spherically symmetric oscillations the probability of an atom to be found at the position $\boldsymbol{r}$ is described by a gaussian with a mean shift of an atom $\sqrt{\overline{u^2}}$:

$$w(\boldsymbol{r}) = \frac{1}{(2\pi\overline{u^2})^{3/2}} \exp(-\boldsymbol{r}^2/2\overline{u^2}). \tag{2.7}$$

The electron density distribution of such atom is equal to

$$\rho_{aT} = \int \rho(\boldsymbol{r} - \boldsymbol{r'}) w(\boldsymbol{r}) d\boldsymbol{r'} = \rho(\boldsymbol{r}) * w(\boldsymbol{r}), \tag{2.8}$$

and the scattering factor becomes

$$f_{aT}(\boldsymbol{r^*}) = \mathcal{F}(\rho(\boldsymbol{r}) * w(\boldsymbol{r})) = \mathcal{F}(\rho(\boldsymbol{r}))\mathcal{F}(w(\boldsymbol{r})) = f_a(\boldsymbol{r^*}) \exp(-B_{iso} r^{*2}/4) \tag{2.9}$$

where $B_{iso} = 8\pi\overline{u^2}$ is usually referred to as Debye-Waller factor.

In the general case atomic thermal motion in the crystal lattice is anisotropic and the thermal factor in this case is represented by an ellipsoid centered on each atom.

2.1.3 Diffraction by a crystal

A crystal is a solid where atoms form a periodic arrangement, which is described by a crystal lattice. The crystal lattice is defined as a set of lattice points each described by the following equation:

$$\boldsymbol{r}_{u,v,w} = u\boldsymbol{a} + v\boldsymbol{b} + w\boldsymbol{c} \tag{2.10}$$

where u, v, w are integers and $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ are noncoplanar vectors called basis vectors. The crystal is formed by a group of atoms, called the unit cell, repeated at each lattice point. The unit cell has a shape of a parallelepiped spanned by the basis vectors $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$. The length of the unit cell edges (a, b, c) and the angles between them (α, β, γ) are referred to as unit cell or lattice parameters.

An infinite lattice can be represented by the following function:

$$L(\boldsymbol{r}) = \sum_{u,v,w=-\infty}^{+\infty} \delta(\boldsymbol{r} - \boldsymbol{r}_{u,v,w}) \tag{2.11}$$

where $\delta(\boldsymbol{r})$ is the Dirac delta function. If $\rho_M(\boldsymbol{r})$ describes electron density in the unit cell then electron density of the infinite crystal is a convolution of $L(\boldsymbol{r})$ with $\rho_M(\boldsymbol{r})$:

$$\rho_\infty(\boldsymbol{r}) = \rho_M(\boldsymbol{r}) * L(\boldsymbol{r}). \tag{2.12}$$

The scattering factor of a crystal then can be calculated from Eqn. 2.5 as

$$\begin{aligned} f_\infty(\boldsymbol{r}) &= \mathcal{F}(\rho_M(\boldsymbol{r}) * L(\boldsymbol{r})) = \mathcal{F}(\rho_M(\boldsymbol{r}))\mathcal{F}(L(\boldsymbol{r})) \\ &= f_M(\boldsymbol{r}^*)\frac{1}{V}\sum_{h,k,l=-\infty}^{+\infty}\delta(\boldsymbol{r}^* - \boldsymbol{r}^*_{h,k,l}) \\ &= \frac{1}{V}\sum_{h,k,l=-\infty}^{+\infty} F_{hkl}\delta(\boldsymbol{r}^* - \boldsymbol{H}_{hkl}) \end{aligned} \tag{2.13}$$

where V is a unit cell volume, $\boldsymbol{H}_{hkl}$ is a reciprocal lattice vector defined as

$$\boldsymbol{H}_{hkl} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*,\ \boldsymbol{a}^* = \frac{\boldsymbol{b}\times\boldsymbol{c}}{V},\ \boldsymbol{b}^* = \frac{\boldsymbol{c}\times\boldsymbol{a}}{V},\ \boldsymbol{c}^* = \frac{\boldsymbol{a}\times\boldsymbol{b}}{V} \tag{2.14}$$

where $\boldsymbol{a}^*$, $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$ are reciprocal cell vectors which constitute reciprocal lattice. As can be seen from Eqn. 2.13, the scattering factor of an infinite crystal can be non-zero only when $\boldsymbol{r}^*$ coincides with a reciprocal lattice point. $F_{hkl} = f_M(\boldsymbol{H}_{hkl})$, a scattering factor of the unit cell in the reciprocal space point $\boldsymbol{r}^* = \boldsymbol{H}_{hkl}$, is referred to as structure factor.

To account for the finite size of the crystal the form function of the crystal must be introduced:

$$\Phi(\boldsymbol{r}) = 1 \text{ inside crystal}, 0 \text{ outside crystal} \tag{2.15}$$

Then the electron density of the crystal Eqn. 2.12 is multiplied by $\Phi(\boldsymbol{r})$ and relation from Eqn. 2.13 becomes

$$f_{cr}(\boldsymbol{r}^*) = \frac{1}{V}\sum_{h,k,l=-\infty}^{+\infty} F_{hkl}D(\boldsymbol{r}^* - \boldsymbol{H}_{hkl}) \tag{2.16}$$

where $D(\boldsymbol{r}^*) = \mathcal{F}(\Phi(\boldsymbol{r})) = \int_V \exp 2\pi i\boldsymbol{r}^*\boldsymbol{r}d\boldsymbol{r}$.

As a result, the delta function in the Eqn. 2.13, corresponding to each reciprocal lattice point, in case of a finite crystal becomes a distribution function $D(\boldsymbol{r}^*)$ identical for all reciprocal lattice points. In the most simple case when the crystal is a parallelepiped with dimensions A_1, A_2 and A_3 the distribution function can be calculated as

$$\begin{aligned} D(\boldsymbol{r}^*) &= \int_{-A_1/2}^{A_1/2}\int_{-A_2/2}^{A_s/2}\int_{-A_3/2}^{A_3/2}\exp[2\pi i(xx^* + yy^* + zz^*)]dxdydz \\ &= \frac{\sin\pi A_1x^*}{\pi x^*}\frac{\sin\pi A_2y^*}{\pi y^*}\frac{\sin\pi A_3z^*}{\pi z^*}. \end{aligned} \tag{2.17}$$

Therefore, each reciprocal lattice node in a diffraction experiment is in fact a limited spatial domain which dimensions in the reciprocal space are equal to $2A_i^{-1}$. As a result, diffracted beams have a limited angular size proportional to A_i^{-1} and the bigger the crystal is the narrower diffraction maxima become. Furthermore, between the principal maxima there are secondary maxima, called fringes, spaced at intervals of A_i^{-1}. This effect is illustrated in Fig. 2.2, showing the scattering factor amplitude of a two-dimensional rectangular crystal (Fig. 2.2a) and diffraction pattern of a photosystem I nanocrystal, demonstrating shape transform fringes recorded experimentally (Fig. 2.2b) [3].

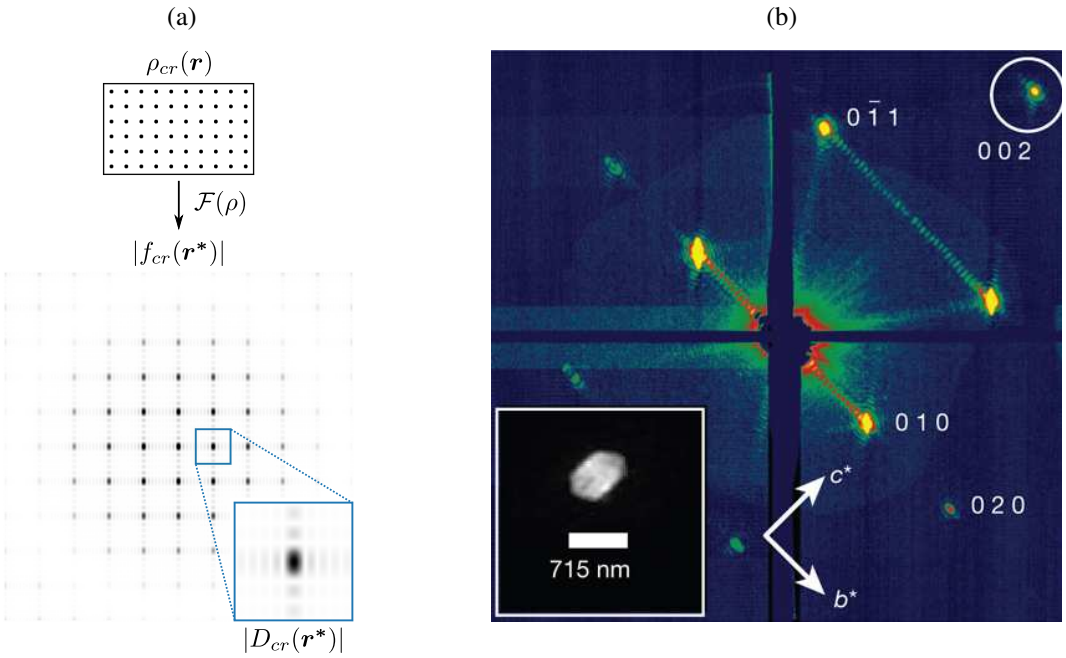


Figure 2.2: (a) Illustration of the scattering factor amplitude $|f_{cr}(\boldsymbol{r}^*)|$ of a two-dimensional crystal consisting of 10×6 unit cells. Inset shows the shape of the reciprocal lattice node produced by the rectangular-shaped crystal. (b) Low-angle diffraction patterns revealing coherent diffraction from the structure of the photosystem I nanocrystals, shown using a logarithmic, false-colour scale. Inset shows a real-space image of the nanocrystal, determined from the shape of the circled diffraction peak. Figure from Chapman *et al.*, 2011 [3].

2.1.4 Bragg's law and Ewald construction

The amplitude of the wave scattered in the direction $\boldsymbol{k} = \boldsymbol{k_0} + \boldsymbol{r}^*$ is proportional to the scattering factor $f_{cr}(\boldsymbol{r}^*)$. As shown above, the maxima of the scattering factor correspond to the reciprocal lattice points $\boldsymbol{r}^* = \boldsymbol{H}_{hkl}$. Thus, most of the X-rays scattered from a crystal will scatter in the directions defined by the reciprocal lattice, called diffraction directions. The diffraction conditions can be then defined as

$$\boldsymbol{k} - \boldsymbol{k_0} = \boldsymbol{H}_{hkl}, \qquad |\boldsymbol{k}| = |\boldsymbol{k_0}| = \frac{1}{\lambda} \tag{2.18}$$

where λ is the wavelength of the radiation.

A simpler method to explain diffraction by a crystal was described by W. L. Bragg in 1912 [2]. If we consider crystal as an array of parallel lattice planes and diffraction as a positive interference between X-rays reflected from these planes (Fig. 2.3) then, to satisfy the diffraction condition at angle 2Θ, the difference path between the X-rays reflected from two neighboring lattice planes should be multiple of λ:

$$2d_{hkl} \sin\theta = n\lambda \tag{2.19}$$

where d_{hkl} is the interplanar distance and n is a positive integer. This equation is known as Bragg's law.

Let us now consider diffraction condition as described by Eqn. 2.18. Since the incident and diffracted X-rays have the same wavelength λ, the ends of the incident and diffracted wave vectors $\boldsymbol{k}$ and $\boldsymbol{k_0}$ lie on the sphere of the radius $1/\lambda$. This geometrical construction is called Ewald sphere. If the origin of the reciprocal space is placed at the end of the vector $\boldsymbol{k}_0$, then the diffraction condition from Eqn. 2.18 is satisfied if a reciprocal lattice point $\boldsymbol{H}_{hkl}$ intersects the Ewald sphere (Fig. 2.4).

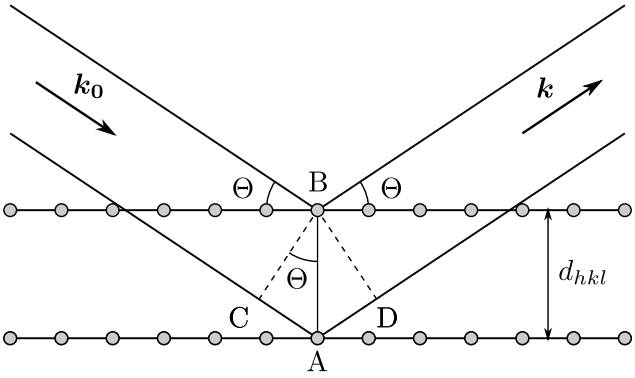


Figure 2.3: Reflection of X-rays from two lattice planes separated by the interplanar distance d_{hkl}. The path difference between waves reflected by two planes is AC + AD = $2d_{hkl} \sin\theta$

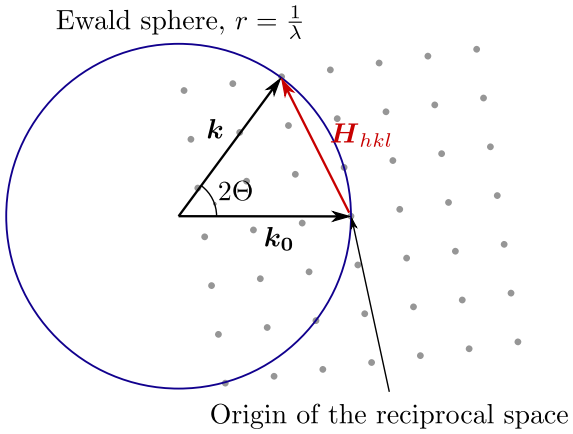


Figure 2.4: Ewald sphere construction: diffraction condition $\boldsymbol{k} - \boldsymbol{k_0} = \boldsymbol{H}_{hkl}$ is satisfied when reciprocal lattice point $\boldsymbol{H}_{hkl}$ lies on the Ewald sphere.

From the Ewald construction it is immediately obvious that equations 2.18 and 2.19 are identical: as can be seen from the definition of the reciprocal lattice, the length of the reciprocal lattice vector $\boldsymbol{H}_{hkl}$ is a multiple of the inverse interplanar distance d_{hkl} and equals to $n/d_{hkl} = 2/\lambda \sin\theta$.

2.1.5 Diffraction intensities and the phase problem

Using Eqn. 2.6 for electron density of the unit cell we can express F_{hkl} as following:

$$F_{hkl} = \sum_{j=1}^{n} f_{aTj} \exp(2\pi i \boldsymbol{r_j} \boldsymbol{H}_{hkl}) \tag{2.20}$$

where n is the number of atoms in the unit cell and $\boldsymbol{r}_j$ and f_{aTj} are the position and scattering factor of j^{th} atom.

In the real diffraction experiment when diffracted X-rays are captured by a detector, the information recorded is the averaged over time scattering intensity at a certain scattering angle. The intensity is proportional to the square of the structure factor modulus:

$$I_{hkl} \sim |F_{hkl}|^2. \tag{2.21}$$

Since $\langle|\exp 2\pi i(\boldsymbol{r}\boldsymbol{H}_{hkl})|^2\rangle = 1$ and atomic scattering factors f_{aTj} monotonically decrease with increasing scattering angle, the average intensity falloff with $\sin\Theta/\lambda$ can be expressed as

$$\langle I(\sin\Theta/\lambda)\rangle = \sum_{j=1}^{n} f_{aTj}^2(\sin\Theta/\lambda). \tag{2.22}$$

This expression is used to estimate the average temperature B-factor. Taking into account that $f_{aTj}^2(\sin\Theta/\lambda) \sim \sum_{j=1}^{n} f_{aj}^2 \exp[-2B_{iso}(\sin\Theta/\lambda)^2]$, dividing both sides of Eqn. 2.22 by $\sum = \sum_{j=1}^{n} f_{aj}^2$ and taking a logarithm we obtain

$$\ln(\langle I\rangle/\textstyle\sum) = \text{const} - 2B_{iso}(\sin\Theta/\lambda)^2. \tag{2.23}$$

Therefore the average B-factor can be directly obtained from the slope of a plot of $\ln(\langle I\rangle/\sum)$ vs $(\sin\Theta/\lambda)^2$ which is commonly used in crystallography and referred to as Wilson plot.

Because of the falloff of the diffraction intensities with increasing of the scattering angle, there is a limit of $\sin\Theta/\lambda$ above which the intensities cannot be measured in a given experiment. The value of d_{min}, defined as $1/2d_{min} = (\sin\Theta/\lambda)_{max}$, is used as a measure of the resolution of the diffraction experiment.

According to Eqn. 2.5, the scattering amplitude of an object is a Fourier transform of its electron density. Thus, knowing both moduli and phases of structure factors, electron density $\rho(\boldsymbol{r})$ in the unit cell can be derived as the inverse Fourier transform:

$$\rho(\boldsymbol{r}) = \mathcal{F}^{-1}(f(\boldsymbol{r}^*)) = \frac{1}{V}\sum_{h,k,l=-\infty}^{+\infty} F_{hkl}\exp 2\pi i(\boldsymbol{r}\boldsymbol{H}_{hkl}). \tag{2.24}$$

Unfortunately, this operation cannot be applied directly to the experimental data because phases of the scattered X-rays are lost in the diffraction experiment and only their intensities or structure factor moduli are measured. This is known as the *phase problem* in crystallography and there are a few methods which can be applied to solve it.

2.1.6 The Patterson function

It is impossible to obtain electron density in the unit cell directly from the structure factor moduli without obtaining the phases, it is possible however to gain useful information by applying inverse Fourier transform to the measured intensities and setting phases to zero:

$$P(\boldsymbol{u}) = \mathcal{F}[|F(\boldsymbol{h})|^2] = \mathcal{F}[F(\boldsymbol{h})F(-\boldsymbol{h})] = \rho(\boldsymbol{r}) \times \rho(-\boldsymbol{r}) = \int_V \rho(\boldsymbol{r})\rho(\boldsymbol{r}+\boldsymbol{u})d\boldsymbol{r} \tag{2.25}$$

where $\boldsymbol{h} = \boldsymbol{H}_{hkl}$ is the reciprocal lattice vector.

$P(\boldsymbol{u})$ is called a Patterson function, it can be calculated directly from the experimental diffraction data and gives the autoconvolution of the electron density in the unit cell. $P(\boldsymbol{u})$ has large values when $\boldsymbol{u}$ is the interatomic distance. If there are N atoms in the unit cell, the Patterson map will have $N(N-1)$ peaks, not considering overlaps due to repeating interatomic vectors, with the intensities proportional to the numbers of electrons in the corresponding atoms.

2.1.7 Friedel's law

The electron density $\rho(\boldsymbol{r})$ is approximately real-valued function, therefore, as it follows from the definition of the Fourier transform

$$F(\boldsymbol{h}) = \mathcal{F}[\rho(\boldsymbol{r})] = \int_V \rho(\boldsymbol{r}) \exp i\boldsymbol{h}\boldsymbol{r} d\boldsymbol{r} \tag{2.26}$$

the structure factors of the centrosymmetric reflections $\boldsymbol{h}$ and $-\boldsymbol{h}$ are complex conjugates:

$$F(\boldsymbol{h}) = F^*(-\boldsymbol{h}). \tag{2.27}$$

That means that the squared amplitudes, or diffraction intensities, are centrosymmetric:

$$|F(\boldsymbol{h})|^2 = |F(-\boldsymbol{h})|^2. \tag{2.28}$$

This statement is known as Friedel's law, and reflection pairs hkl and $\bar{h}\bar{k}\bar{l}$ with the opposite indices are called Friedel pairs.

2.1.8 Anomalous scattering

So far we considered electrons in an atom as free electrons. In reality, the electrons are bound to the atoms, occupying atomic orbitals, and can be considered as oscillators with characteristic orbital frequency. In particular, if the frequency of the incident wave is close to the natural frequency, resonance will occur. The scattering in this case is called anomalous. The classical motion equation of the electron in the electric field E_0 of the incident wave with the frequency $\omega/2\pi$ can be expressed as

$$\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + \omega_0^2 x = \frac{eE_0}{m} \exp i\omega t \tag{2.29}$$

where ω_0 is the natural angular frequency of the electron and γ is the damping coefficient.

The solution of this equation is

$$x(t) = \frac{eE_0}{m} \frac{\exp i\omega t}{\omega_0^2 - \omega^2 + i\gamma\omega}. \tag{2.30}$$

The dipole moment of this oscillating electron is ex, and the electromagnetic wave produced by such oscillating dipole has the amplitude

$$E = \frac{e^2 E_0 P}{mc^2 r} \frac{\omega^2}{\omega_0^2 - \omega^2 + i\gamma\omega} \tag{2.31}$$

where P is the polarisation coefficient. Thus, compared to Thomson formula (Eqn. 2.2), the scattering amplitude of the electron gains a frequency-dependent factor

$$\frac{E}{E_{Th}} = \frac{\omega^2}{\omega_0^2 - \omega^2 + i\gamma\omega}. \tag{2.32}$$

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

$$f_a = f_0 + \Delta f' + if'' \tag{2.33}$$

where f_0 is the 'normal' scattering factor in the absence of anomalous scattering, $\Delta f'$ and f'' are called the real and imaginary dispersion corrections respectively.

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 due to imaginary term f'' Friedel's law is in general not valid in the presence of the anomalous scattering.

2.2 Crystal structure determination

In small-molecule crystallography, for crystal structures with typically less than 1000 atoms per unit cell, the phase problem is usually solved by so called direct methods. Direct methods obtain structure factor phases directly from experimental amplitudes, utilizing two important approximations for the electron density: it is everywhere positive (positivity) and it is composed of electron densities of discrete atoms (atomicity). Based on these two assumptions, statistical relationships between the sets of structure factors can be derived, which are then used to deduce the probable values for the phases. The statistical relationships become weaker as the number of atoms in the unit cell increases, and the atomicity assumption is only practically relevant when the resolution of the measured intensities is high enough to resolve individual atoms, i.e. $\lesssim 1.2$ Å. In protein crystallography, where the number of atoms in the unit cell is several thousand and the resolution rarely exceeds 1.5 Å, direct methods are generally not applicable except for some cases where they are used to find the positions of heavy atoms.

2.2.1 Experimental phasing in macromolecular crystallography

There are two main techniques used for *de novo* phasing in protein crystallography: isomorphous replacement and anomalous diffraction. Both of them comprise several variations, including single isomorphous replacement (SIR) or multiple isomorphous replacement (MIR) as well as single-wavelength anomalous diffraction (SAD) and multi-wavelength anomalous dispersion (MAD).

Historically, the first structures of biological macromolecules were solved by John C. Kendrew [4] and Max Perutz [5] using isomorphous replacement method, first developed by John M. Robertson for small-molecule crystallography [6]. This method relies on crystallization of the derivative - a target compound with one or more heavy atoms incorporated into its structure. It requires for the derivative and native crystals to be isomorphous, i.e. the structure of the molecules and the lattice should be identical except for the addition of the heavy atom. The diffraction intensities of the native crystal can be subtracted from the intensities of the derivative, and the locations of the incorporated heavy atoms can be determined using Patterson map (Section 2.1.6). The phases of the structure factors of the native and derivative crystals are then calculated using the phases of the heavy atom substructure. The difference between SIR and MIR is that in MIR more than one derivative is used which allows for the unambiguous phase determination.

Anomalous phasing is based on the effect of the anomalous scattering, explained in more detail in section 2.1.8. The majority of atoms composing a protein are light atoms, such as H, C, N and O, giving very low anomalous scattering. To obtain a measurable anomalous signal, heavier atoms, which are either naturally present in the protein or intentionally introduced to the structure, are used. From the differences in the measured intensities of Friedel pairs of reflections only (SAD) or in combination with the differences between the intensities of the same reflections at two different X-ray wavelengths (MAD), it is possible to obtain phases and determine the heavy atom substructure, which is then used to solve the target protein structure similar to isomorphous replacement method.

2.2.2 Molecular replacement

An alternative approach to solving the phase problem in macromolecular crystallography pioneered by Michael G. Rossmann [7] is the molecular replacement method (MR). It takes advantage of the similarities

in the structures of different proteins, by using the phases of a previously known structure as initial estimates for the phases of the new structure. The model and the target protein should have a certain degree of similarity which is usually evaluated based on their sequence identity. The model structure is then put into the target unit cell and probed in all possible positions and orientations for the best fit between the predicted and experimentally observed intensities. Once the best match is found, the phases of the model structure together with the measured structure factor moduli are used to calculate the initial electron density map and the model of the target protein can be built.

In some cases proteins with low sequence identity exhibit high degree of structural similarity. For example, G protein-coupled receptors (GPCRs) - a largest family of membrane proteins in eukaryotes which are responsible for most of cell responses to outside signals - share several structural features regardless of their sequence identity. All GPCRs contain seven transmembrane helices and additional loops both inside and outside the cell (Fig. 2.5). While the structures of intracellular and extracellular loops vary widely between different GPCRs, the fold in transmembrane region is very similar for all proteins in the GPCR family, which often leads to relatively straightforward phasing using MR [8].

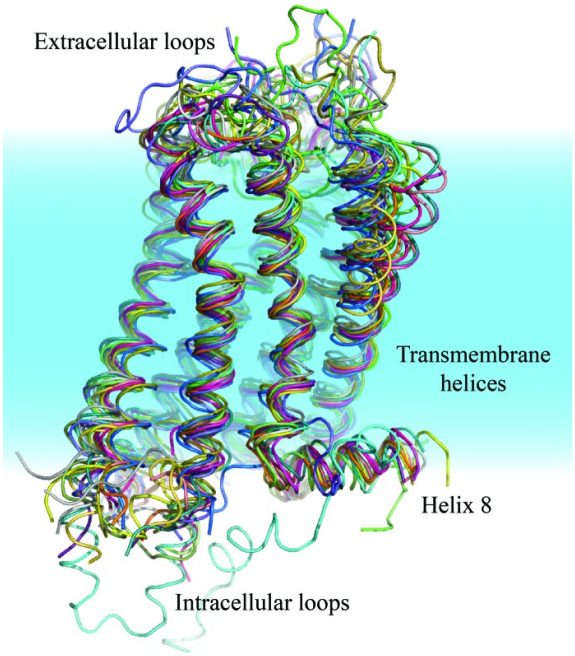


Figure 2.5: Alignment of seven different GPCR structures. Structures show exceptional fold conservation, even with pairwise sequence identities that are lower than 25% in some cases. Figure from Kruse *et al.* 2013 [8].

As the number of known structures increases the use of MR becomes more common. It is the most widely used phasing method now accounting for up to 70% of all structure deposited in the Protein Data Bank (PDB) [9].

2.2.3 Structure refinement and validation

After the initial phases are obtained the structure has to be refined. The crystallographic data in macromolecular crystallography is underdetermined: the number of unknowns, which include three positional arguments, isotropic B-factor and sometimes occupancy for each atom, is usually on the order of or often

even less than the number of the experimentally measured intensities. Therefore, the chemical restraints have to be applied to the model during refinement, using prior knowledge of chemically likely bond lengths and bond angles between atoms.

The goal of the refinement is to find a model which describes the experimental data in the best possible way. A figure of merit, used to evaluate the quality of the model by comparing the structure factors calculated from the model to the observed structure factors, is the R-factor:

$$R = \frac{\sum_{hkl} ||F_{obs}| - |F_{calc}||}{\sum_{hkl} |F_{obs}|}. \tag{2.34}$$

The refinement is aimed to improve the model by reducing the R-factor. Since the model is used together with the experimental data to calculate electron density, there is a danger of over-fitting the model if the R-factor is used to assess its quality. To avoid this, a so called *free* set of 5-10% of reflections are removed from the refinement and used to calculate R_{free}. It shows how well the model predicts the structure factors of reflections which are not used in the refinement. The remaining 90-95% of the reflections are called *work* set and from them R_{work} is calculated. To not over-fit the data, the R_{work} and R_{free} should stay similar: as a rule of thumb, the difference between them should not exceed 0.05. The values of R_{work}/R_{free} typically range between 0.15-0.30 depending on the quality of the data.

2.3 Radiation damage

In the process of elastic scattering, responsible for the diffraction effects described previously in this chapter, the photon energy is conserved, i.e. no energy is deposited in the sample. However, at the X-ray energies typically used in macromolecular crystallography (∼5-15 keV), elastic scattering cross section is orders of magnitude smaller than the cross section of inelastic effects, primarily of photon absorption (Fig. 2.6). This means that for every coherently scattered photon contributing to the diffraction, there will be several tens of photons absorbed by the atoms each ejecting a photo-electron. This electron will have enough energy to cause few hundreds of ionization events until it thermalizes, producing many secondary electrons. In the ionized atom electron from a higher energy level will fall into the vacancy left by the photo-electron, which will result in the energy release either in the form of characteristic fluorescent photon or an outer shell electron ejected from the atom in a process called Auger decay.

Photoelectric absorption causes the energy to be lost in the crystal resulting in formation of radicals, deterioration of crystal lattice and temperature rise in the sample. This effect is referred to as radiation induced damage. The measure of the energy loss in the sample per unit mass is called the dose, its SI unit is Gy = J/kg. When the crystal thickness d is much smaller than the attenuation depth l of the X-rays, which is usually the case in macromolecular crystallography, the dose absorbed by a crystal can be calculated as following:

$$D = \frac{E_{abs}}{m} = \frac{N_{ph} h\nu (1 - e^{-d/l})}{\rho V} \simeq \frac{N_{ph} h\nu l}{\rho S} = \frac{I_0}{l\rho} \tag{2.35}$$

where ρ is the density of the sample, $V = Sd$ is the irradiated volume and $I_0 = \frac{N_{ph} h\nu}{S}$ is the energy per unit area or fluence of the incident beam. The dose can then be roughly estimated, using typical values of the attenuation depth and the density. For a protein crystal containing no heavy atoms and X-rays with the wavelength of 1 Å, the attenuation depth $l \simeq 3600$ μm and the density is about 1.35 g/cm^3. For a more

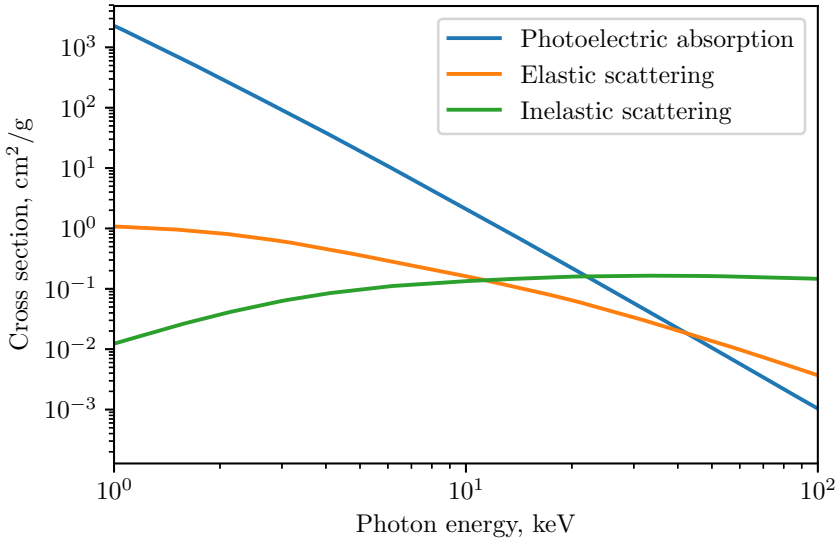


Figure 2.6: Atomic cross sections of carbon for photoabsorption, elastic and inelastic (Compton) scattering.

accurate calculation of the dose for an arbitrary wavelength and crystal contents there exists a commonly used software called RADDOSE [10].

The first study of radiation damage in macromolecular crystallography was published by Blake and Phillips in 1962 [11]. The radiation damage is generally divided into two kinds: global damage and specific damage. The first sign of the global damage is the fading of the diffraction intensity, particularly at a high resolution, with increase of the absorbed dose. In addition to the degradation of the resolution, increase of the Wilson B-factor, increase of the unit cell dimensions and increase in mosaicity (Section 3.3.1) are often observed. These metrics can be used as the indicators of the global damage [12]. The specific structural damage is inflicted on particular covalent bonds and can be observed in electron density maps. It was predicted already by Blake and Phillips [11], as they observed that the structure factors of some reflections increased while others decreased with the radiation dose, meaning that there were local structural changes occurring in addition to global radiation damage. Their hypothesis has since been confirmed, with the most prominent example being the cleavage of the disulfide bonds [13].

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, Henderson estimated a dose limit for macromolecular crystallography of three-dimensional crystals to be 20 MGy [14]. The Henderson limit was experimentally measured at 100 K to be 43 MGy [15], however the value of 30 MGy, corresponding to 0.7 instead of 1/2 of the preserved diffraction intensity, is generally used as the dose limit when planning diffraction experiment. Later study by Howells *et al.* [16] 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 the maximum tolerable dose vary substantially between different protein samples. In general, $D_{1/2}$ decreases by about two orders of magnitude when the temperature is increased from 100 K to 300 K [11, 17], giving the dose limit of 300 kGy. This fact gave rise to cryo-temperature crystallography, which stays the predominant technique for macromolecular structure determination since the early 1990s. Outrunning radiation damage is one of the main drivers of the development of experimental methods, which will be the subject of the next chapter.

CHAPTER 3

Experimental methods in X-ray crystallography

Previous chapter provided an introduction to X-ray crystallography showing how X-ray diffraction on crystals can be used to solve their atomic structure. This chapter describes experimental aspects of X-ray crystallography. The first half gives an overview of different X-ray sources and discusses their main properties. The second half describes different data collection strategies used in crystallography to obtain a full set of diffraction intensities to solve crystal structure.

3.1 X-ray sources

Since their discovery by Wilhelm Röntgen in 1895, X-rays found many applications in various areas of science including physics, chemistry, biology and industrial research, as well as in everyday life where they range from medical uses for imaging or radiotherapy to airport security scanners.

Röntgen's discovery was possible thanks to invention of the electrical discharge tube by William Crookes and others in the late 19^{th} century. The Crookes tube consisted of two metal electrodes in a gas bulb between which a high electric voltage was applied. The high voltage ionizes gas molecules in the tube creating positive charged ions and free electrons which in their turn ionize more molecules in a chain reaction. The positive ions attracted to the cathode knock electrons out of its surface, which are then accelerated by high voltage and hit the anode and the walls of the tube at a very high velocity. When deflected by atomic nuclei of the anode or the tube walls material, the electrons emit X-rays of a broad energy spectrum in the process called bremsstrahlung. They also knock electrons of the atoms to higher energy levels, these electrons then return to their initial levels emitting X-rays with characteristic energies specific to each element. The resulting X-ray energy spectrum of an X-ray tube is shown in Fig. 3.1.

In modern X-ray tubes the electrons are produced by thermionic emission from a cathode heated by electric current. The anode is rotated in vacuum to allow for more efficient cooling. Such rotating anodes can produce up to 10^{10} photons/s/mm^2 and are commonly used in both medicine and as laboratory sources for scientific research.

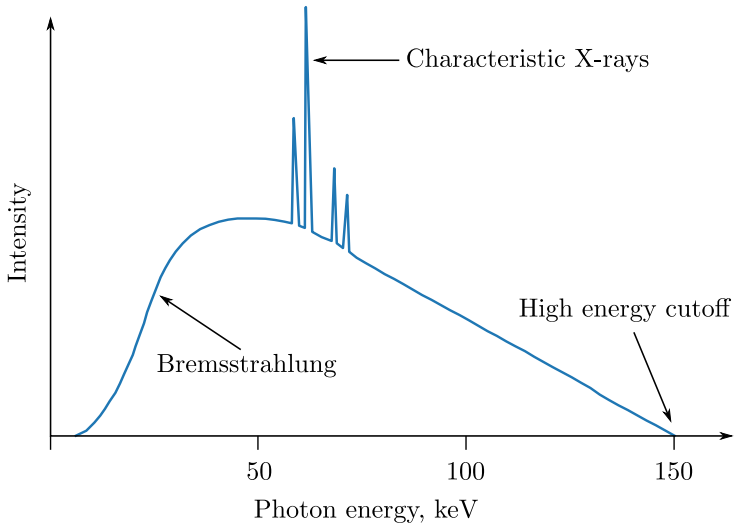


Figure 3.1: Typical X-ray energy spectrum of a tube source. The high energy cutoff is determined by the kinetic energy of electrons.

3.1.1 Synchrotron light sources

A different widely used source of X-rays is synchrotron radiation - a radiation emitted by relativistic charged particles, typically electrons, accelerated radially. It is produced, for example, when electrons are forced to move along a curved path by a magnetic field. The first experiments with such X-rays were performed using parasitic radiation from storage rings built for particle physics experiments, which are now called first generation synchrotrons.

The second generation synchrotrons, dedicated specifically to produce synchrotron radiation, were built in the early 1970s. They used single magnets, called bending magnets, to curve a trajectory of the electrons so that the electrons could be stored on a quasi-circular orbit. The X-rays were produced while the electrons accelerated in the magnetic field of the bending magnets.

In the third generation synchrotrons, which are used today, the electrons are kept on their orbit by bending magnets, but radiation is produced by periodic magnetic structures such as wigglers and undulators. They don't bend the orbit of the electrons therefore they can use much higher magnetic fields giving X-rays of higher power and frequency.

This section provides an introduction to synchrotron radiation and its properties. It is necessary to define here figures of merit used to describe and compare X-ray sources. The flux of a source is a number of photons per second per unit area:

$$\Phi = \frac{dN_{ph}}{dSdt} \tag{3.1}$$

Brightness is the flux per unit solid angle:

$$B = \frac{d\Phi}{d\Omega} \tag{3.2}$$

Brilliance is the number of photons within a bandwidth of 0.1% centered around a certain frequency per second per unit area per unit solid angle:

$$B_r = \frac{d^2\Phi}{d\omega d\Omega} \tag{3.3}$$

3.1.1.1 Bending magnet radiation

When a charged particle is accelerating, for example when it moves along a curved trajectory due to the Lorentz force in a magnetic field, it emits electromagnetic radiation. The power of the radiation in the non-relativistic case is given by the Larmor's formula:

$$P = \frac{q^2}{6\pi\varepsilon_0 c^3} a^2 \tag{3.4}$$

where ε_0 is the vacuum permittivity, c is the speed of light, q is the charge of the particle and a is its acceleration. The power distribution of the radiation is proportional to $\sin^2\Theta$, where Θ is the angle between the acceleration and the direction of observation, so the power is distributed over broad angular range.

If the particle velocity is close to the speed of light the angular range of the radiation is compressed in the forward direction of the particle movement, as can be shown by the Lorentz transformation:

$$\tan\Theta = \frac{\sin\Theta^*}{\gamma(\beta + \cos\Theta^*)} \tag{3.5}$$

where Θ^* is the angle observed in the frame of reference of a moving particle, $\beta = \frac{v}{c}$ and $\gamma = \frac{1}{\sqrt{1-\beta^2}}$ is the relativistic factor. As $\beta \simeq 1$, the radiation is collimated to a small angle $\Theta \simeq 1/2\gamma$: the more particle is accelerated, the more focused the radiation gets.

After the relativistic effects are taken into account, the generalized Larmor's formula for the total radiated power becomes

$$P_\gamma = \frac{2q^2 c\beta^4\gamma^4}{r^2} \tag{3.6}$$

where r is the bending radius of the orbit caused by the Lorentz force in the magnetic field.

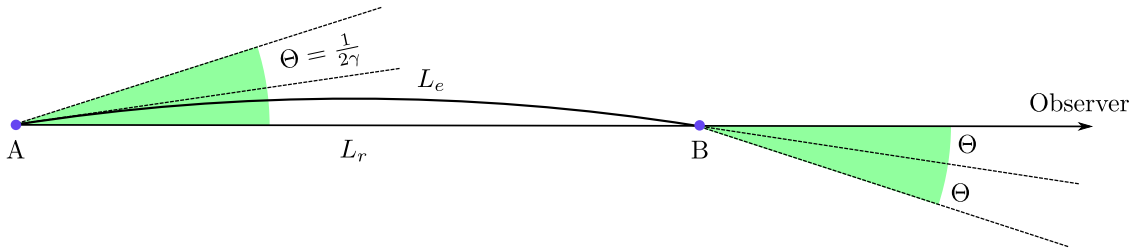


Figure 3.2: The electron travelling between points A and B emits radiation towards the observer in a cone with an opening angle Θ.

The spectral width of the radiation produced by electrons in a bending magnet can be qualitatively estimated through the Heisenberg's uncertainty principle $\Delta E \Delta t \geq \hbar/2$. The time Δt during which the radiation is detected by the observer (Fig. 3.2) is the difference between the time the electron passes the emission angle Θ and the time radiation travels the interval AB:

$$\Delta t = \frac{L_e}{v} - \frac{L_r}{c} = \frac{2r\Theta}{v} - \frac{2r\sin\Theta}{c} \simeq \frac{r}{\gamma\beta c}(1-\beta) \simeq \frac{r}{2c\gamma^3} \tag{3.7}$$

where r is the radius of relativistic electron orbit in the magnetic field B: $r = \frac{\gamma mc}{eB}$. The uncertainty in observed photon energies is then

$$\Delta E \geq \frac{2e\hbar B\gamma^2}{2m}. \tag{3.8}$$

The spectrum of the bending magnet will therefore be a broad spectrum centered around the critical energy

$$E_c = \frac{3e\hbar B\gamma^2}{2m}, \tag{3.9}$$

which is defined as the energy which divides the integral radiation power in half (Fig. 3.3).

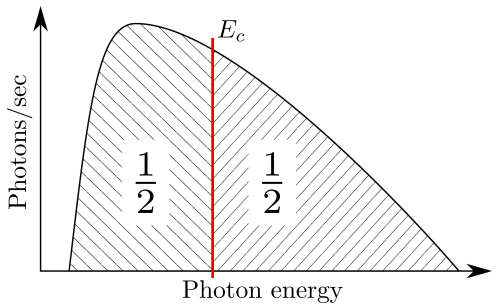


Figure 3.3: Critical energy E_c divides the integral radiation power in half.

3.1.1.2 Insertion devices

As can be seen from the Larmor formula (Eqn. 3.6), the radiation power is inversely proportional to the square of the radius of the electron orbit. Thus, there exists a limitation of the radiation power produced by the bending magnet: the radius can not be decreased as the electrons should be kept in the storage ring. In order to overcome this limitation, the periodic arrays of magnets with alternating polarity called insertion devices are used. They significantly increase the radiation power while keeping electrons on their intended orbit.

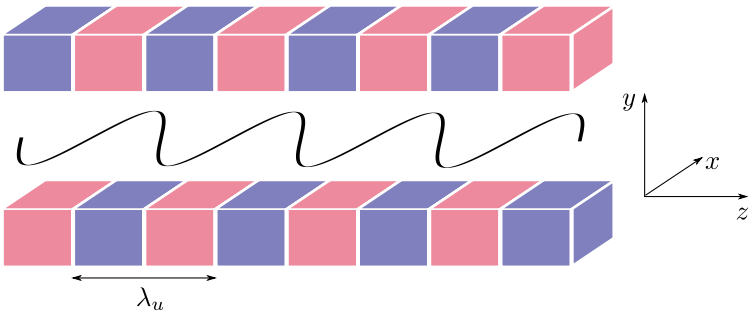


Figure 3.4: A schematic of the periodic magnetic structure of an insertion device with a period λ_u. Electrons travelling through the periodic magnetic field follow a sinusoidal trajectory shown in black.

The insertion devices (Fig. 3.4) create a sinusoidal magnetic field described as

$$\boldsymbol{B}(z) = B_0 \cos\left(\frac{2\pi}{\lambda_u} z\right)\hat{\boldsymbol{y}}. \tag{3.10}$$

The electrons in this magnetic field experience Lorentz force $F_l = e\boldsymbol{v} \times \boldsymbol{B}$:

$$\frac{d\boldsymbol{p}}{dt} = e\boldsymbol{v} \times \boldsymbol{B} \quad \Rightarrow \quad \frac{dp_x}{dt} = -ev_z B_y = -e\frac{dz}{dt} B_o \cos\left(\frac{2\pi}{\lambda_u} z\right). \tag{3.11}$$

Integrating this expression over time we obtain

$$m\gamma v_x = -\frac{eB_0\lambda_u}{2\pi}\sin\left(\frac{2\pi}{\lambda_u} z\right) \quad \Rightarrow \quad v_x = -\frac{eB_0\lambda_u}{2\pi m\gamma}\sin\left(\frac{2\pi}{\lambda_u} z\right) = -\frac{Kc}{\gamma}\sin\left(\frac{2\pi}{\lambda_u} z\right) \tag{3.12}$$

where $K = \frac{eB_0\lambda_u}{2\pi mc}$ is the non-dimensional undulator parameter. The maximum deflection angle from the $\hat{\boldsymbol{z}}$ direction can be estimated as $\phi_{max} \simeq \max\left(\frac{v_x}{v_z}\right) \simeq \max\left(\frac{v_x}{c}\right) = \frac{K}{\gamma}$.

Considering that the opening angle of the radiation emitted by the electron is $\frac{1}{\gamma}$, K gives the ratio between the maximum deflection angle of the electron trajectory and the angle of emission. Therefore two important cases are distinguished: when $K >> 1$ the device is called a wiggler. The wiggler shifts critical energy E_c towards the higher values compared to a bending magnet and increases the total radiation power proportional to the number of magnetic periods in the device, but the spectrum of the resulted X-rays stays qualitatively similar to the one of the bending magnet.

The device where $K \leq 1$ is called an undulator. In an undulator the maximum deflection angle is smaller than the angle in which the radiation is emitted, thus the X-rays emitted by electron at different times will overlap and interfere while electron travels along the undulator.

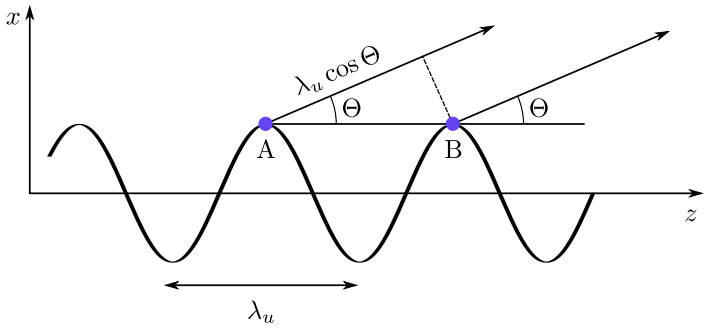


Figure 3.5: For constructive interference between the waves emitted by the same electron in an undulator, the path difference between the waves originated from the points separated by the undulator period must be a multiple of the wavelength.

Let us consider two waves emitted by the electron at an angle Θ from points A and B separated by the period of the undulator λ_u (Fig. 3.5). In order for the emitted wavelength λ to experience constructive interference, the path difference between these two waves should be a multiple of this wavelength. The time delay between points A and B is $\Delta t = \lambda_u/\overline{v}_z$, where $\overline{v}_z$ is the average longitudial velocity of electron. During this time the light emitted in A will propagate a distance of $c\Delta t$. The path distance between two waves then will be

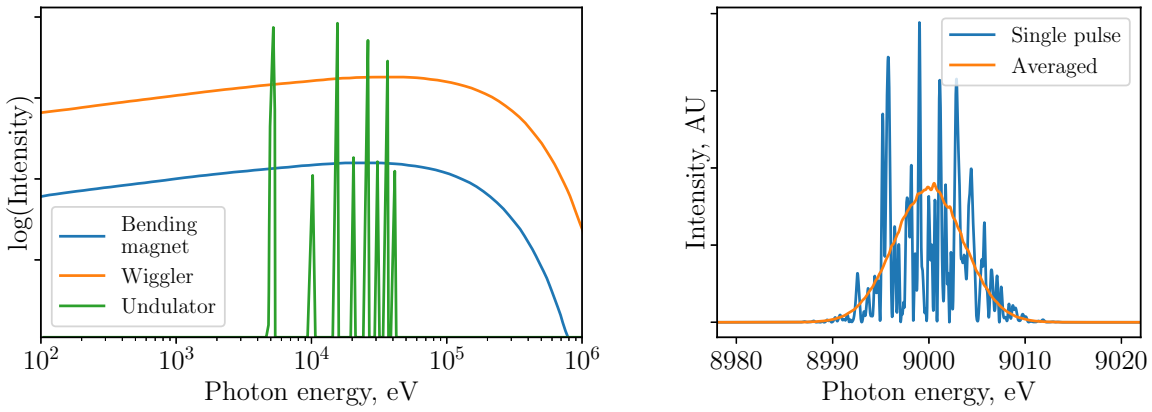
$$\frac{c\lambda_u}{\overline{v}_z} - \lambda_u \cos\Theta = n\lambda \tag{3.13}$$

The average velocity along $\hat{z}$ can be calculated from Eqn. 3.12:

$$\begin{aligned} v_z = \sqrt{v^2 - v_x^2} &= \sqrt{v^2 - K^2\frac{c2}{\gamma^2}\sin^2\left(\frac{2\pi}{\lambda_u} z\right)} = c\sqrt{1 - \frac{1}{\gamma^2}\Big[1 + K^2\sin^2\left(\frac{2\pi}{\lambda_u} z\right)\Big]} \\ &\simeq c\Big\{1 - \frac{1}{2\gamma^2}\Big[1 + K^2\sin^2\left(\frac{2\pi}{\lambda_u} z\right)\Big]\Big\} \quad \Rightarrow \quad \overline{v}_z = c\Big[1 - \frac{1}{2\gamma^2}\Big(1 + \frac{K^2}{2}\Big)\Big] \end{aligned} \tag{3.14}$$

When inserted into Eqn. 3.13 it gives the resonant wavelength of an undulator:

$$\lambda = \lambda_u \left[\frac{1}{1 - \frac{1}{2\gamma^2}\left(1 + \frac{K^2}{2}\right)} - \left(1 - \frac{\Theta^2}{2}\right) \right] \simeq \frac{\lambda_u}{2\gamma^2}\left[1 + \frac{1}{2}K^2 + (\gamma\theta)^2\right] \tag{3.15}$$



(a) (b)

Figure 3.6: (a) Schematic spectra of a bending magnet, a wiggler and an undulator. (b) Typical SASE spectrum of a single FEL pulse and averaged over 5000 pulses. Figure courtesy of S. Serkez.

Undulator usually has much more magnetic periods compared to a wiggler and weaker magnetic fields to keep the deflection angle of the electrons low. Due to interference the intensity of the radiation produced by an undulator increases proportional to the square of the number of magnetic periods unlike the wiggler where the dependence is only linear. The spectrum of an undulator has sharp peaks at the multiples of the resonant frequency - harmonics of the undulator. The width of the spectral peak is defined by the number of the undulator periods N_u as $\frac{\Delta E}{E} = \frac{1}{N_u}$. Schematic representation of the typical X-ray spectra of different synchrotron insertion devices is shown in Fig. 3.6a.

3.1.2 X-ray free electron lasers

The radiation emitted by electrons in the synchrotron bunch sums up incoherently as the bunch has no internal order. The intensity of the radiation is proportional to the number of electrons in the bunch and the pulse length is defined by the length of the bunch. Under favourable conditions, the energy can be transferred back and forth between the electrons and the generated electromagnetic radiation propagating along the undulator. Depending on the phase difference between electrons and the electromagnetic wave, half of the electrons gain energy from the radiation while the other half looses it. As a result, a periodic density modulation occurs in the electron bunch of the same period as the radiation field. This process is called microbunching. If the undulator is sufficiently long, the electrons in microbunches radiate coherently amplifying the radiation field and thus enhancing the microbunching even more. This leads to the exponential growth of the radiation power with the number of undulator periods until it saturates, when the balance between microbunching and repulsion forces between electrons is reached (Fig. 3.7).

The process of electron beam microbunching and subsequent amplification of the emitted radiation, referred to as Self-Amplified Spontaneous Emission (SASE), is a working principle of X-ray free electron lasers (FELs). A schematic view of an FEL is shown in Fig. 3.8a. Electron bunches in an FEL are created

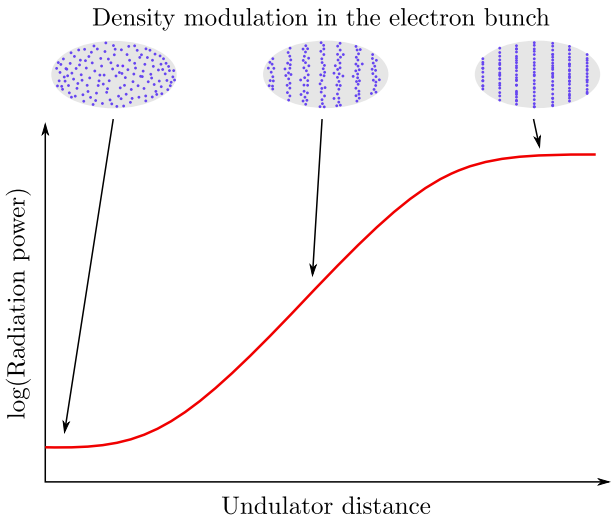


Figure 3.7: Growth of the radiated power and the electron beam microbunching with the undulator distance.

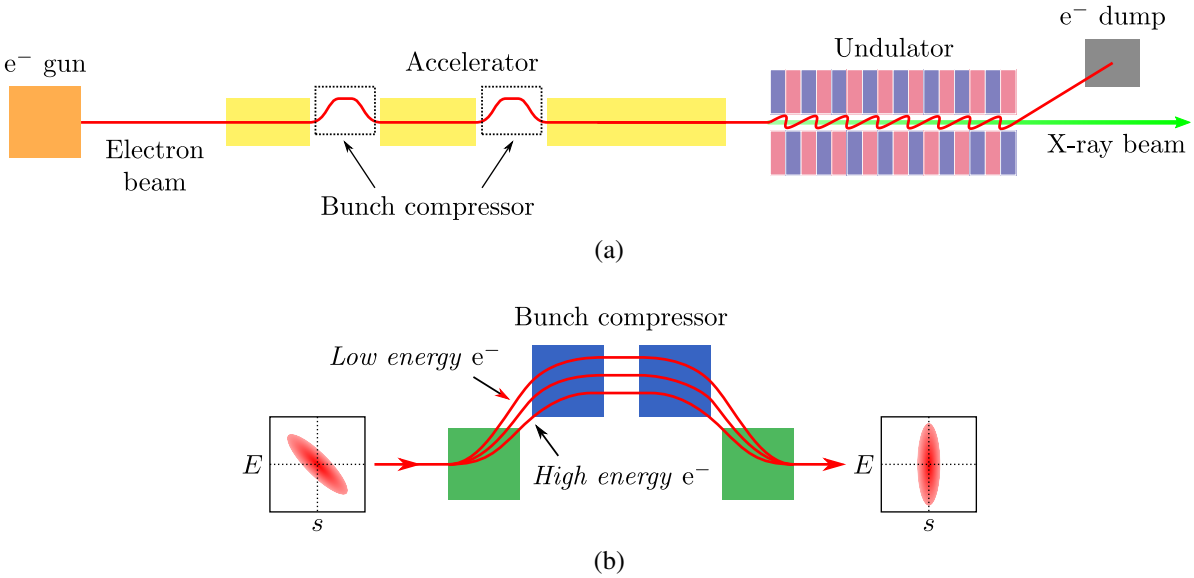


Figure 3.8: (a) Schematic layout of an FEL. Electron bunches are created in the electron gun and accelerated to relativistic velocities in a series of accelerators. To reduce the longitudinal size of the bunch the electrons are passed through the bunch compressor. The X-rays are then produced in a long undulator section. (b) Working principle of the bunch compressor. The left and right plots show the correlation between the energy of the electrons E in the bunch and their position s along the beam direction before and after passing through the chicane. After the accelerator electrons with the higher energy come in the tail of the bunch. In the magnetic chicane they travel a shorter distance and catch up with the lower energy electrons in the head of the bunch.

in the electron gun optimized for low emittance and injected into a linear accelerator, where they are accelerated to relativistic velocities by radio-frequency (RF) cavities. The bunches then pass through a long undulator section where the X-ray radiation is emitted. In order to achieve saturation of radiation power, the undulators used at FELs are much longer than the ones at a synchrotron. The resulting photon flux is proportional to the square of number of electrons.

To reduce the duration of the X-ray pulse electron bunches are shortened in a bunch compressor. After

the acceleration in the RF cavities the electrons in the tail of the bunch have higher energy compared to the leading ones. When they pass through the bunch compressor - a magnetic chicane consisting of four magnets, the electrons in the tail move on a shorter trajectory and are able to catch up to those in the head (Fig. 3.8b). This results in the creation of very short X-ray pulses of a duration down to tens of femtoseconds, which is orders of magnitude shorter than those of a synchrotron.

SASE starts up from the shot noise in the electron beam, which results in the X-ray spectrum having many sharp spikes corresponding to coherent radiation, unique for each pulse. Example of a simulated spectrum of a single FEL pulse as well as the averaged spectrum is shown in Fig. 3.6b. In comparison with insertion devices at the synchrotrons, X-ray spectrum produced by an FEL is relatively narrow with a typical bandwidth of $\frac{\Delta E}{E} = 0.2\%$. However, most of synchrotron beamlines, especially the ones dedicated for crystallography experiments, are equipped with a monochromator which reduces the undulator bandwidth to 0.01%. Compared to this FEL spectrum is significantly broader.

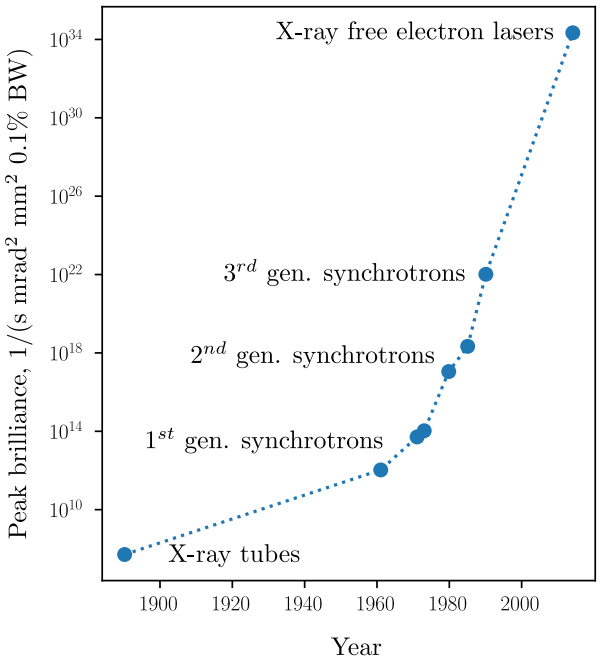


Figure 3.9: Development of X-ray sources over the years.

In the course of over hundred years of development, X-ray sources have gained 30 orders of magnitude in the peak brilliance (Fig. 3.9) with FELs being the brightest X-ray source as of today. Since the first hard X-ray FEL, The Linac Coherent Light Source (LCLS) at SLAC National Accelerator laboratory started operation in 2009, FELs have been opening new avenues in various areas of science. They triggered a development of new experimental techniques, one of which, serial crystallography, is the main topic of this work.

3.2 X-ray monochromators

All widely used X-ray sources, such as X-ray tubes and synchrotrons, generate polychromatic X-rays (Fig. 3.1 and 3.6a). Even harmonics of an undulator, although relatively narrow compared to

bremsstrahlung or wiggler spectra, still have a typical bandwidth of $\frac{\Delta E}{E} \simeq 0.05$. However, nowadays the majority of crystallography experiments are performed with monochromatic X-rays. A narrow range of X-ray energies is selected from the wide spectrum using a crystal monochromator - a single crystal with one face parallel to a major set of lattice planes (hkl). The principle of the crystal monochromator is based on the Bragg's law (Section 2.1.4): reflection hkl is observed at an angle Θ which depends on the wavelength λ: $\sin\Theta = \frac{\lambda}{2d_{hkl}}$. Therefore, by adjusting the incident angle Θ it is possible to select a particular wavelength from the polychromatic beam. To preserve the direction of the beam and narrow the bandwidth the X-rays can be diffracted twice by similar crystals as shown in Fig. 3.10.

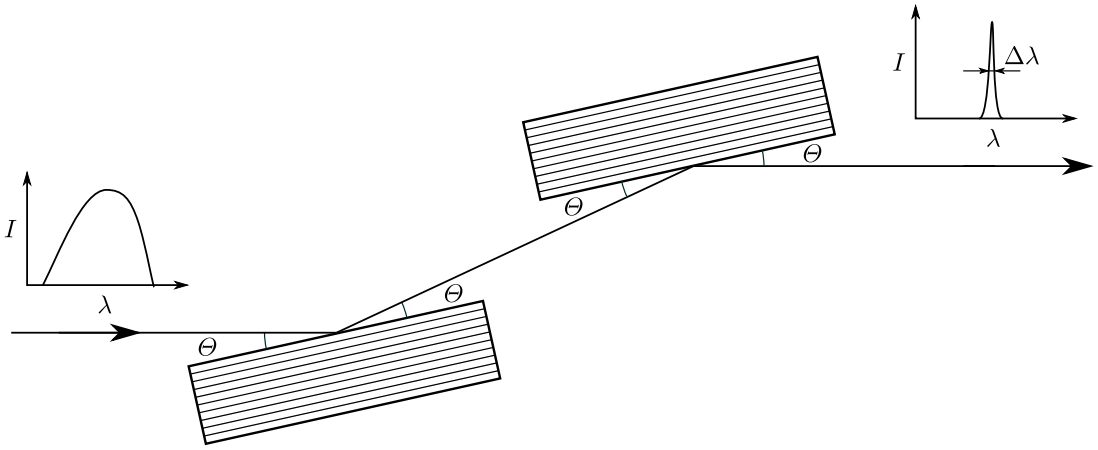


Figure 3.10: Schematic diagram of a double-crystal X-ray monochromator.

The bandwidth of the resulting radiation can be derived from the Bragg's law as

$$\frac{\Delta E}{E} = \frac{\Delta\lambda}{\lambda} = \cot\Theta\Delta\Theta. \tag{3.16}$$

Here, $\Delta\Theta$ is the Darwin width - a FWHM of the reflectivity curve derived in the dynamic theory of diffraction [18] as

$$\Delta\Theta = C\frac{\sqrt{F_{\boldsymbol{h}}F_{\bar{\boldsymbol{h}}}}}{V\cos 2\Theta}, \tag{3.17}$$

where $F_{\boldsymbol{h}}$ is the structure factor for the corresponding Bragg reflection, V is the unit cell volume and C is the proportionality constant. Thus, the bandwidth of the monochromator is

$$\frac{\Delta E}{E} = C\frac{\sqrt{F_{\boldsymbol{h}}F_{\bar{\boldsymbol{h}}}}}{2V\sin^2\Theta} = \frac{2C}{\lambda^2}\frac{d^2\sqrt{F_{\boldsymbol{h}}F_{\bar{\boldsymbol{h}}}}}{V}, \tag{3.18}$$

i.e. it is proportional to the square of the d-spacing between the lattice planes and the effective scattering density of the material for the corresponding Bragg reflection $|F_{\boldsymbol{h}}|/V$. A typical bandwidth of the most widely used Si(111) or Ge(111) monochromators is around 10^{-4}.

Broader bandwidth can be obtained using multilayers - layered periodic microstructures consisting of alternating heavy and light layers. Double multilayer monochromators have the same geometry as the double crystal monochromators (Fig. 3.10), but due to larger d-spacing and higher scattering density they can produce broader bandwidth of 10^{-2} or greater. For example, theoretical reflectivity curves of a perfect Si(111) crystal and W/B_4C multilayer with 25 Å bilayer thickness are compared in Fig. 3.11. Crystallographic experiment using multilayer monochromator is described in Section 6.4.

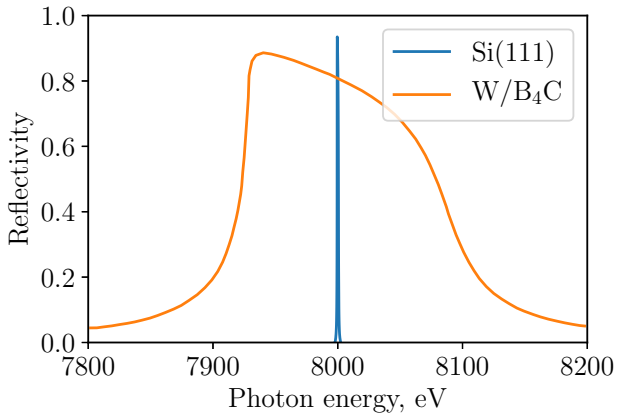


Figure 3.11: Theoretical reflectivity curve of a Si(111) crystal Bragg peak and first order Bragg peak of W/B_4C multilayer with the respective layer thickness of 7.5/17.5 Å [19].

3.3 Data collection techniques in X-ray crystallography

3.3.1 Reflection partiality

As it was shown in Section 2.1.3, the reflections of a perfect crystal have a finite dimensions in the reciprocal space defined by the shape of the crystal. Furthermore, in the real crystal atoms generally don't align into a perfect lattice. They form a mosaic crystal with many domains separated by lattice defects (Fig. 3.12). The measure of misalignment of the individual domains is called mosaicity of the crystal. Since different domains will diffract at different angles, the crystal mosaicity further increases the size and deforms the shape of the reciprocal lattice peaks.

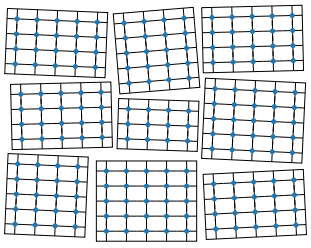


Figure 3.12: Mosaic model of a crystal. Misalignment of individual domains leads to broadening of the reciprocal lattice peaks.

In the experiment performed using monochromatic radiation and a crystal in a certain orientation, the reciprocal lattice points which intersect the Ewald sphere will give rise to a diffraction (Section 2.1.4). In a real experiment, even with a monochromator, X-rays have a finite bandwidth and are not perfectly parallel, having a certain degree of divergence. Therefore, Ewald sphere is not infinitely thin but has a certain thickness dependent on the scattering angle. Considering that both reciprocal lattice nodes have certain dimensions and the Ewald sphere has a certain thickness, in such experiment only diffraction from a cross-section between these reflections and the Ewald sphere can be recorded. Therefore, reflection intensity can be measured only partially unless the whole reciprocal lattice node lies fully within the Ewald sphere. This is illustrated in Fig. 3.13, where from four reflections intersecting the Ewald sphere at a given experimental geometry only the intensity of reflection A can be fully measured. The fraction of

the full reflection intensity observed in the experiment is called reflection partiality:

$$p = \frac{I_{obs}}{I_{full}} \tag{3.19}$$

To solve a crystal structure it is necessary to collect a complete set of structure factor amplitudes up to the highest possible resolution. In order to obtain the structure factor amplitude of a reflection either the diffracted intensity from the entire volume of the reciprocal lattice node or accurately calculated partiality is required. Estimation of reflection partialities is very difficult as it requires accurate knowledge of experimental parameters including beam bandwidth and divergence as well as the unit cell parameters of the crystal, its orientation and its mosaicity. While there are methods to estimate these parameters, they still rely heavily on having a large number of fully integrated reflection intensities [20, 21]. Therefore, experimental methods in crystallography usually attempt to measure reflections fully.

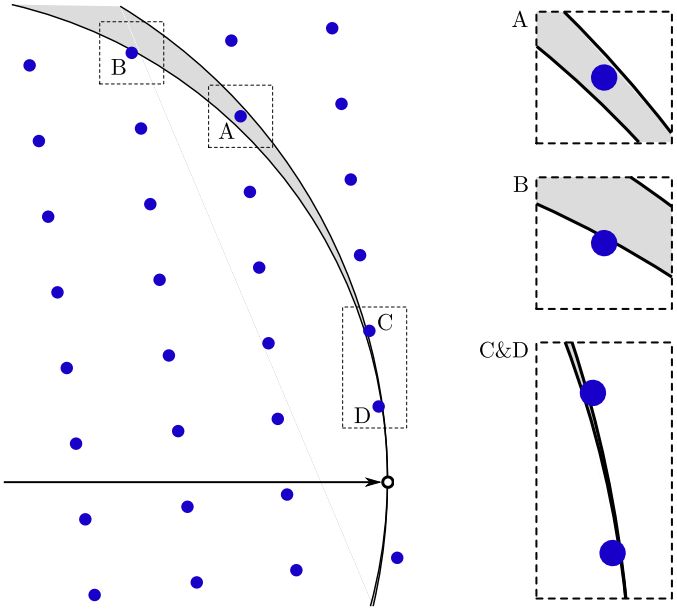


Figure 3.13: Illustration of the partiality problem in crystallography: the intensity of reflection A is integrated fully while the intensities of reflections B, C and D are only partially integrated.

This section gives an overview of the main data collection techniques used in crystallography to obtain a complete dataset of the full diffraction intensities. A special case, serial crystallography, where full diffraction intensities are obtained solely from partially recorded reflections is described in the next section.

3.3.2 Laue crystallography

One of the ways to obtain the full diffraction intensities is to use polychromatic radiation. This technique, also called the Laue method, was used in the original discovery of X-ray diffraction by crystals. Fig. 3.14 shows the Ewald construction in the case of polychromatic radiation with the wavelengths ranging between λ_{min} and λ_{max}. In this case the Ewald sphere becomes a shell between two limiting spheres with radii of $1/\lambda_{min}$ and $1/\lambda_{max}$. Different cross-sections of the reciprocal lattice nodes give rise to diffraction as they are excited by the X-rays of different wavelengths. Therefore the intensities of the reflections which lie fully within the shell between the limiting Ewald spheres are fully integrated. By measuring diffraction of

a crystal in a sufficient number of different orientations, defined by the radiation bandwidth, a complete dataset of diffraction intensities can be collected and used for structure determination.

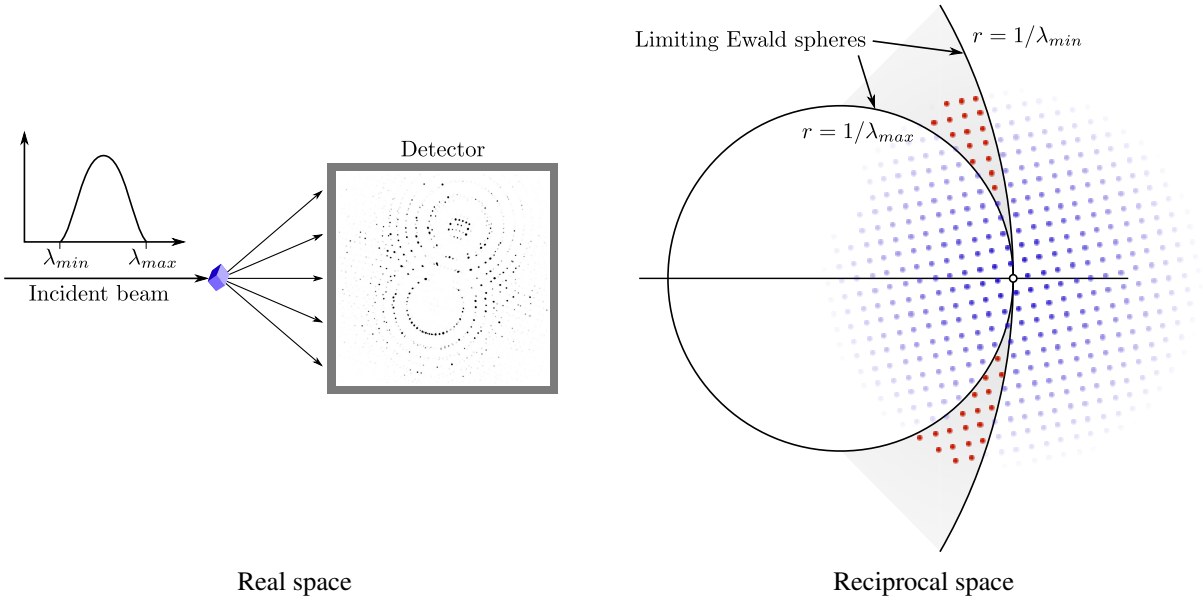


Figure 3.14: Laue method. If a stationary crystal is put into the polychromatic X-ray beam with the wavelengths ranging between λ_{min} and λ_{max}, the Ewald sphere becomes a shell (shown in gray) between two limiting spheres with radii of $1/\lambda_{min}$ and $1/\lambda_{max}$. All reflections lying fully within the shell (shown as red circles) will be fully integrated.

In practice, the polychromatic diffraction data is much more difficult to interpret compared to monochromatic data and setting up the experiment using polychromatic radiation at a modern X-ray source, such as 3rd generation synchrotron beamline, presents a number of challenges, therefore the monochromatic techniques are much more widely used. However, thanks to the rising demand in the field of macromolecular crystallography, namely time-resolved crystallography (Section 3.4), for the short exposure times achievable at a synchrotron source only with the use of polychromatic radiation, the Laue method is becoming popular again, in particular in combination with the serial diffraction method described in the next section. The data analysis of serial Laue diffraction constitutes a significant part of this work and is discussed in more detail in Chapters 6 and 7.

3.3.3 Single crystal rotation

Conventional data collection with monochromatic X-rays uses the rotation series method. In order to record the full reflection intensities, the crystal is rotated with respect to the incident beam. A diffraction pattern is recorded for each small angle increment $\Delta\phi$ usually between 0.1 and 1 degree. Each reciprocal lattice node will then cross the Ewald sphere completely and its full diffraction intensity will be recorded either in one rotation pattern or over several consecutive patterns (Fig. 3.15).

The single crystal rotation is by far the most widely used crystallographic data collection method at both laboratory sources and synchrotron radiation facilities, in macromolecular crystallography in particular. The analysis of the rotation data is relatively straightforward with many existing software packages for automatic processing. With the development of high intensity X-ray sources and modern X-ray detectors, the collection of a complete dataset typically takes below 2 minutes [22]. As a result, it is

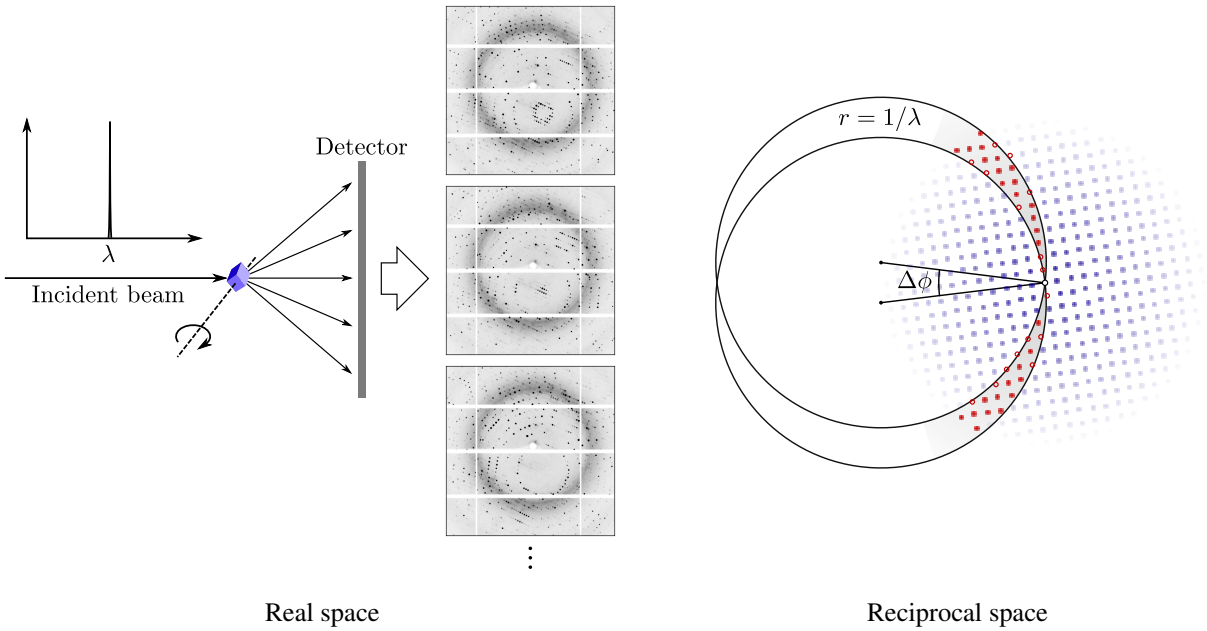


Figure 3.15: Single crystal rotation. The rotation motion of the crystal in the beam can be illustrated as the rotation of beam with respect to the stationary reciprocal lattice. If the diffraction pattern is recorded during the rotation over the angle $\Delta\phi$, the intensities of the reflections lying fully within the volume swiped by the Ewald sphere (shown as filled red circles) will be fully integrated. The reflections intersecting the Ewald sphere but not lying fully within the volume (shown as unfilled red circles) will be partially integrated, their full intensity will be recorded over several consecutive patterns.

highly reliable technique for macromolecular structure determination with many synchrotron beamlines available around the world dedicated specifically for it.

3.3.4 Powder diffraction

The alternative approach to the single crystal methods is data collection from a polycrystalline material or powder. An ideal powder sample is composed of a very large number of small randomly oriented crystals. Each reciprocal lattice vector $\boldsymbol{H}_{hkl}$ will then be found in all possible orientations with respect to the incident X-ray beam, forming a sphere of radius $|\boldsymbol{H}_{hkl}|$ (Fig. 3.16). Thus, instead of a one point intersecting the Ewald sphere there will be a circle corresponding to each reciprocal lattice point. The powder diffraction pattern recorded on the two-dimensional detector placed perpendicular to the incident beam will consist of the series of concentric rings. Depending on the symmetry of the crystal, several reciprocal lattice points can contribute to the same diffraction ring, which must be taken into account during the analysis of powder diffraction.

Although single crystal diffraction is undoubtedly a better way to obtain reflection intensities and solve the structure, powder diffraction must be used in some situations, for example when the crystals of sufficient size are not available. The limitation of the powder method is that the rings begin to overlap at higher resolution and become impossible to resolve, which is especially the case in macromolecular crystallography where the unit cells are large. Nevertheless, the method has been successfully applied to solve protein structures and stays as a valuable complementary technique to single-crystal measurements [23].

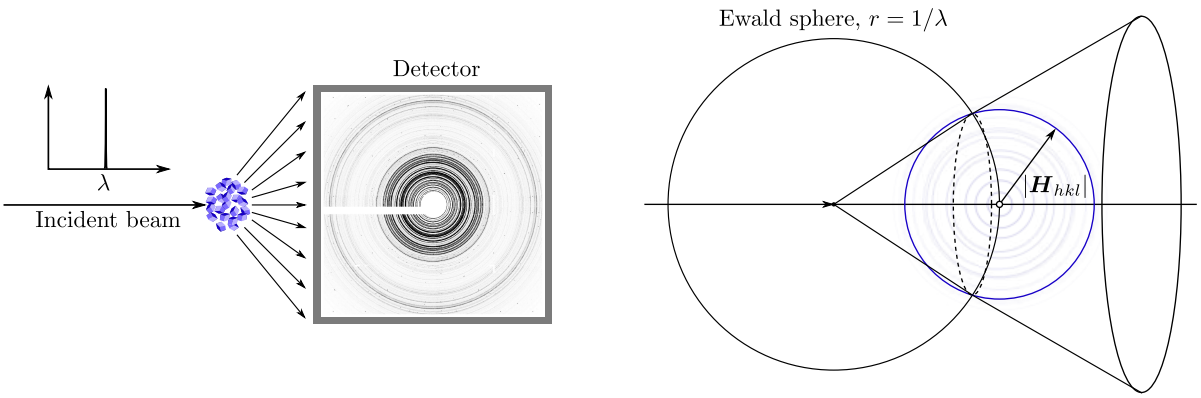


Figure 3.16: Powder diffraction. If the sample is a powder, the reciprocal lattice becomes a series of concentric spheres, shown in blue, corresponding to each reciprocal lattice vector $\boldsymbol{H}_{hkl}$, centered at the origin of the reciprocal space. The intersection of the sphere of radius $|\boldsymbol{H}_{hkl}|$ with the Ewald sphere is shown as a dashed circle. It defines a cone of all directions in which diffraction is observed.

3.4 Time-resolved crystallography

Conventional crystallography experiments allow to obtain static protein structures representing a time average over the duration of the X-ray exposure and a space average over all the molecules in a crystal. To understand how proteins function, knowledge of their static structures is necessary but often insufficient. While proteins are at work they usually experience conformational changes, i.e. subtle changes in their 3D structures, which have to be determined in order to characterize the working mechanism of the protein.

Crystallography provides several ways to examine the dynamics of biological macromolecules by capturing intermediate states of the molecules during the externally triggered reactions. In general, the approaches to study protein dynamics can be divided into two types: the first is to trap an intermediate by extending its lifetime to a typical data collection time with monochromatic X-rays [24]. This can be achieved, for example, by triggering the reaction in the crystal at ambient temperature and then rapidly cooling the crystal when a significant fraction of molecules accumulates in the intermediate state. A disadvantage of this method is the loss of information about the rate at which reaction occurs. The second approach, called time-resolved crystallography, uses X-ray exposure shorter than the actual lifetimes of the intermediates, capturing proteins in action at ambient temperature [25, 26]. Historically, time-resolved crystallography employed Laue method for data collection (Section 3.3.2). Using polychromatic beam allowed to achieve necessary photon flux to collect crystallography data with the X-ray exposures as low as 100 ps - a typical X-ray pulse duration at the third generation synchrotron source.

The first important part of a time-resolved experiment is reaction triggering: the reaction has to be initiated in as many molecules in the crystal as possible during the time shorter than the lifetime of the intermediates. A majority of the time-resolved crystallography experiments use so-called pump-probe method, where the reaction is triggered by a laser pulse. After a certain time delay after the pump laser pulse, the sample is probed by an X-ray pulse providing diffraction data from the molecules undergoing

the reaction at a sub-nanosecond time-resolution. The pump-probe sequence can be repeated several times for each diffraction pattern to reach the desired signal-to-noise ratio and the data has to be collected at a sufficient number of different crystal orientations to properly sample the reciprocal space. The experiment is usually repeated at different time delays to capture all reaction intermediates and measure their lifetimes. Conformational changes of the molecules are determined from the differences in the structure factors between the activated and dark states, so the dark structure data has to be collected from each crystal to avoid systematic errors due to lack of isomorphism between different crystals [25].

Since the reaction has to be triggered by the laser and it has to be repeated for each crystal orientation and pump-probe time delay, this method is limited to photosensitive proteins that undergo reversible reactions which relax back to the resting state in seconds or less. It also requires relatively large crystals, typically above 100 μm in size [27–30], which should be able to sustain significant radiation damage due to repeated X-ray exposures [25]. Nevertheless, since the first experiments in the 1990s, the method of time-resolved Laue crystallography has been highly successful. It provided important insight into dynamics of various proteins including photoactive yellow protein [29, 31, 32], myoglobin [27, 28, 33–36] and hemoglobin [37], at time scales starting from hundreds of picoseconds. The following section describes how the limitations of the time-resolved Laue crystallography have recently been overcome by the new method of serial crystallography.

3.5 Serial crystallography

Since the 3rd generation synchrotrons started their operation, the large improvement in the X-ray brightness allowed crystallographic data collection from very small protein crystals of a few micrometers in size. However, the dose rates inflicted on the sample made the radiation induced damage the major bottleneck for structure determination, even at cryo-temperatures (Section 2.3). The high resolution reflections usually disappear long before the full dataset has been collected. The traditional approach to overcome this limitation is to use several crystals or different parts of one crystal to distribute the absorbed radiation over the larger sample volume.

As the X-ray free electron lasers became available (Section 3.1.2), an alternative technique solving the radiation damage issue has been developed, called serial crystallography. FELs provide extremely bright ultrashort X-ray pulses with the energy of 2-4 mJ and the duration on the order of tens of femtoseconds. In 2000, Neutze *et al.* simulated the interaction of the FEL radiation with the biological macromolecules and predicted that the FEL pulses are short enough to provide useful diffraction data before the radiation damage destroys the sample [38]. The first experimental demonstration confirming the validity of this principle was done using the FLASH soft X-ray FEL in 2005: the reconstructed image from a coherent diffraction pattern from a two-dimensional specimen showed no radiation induced damage even though the sample was completely vaporized after a single FEL pulse [39]. While it is possible to collect a diffraction pattern from a sample unaffected by the radiation damage, the intensity of a pulse is so high that the sample immediately vaporizes. This process is usually referred to as ‘diffraction before destruction’.

As a result, only one diffraction pattern can be collected from each individual crystal, providing only minimal part of the information required to solve the structure. In serial crystallography this issue is addressed by delivering a new crystal into the beam for each new X-ray pulse collecting single diffraction images from many thousands of individual crystals. Serial crystallography was first successfully demonstrated by Chapman *et al.* in 2009 [3]. In the experiment, performed at AMO beamline at LCLS,

the structure of photosystem I was obtained at 8.5 Å resolution using diffraction data from over 15000 nanocrystals. The resolution in this case was limited by the experimental geometry and relatively low X-ray energy of 1.8 keV (6.9 Å wavelength). Following this groundbreaking experiment, serial crystallography method rapidly progressed in the last 10 years, becoming an established technique for macromolecular structure determination at FELs. Furthermore, after the first demonstration by Stellato *et al.* in 2014 [40], serial crystallography has also been adapted to synchrotron sources.

Typical experimental setup for serial crystallography is illustrated in the Fig. 3.17. Originally, serial crystallography experiments were performed using liquid jets to inject crystals into the interaction region. Several other methods of delivering crystals into the beam have since been developed and are described in more detail in the next section. The X-rays are focused onto the jet and the diffraction image is recorded on the detector after each pulse, regardless whether the crystal has been hit or not. Hybrid pixel-array photon-counting detectors, commonly used for conventional crystallographic data collection at synchrotrons, have limited count rates of several megahertz [41]. They are not suited for the data collection at FELs, where one pulse of about 50 fs duration typically contains 10^{12} photons. CCD detectors are an older technology, with higher noise and slower readout, but they are integrating and hence were used in the first FEL experiments. Recently, hybrid charge-integrating pixel detectors, such as CSPAD [42], AGIPD [43] and JUNGFRAU [44], have been developed specifically for FELs, but can also be beneficial in high-flux synchrotron experiments, as it will be shown in Chapter 6.

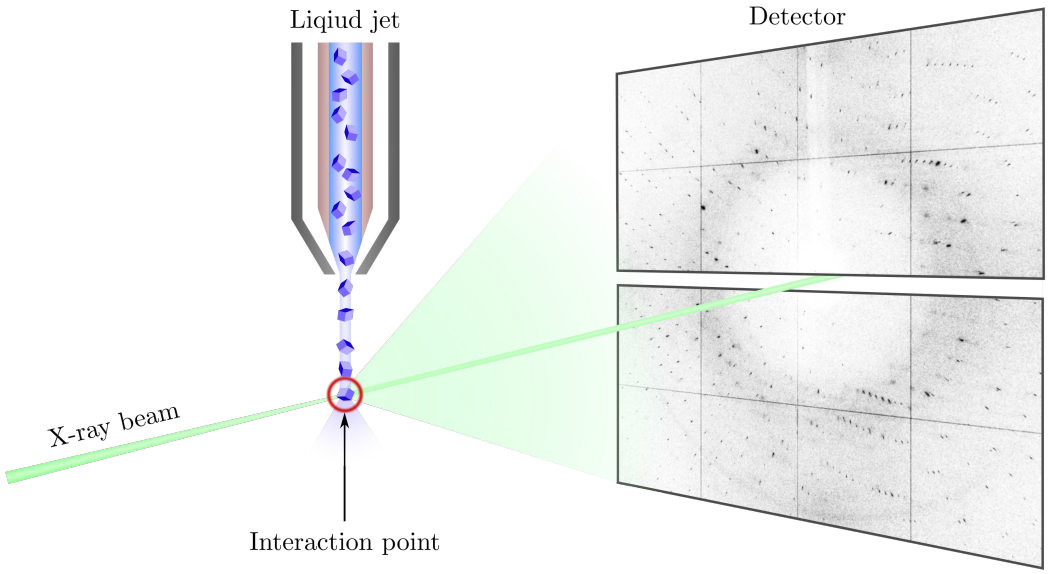
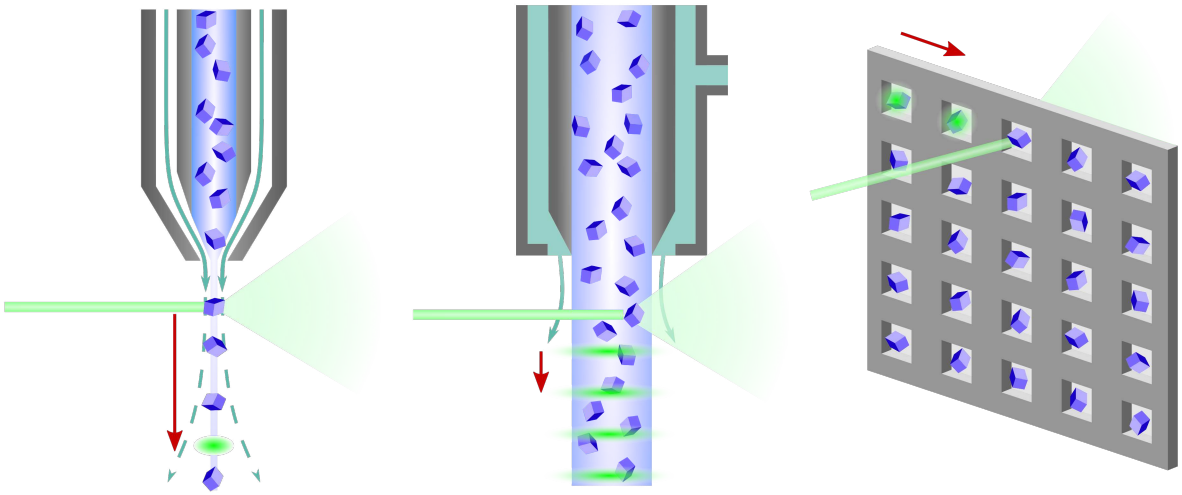


Figure 3.17: Scheme of a typical setup for serial crystallography experiment.

The important application of serial crystallography is time-resolved measurements. Serial crystallography is perfectly suited to study irreversible reactions since the sample is only once exposed by X-rays. Very short FEL pulses provide a possibility to observe laser induced protein dynamics at a time scales down to sub-picoseconds [45, 46]. The use of small crystals gives a benefit of larger crystal volume fractions activated by the optical laser in the pump-probe serial crystallography experiments compared to traditional time-resolved Laue experiments, as well as larger diffusion volumes in the mix-and-inject experiments, allowing to study ligand-triggered biological reactions starting at a sub-millisecond times [8, 47, 49].



(a) GDVN (b) LCP-jet (c) Fixed-target

Figure 3.18: Sample delivery methods in serial crystallography: (a) liquid jets give low background, but have high sample consumption because of their high speed, moving the sample a long distance between X-ray pulses; (b) extrusion jets are slower, giving higher sample efficiency, but at the cost of higher background; (c) raster-scanned fixed-targets can move fast enough to hit every point, giving optimum sample efficiency and data collection rate.

3.5.1 Sample delivery

Serial crystallography employs a variety of methods to deliver individual micro- or nanocrystals to the X-ray beam for data collection [50, 51]. In order to efficiently utilize every X-ray pulse, the sample must be introduced and replenished to the beam at a high speed depending on the pulse repetition rate. The first established sample delivery method exploits flow-focusing injectors, known as gas dynamic virtual nozzles (GDVN) [52]. They create micrometer-sized jets accelerating liquid coming from a capillary of an order of magnitude larger diameter by means of a coaxial flow of the sheath gas (Fig. 3.18a). The flow rate needs to be sufficiently high to achieve and sustain a stable jet, which leads to very high flow speeds of several meters per second. As a result, the vast majority of the sample goes by between the X-ray pulses without being exposed to the beam and not contributing to the data collection. While the advantage of the flow-focusing injectors is the low scattering background from only few micrometers of liquid, which is ideal for measuring sub-micrometer sized crystals, the high sample consumption remains the main bottleneck of the method. The sample flow rate can be reduced by a factor of $\sim$5 using the so called double-flow focusing nozzle (DFFN), where the second sheath liquid is employed to initially focus the sample liquid prior to the final gas focusing [53]. Currently, liquid jets are the only sample delivery method sufficiently fast for FEL sources operating at megahertz repetition rates, such as EuXFEL and LCLS-II, and have already been successfully employed for the megahertz data collection at the former [54, 55].

One way to significantly reduce the sample consumption is to use a more viscous liquid. Lipidic cubic phase (LCP), a bicontinuous cubic phase composed of monoolein and water, is a highly viscous medium which has been proven to facilitate crystallization of membrane proteins [56]. Membrane proteins perform key roles in the cell's interaction with the surroundings and thus constitute the majority of the targets in drug design. The crystals grown in LCP are usually relatively small in size, and their achievable structure

resolution is limited by the radiation damage, which makes them an ideal sample for serial crystallography at FELs. For this purpose, a special injector was designed, which is able to extrude LCP from a small nozzle of about 50 μm in diameter (Fig. 3.18b) [57]. The extrusion velocity can be varied and made slow enough to probe most of the sample by the X-ray pulses but avoiding the radiation damage inflicted by the preceding pulse, leading to very high sample efficiency. However, the background scattering produced by the LCP jet is much stronger compared to liquid jets, therefore larger crystals have to be used, typically more than 5 μm in size. Several important membrane protein structures have been solved using the extrusion injector, including the angiotensin receptor and the rhodopsin-arrestin complex [58, 59].

Fixed-target techniques are the most sample efficient way to deliver crystals into the beam. The concept is similar to the traditional sample mounts, such as sample loops, used in conventional macromolecular crystallography, where the crystal is fixed on a solid support which is manipulated to move and rotate the crystal in the X-ray beam. By creating a bigger support allowing to hold large number of crystals, preferably in well-defined positions, and moving the support with respect to the X-ray beam in such way that each X-ray pulse hits individual crystal, it is possible to reach 100% sample efficiency and hit rate (Fig. 3.18c).

Several different fixed-target designs for serial crystallography have been developed in the recent years [60–64]. The crystals are usually placed on a thin membrane chip made of silicon or silicon nitride periodically patterned with microscopic wells or pores. If the pores are of an appropriate size, the crystals get trapped when the excess mother liquor is removed by blotting [62], resulting in very low background and hit rates close to 100% [2, 64, 65]. It is necessary to avoid degradation of the crystals from dehydration, which is usually achieved by sealing the chip between two membranes or by keeping it under the stream of humidified gas. The first approach is suitable for in vacuum measurements, while the second one has the advantage of giving lower background as it doesn't introduce any extra material into the beam. The chip is mounted onto the fast scanning stage allowing for a high-throughput data collection at the rate of 120 Hz at LCLS [65] and up to 1 kHz at the synchrotron as will be presented in Chapter 6.

3.5.2 Solving partiality problem

In serial crystallography a single snapshot diffraction pattern is acquired from each crystal at random orientation. Since it is not possible to rotate the crystal in the beam as it is immediately destroyed by an extremely powerful X-ray pulse, and the radiation doesn't have the bandwidth large enough to record the full reflection intensities, the reflections in each diffraction pattern are in general only partially recorded (Section 3.3.1). The problem of partiality is addressed in serial crystallography by averaging the integrated intensities from sufficiently large number of measurements [67]. This approach, called the Monte Carlo method, has been shown to produce useful intensities since the first FEL experiments [68]. For a pulse duration short enough to out-run radiation damage and a given pulse fluence, the number of patterns required to complete a high-quality dataset of structure factors depends on factors such as crystal size, symmetry, radiation bandwidth and beam convergence. For a typical FEL bandwidth of 0.2%, more than 10000 still diffraction patterns are required to obtain a high-quality dataset using Monte Carlo approach [69]. Using a much smaller bandwidth of only 0.01%, serial crystallography experiments with monochromatic synchrotron radiation require at least several ten thousands of still diffraction images to obtain a complete high-quality dataset [40].

Although several approaches have been attempted to account for partiality in serial crystallography [70–

72], none of them offer a general solution of partiality problem applicable in every experimental case. Large number of varying from shot to shot parameters, such as crystal size and mosaicity as well as the properties of the FEL X-ray beam (intensity, energy and bandwidth) due to stochastic nature of SASE, make it difficult to define exact experimental geometry for each diffraction pattern, which is necessary to estimate reflection partialities.

One way to overcome this problem is to increase the X-ray spectral bandwidth [69], which can be done at an FEL by tuning the electron beam or at a synchrotron by using the polychromatic radiation produced by an undulator (Section 3.1.1.2). Serial crystallography measurements with a full harmonic of the undulator with a bandwidth of 5.7% were recently demonstrated [2], where only 50 diffraction patterns were sufficient to obtain a high-quality dataset. This method is further explored in Chapters 6 and 7.

CHAPTER 4

Data analysis in serial crystallography

The data collection strategy in serial crystallography differs drastically from conventional crystallography, so analysis of serial crystallographic data requires specifically developed software. Two differences need to be addressed in particular. First, large amounts of diffraction images collected in a short time have to be handled efficiently. For example, measurements at LCLS with CSPAD detector at the repetition rate of 120 Hz generate approximately 2.5 Tb of data in one hour. Second, estimations of the structure factor moduli have to be obtained from a set of still diffraction patterns of randomly orientated crystals, for which conventional software cannot be applied. Once the structure factor moduli are determined, the crystal structure can then be solved using the standard crystallographic programs.

This chapter gives an overview of all data processing steps required to obtain a final set of reflection intensities from the dataset of raw diffraction images.

4.1 Review of processing serial crystallography data

4.1.1 Pre-processing and hit-finding

In serial crystallography crystals are delivered into the X-ray beam by means of a liquid jet or on a moving fixed-target support. As a result not every image recorded by the detector contains crystal diffraction. A typical value of the percentage of recorded images containing crystal diffraction, referred to as hit fraction, is about 5-10%, but hit fractions even lower than 0.1% have been observed in some experiments. Using the fixed-target sample delivery approach much higher hit fractions up to almost 100% can be achieved, but in such cases the majority of diffraction patterns are multiple hits, i.e. containing overlapping diffraction from two or more crystals, which complicates data processing. As of today, the majority of serial crystallography experiments are performed using liquid or viscous jets and result in collecting large amounts of blank images. Therefore, in order to reduce the disk space required to store the data and to speed up the subsequent processing, the first step in the analysis is to sort out images containing crystal diffraction.

Cheetah is the most widely used data reduction and pre-processing software for serial crystallography experiments at FELs [73]. The primary functions of *Cheetah* are performing detector corrections, identifying Bragg peaks, sorting crystal diffraction patterns and converting them into a facility-independent format for subsequent analysis. Detector corrections include identifying and flagging bad and saturated pixels, dark and common mode correction of each module and individual gain corrections for each pixel.

After the corrected image is obtained, *Cheetah* searches for possible Bragg peaks in the image, in most cases using so called *peakfinder8* algorithm. It finds all clusters of more than $n_{\min}$ but fewer than $n_{\max}$ connected pixels with values above a radially dependent threshold, determined from the radially averaged background intensity. If the number of found peaks with a sufficiently high signal-to-noise ratio exceeds a certain minimum number n_{peaks}, the frame is identified as hit. All found hits are then saved in HDF5 format together with positions and intensities of the found peaks and various instrument and experiment information.

For the FEL data described in Chapter 5, pre-processing was performed using *Cheetah*. For the synchrotron serial crystallographic data presented in Chapters 6 and 7, pre-processing was done with similar routines implemented in Python using *peakfinder8* algorithm from *Cheetah* for hit-finding.

4.1.2 *CrystFEL*: from diffraction images to *hkl* intensities

Once the patterns containing crystal diffraction are sorted, the next analysis steps include determination of crystal orientation, integration of reflection intensities in each pattern and merging integrated intensities into a final dataset. All these steps are performed by the software suite *CrystFEL*, which is a free and open-source software specifically developed for data processing in serial crystallography [11, 74]. The alternative packages are *cctbx.xfel* [76], *nXDS* [77] and *DIALS* [78]. Since *CrystFEL* is much more widely used in the community, accounting for 4 times more structures deposited in the PDB compared to all three other packages combined, and this thesis is largely concerned with development of *CrystFEL*, the alternative packages will not be further discussed here.

The main data processing steps performed by *CrystFEL* are summarized in the flow diagram in Fig. 4.1 and described in detail below.

4.1.3 Indexing

Determination of crystal orientation from a diffraction pattern, also called indexing as it is equivalent to assigning Miller indices to the found Bragg reflections, is performed by *CrystFEL* program *indexamajig*. It starts with projecting all peaks, found in a diffraction pattern during the pre-processing or within *indexamajig* itself, onto the Ewald sphere and determining reciprocal space coordinates of their corresponding reflections. Either the peak positions on the detector or the calculated reciprocal space coordinates are then passed to various indexing algorithms, which attempt to find a three-dimensional periodic lattice coinciding with the observed reflections.

Originally, indexing in *indexamajig* was solely performed by invoking external algorithms, such as *MOSFLM* [79], *DirAx* [80] and *XDS* [81], integrated in conventional crystallographic programs. My first contribution to *CrystFEL* was development of *asdf* - the first indexing algorithm implemented internally in *CrystFEL*. Description of the algorithm and its comparison to *MOSFLM* and *DirAx*, the only two indexers which worked reliably in *CrystFEL* at the time when *asdf* was introduced, are provided later in this chapter. Several other indexing methods have since been added to *CrystFEL*, including *TakeTwo* [82] and *Felix* [10] developed specifically for indexing multiple crystals in one diffraction pattern, *XGANDALF* [84] especially beneficial for indexing weak patterns with small number of detected spots and *pinkIndexer* [5] suitable for more difficult cases such as Laue diffraction or electron crystallography.

When the indexing solution is found the resulting lattice parameters are compared to the parameters of the expected unit cell. In the case where the expected unit cell is unknown several indexers (*MOSFLM*,

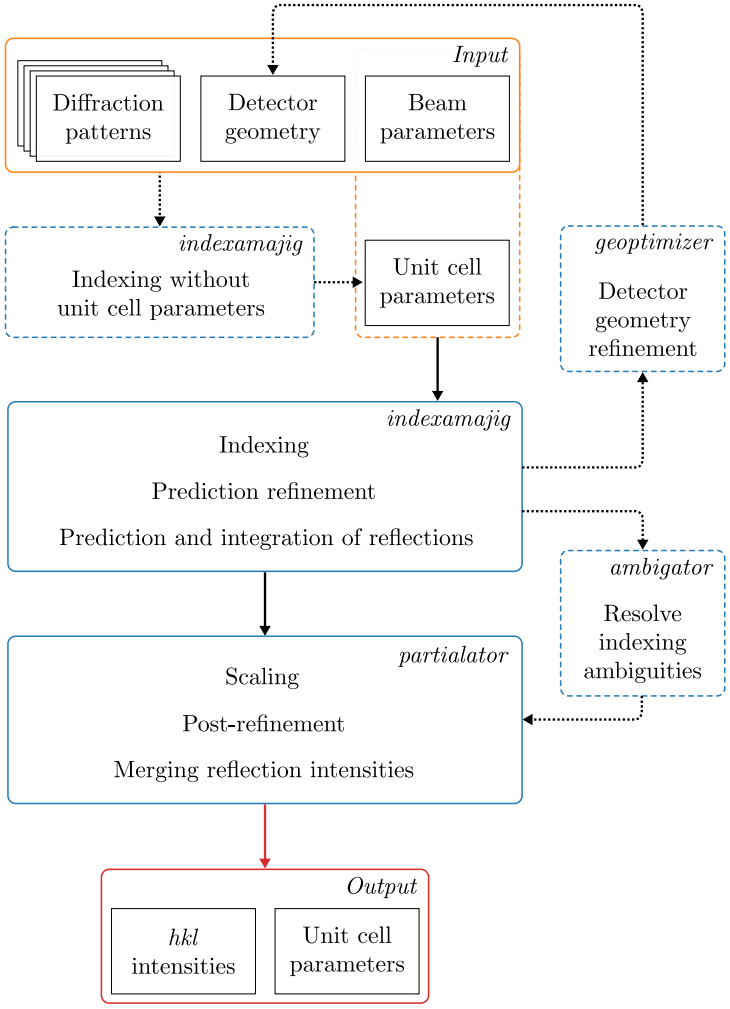


Figure 4.1: Serial crystallography data processing pipeline with *CrystFEL*.

DirAx, *asdf* and *XGANDALF*) can index diffraction patterns without prior knowledge of the parameters and give their initial estimations.

The found orientation matrix is then used to calculate the locations of the Bragg spots on the detector. The predicted locations are compared to positions of the found peaks: if a certain percentage of the found peaks match with the predicted reflections, indexing is considered successful. Percentage of indexed hits, also called indexing fraction, quantifies the success of the indexing procedure.

Using the found Miller indices of all correctly predicted peaks, i.e. peaks which have a matching predicted reflection within a certain distance, the crystal lattice and the detector center are then refined by minimizing the distances between the found and predicted spot positions in a process called prediction refinement [11]. Afterwards the reflection profile radius is calculated such that 98% of the spots that were assigned indices are predicted. Similarly, the individual diffraction resolution of each crystal is estimated to be at the 98th percentile of the scattering angles of predicted peaks. Using this value individual resolution cut-off can be applied to each crystal during integration and merging of intensities, the benefits of such approach are explored in more detail in the next chapter.

X-ray detectors used in serial crystallography, especially at FELs, are often segmented and designed

in a way to allow the modules to be moved relative to each other for different experiments. Therefore, in addition to refinement of the detector center, accurate determination of the exact position of each module is often necessary. By comparing the locations of observed and predicted peaks, the position, rotation and distance of each module relative to the interaction region are refined by *CrystFEL* program *geoptimiser* [86]. Usually several cycles of indexing and subsequent geometry refinement are required to obtain the best estimation of the detector geometry.

A significant limitation of serial crystallography in the first few years were indexing ambiguities. When the symmetry of the Bravais lattice is higher than the symmetry of the space group, the crystal in two or more different orientations would produce diffraction patterns with Bragg spots in identical locations but with different intensities. The indexing would then result in finding any of these orientations while only one of them is correct. The *ambigator* program in *CrystFEL* includes a simplified version of Brehm-Diedrichs algorithm [87], which can resolve such ambiguities using a clustering approach by calculating the correlations between the integrated reflection intensities [11].

4.1.4 Integration and merging of intensities

The main output of *indexamajig* program is a list of predicted reflections with their integrated intensities for each indexed crystal. Reflection intensities are measured using so called 'three-rings' integration method: three concentric rings centered at the predicted reflection position determine the peak, buffer and background estimation regions (Fig. 4.2). The integrated intensity is calculated as a sum of pixel values inside the smaller circle minus the background estimated from the annulus between the middle and outer circles. The importance of accurate peak prediction becomes apparent here: the closer the predicted reflection position is to the actual position of the Bragg peak the smaller inner ring radius can be used for its integration improving the signal-to-noise ratio of the integrated intensities.

4.1.4.1 Monte Carlo approach

The last step after indexing and integration is merging integrated reflections into a final set of hkl intensities, which can then be used for crystal structure determination. The simplest way to merge the intensities is Monte Carlo approach which averages integrated intensities of each symmetrically unique reflection from different crystals (Section 3.5.2). It is performed by *CrystFEL* program *process_hkl*. Given a large enough number of merged diffraction patterns to sample all possible crystal orientations, this method is equivalent to angular integration [67, 68], accomplished in conventional crystallography by rotating the crystal during the X-ray exposure.

4.1.4.2 Scaling and post-refinement

The convergence of the Monte Carlo process, i.e. the number of diffraction patterns required to achieve a desired uncertainty level of the merged intensities, is defined by several factors: detector noise and background scattering level, reflection partiality (Section 3.3.1), deviations of the crystal size and quality as well as fluctuations of the beam energy and spectrum, especially in the case of SASE beam at an FEL (Section 3.1.2). The effects of partiality can be reduced by increasing the bandwidth of the radiation or the beam convergence angle [69], although this would lead to several other experimental and data analysis challenges as discussed in Chapter 6. Alternatively, some of these factors, such as variations in the beam

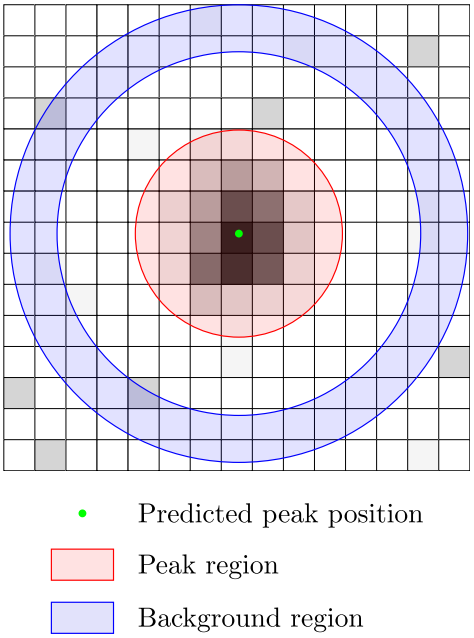


Figure 4.2: 'Three-rings' integration method in *CrystFEL*. The background is estimated as the average pixel value in the background region and subtracted from each pixel value in the peak region. Reflection intensity is calculated as a sum of pixel values in the peak region minus the average background.

intensity or crystal size as well as partiality can be accounted for during merging, improving the quality of the resulting intensities.

The more advanced intensity merging methods are implemented in *CrystFEL* program *partialator*. The key procedures performed by *partialator* prior to final merge of intensities include scaling, partiality correction and post-refinement [11, 70]. For each crystal two scaling factors, linear and Debye-Waller term $\exp(-B(\sin\Theta/\lambda)^2$ (see Eqn. 2.23), are determined in such way that, when applied, they bring the individual measurements of reflection intensities into as close an agreement as possible. The linear term accounts for variations on crystal size and beam intensity while B-factor scaling compensates deviations in crystal quality. The partiality is calculated using geometrical model described by White 2014 [70], which estimates the partiality as a fraction of the reciprocal lattice node volume excited by the X-rays (Fig. 3.13). Post-refinement means an iterative refinement of scaling and geometrical parameters for each crystal to achieve the best possible agreement of corrected reflections intensities.

While the partiality correction is still an experimental feature and only works in favorable cases [71, 88], the iterative scaling procedure consistently improves data quality and reduces the total number of merged crystals required to obtain high-quality intensities [11]. There are options to apply both scaling and partiality corrections or only any one of them. After the determined scaling factors and partiality estimates are applied to all crystals, the final intensity of each symmetrically unique reflection hkl is again calculated as the mean value of corrected intensities of all measurements. The errors in the merged intensities are estimated as a standard error of the mean:

$$\sigma_{hkl} = \frac{\sqrt{\sum(I_{hkl} - \overline{I_{hkl}})^2}}{N_{hkl}}, \tag{4.1}$$

where I_{hkl} is an individual measurement of reflection hkl, $\overline{I_{hkl}}$ is the mean of all such measurements or the final merged hkl intensity and N_{hkl} is the number of measurements of hkl reflection [74].

4.1.5 Evaluation of the data quality

Serial crystallography uses several traditional crystallographic figures of merit to evaluate quality of the final intensity dataset. They include

- average signal-to-noise ratio $\langle I_{hkl}/\sigma_{hkl}\rangle$,
- redundancy - average number of individual measurements of each reflection $\langle N_{hkl}\rangle$,
- completeness - fraction of all possible reflections which intensities were measured N_{meas}/N_{poss}.

In addition to these, due to the nature of serial crystallography approach of merging diffraction data from thousands of crystals, other metrics quantifying self-consistency of the data are used. They are based on randomly splitting the initial full dataset into two subsets of the same size. Each subset is then merged independently and two resulting sets of the merged intensities I_1 and I_2 are compared by calculating the following metrics:

- R_{split}, R-factor similar to R_{merge} used in conventional crystallography [74]

$$R_{split} = \frac{1}{\sqrt{2}} \frac{\sum_{hkl} |I_1 - I_2|}{\frac{1}{2}\sum_{hkl}(I_1 + I_2)}, \tag{4.2}$$

- Pearson correlation coefficient $CC_{1/2}$

$$CC_{1/2} = \frac{\sum_{hkl}(I_1 - \overline{I_1})(I_2 - \overline{I_2})}{\sqrt{\sum_{hkl}(I_1 - \overline{I_1})^2 \sum_{hkl}(I_2 - \overline{I_2})^2}}, \tag{4.3}$$

- CC^* derived from $CC_{1/2}$ to estimate correlation of the merged dataset with the true (usually unmeasurable) intensities using the assumption that errors in the subsets are random [89]

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}. \tag{4.4}$$

Traditionally, resolution of the diffraction dataset is estimated by signal-to-noise ratio with the typical $I/\sigma(I)$ cut-off values around 3-5. In serial crystallography a cut-off based on other metrics is used. Crystallographic data is usually significant if $CC_{1/2} \gtrsim 0.15$ or, equivalently, $CC^* \gtrsim 0.5$, therefore these values are often used to estimate resolution of the final dataset.

4.2 New indexing algorithm in *CrystFEL*

As mentioned before, my first contribution to *CrystFEL* was implementation of *asdf* indexing algorithm. The goal was not to develop a new indexing method but rather to implement one natively within *indexamajig* providing the possibility to use the program without any external software. Thus, due to its better speed compared to FFT based algorithms such as *MOSFLM*, *DirAx* was chosen as the base algorithm with a few modifications to improve the performance.

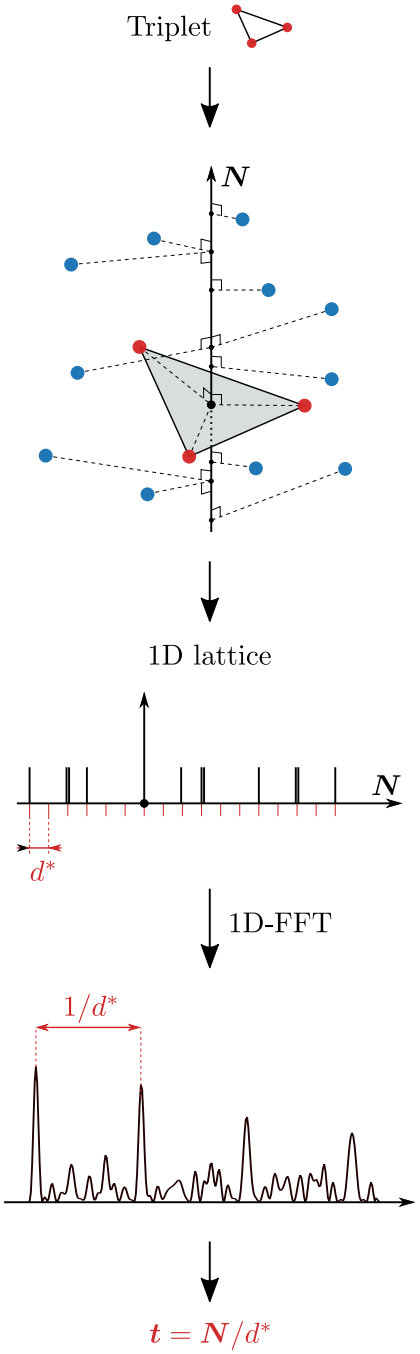


Figure 4.3: Finding possible direct lattice vector in *asdf*. For each triplet (red circles) a normal vector $\boldsymbol{N}$ is calculated. All other reciprocal lattice points (blue circles) are projected onto $\boldsymbol{N}$, forming a one-dimensional lattice. Period d^* of the 1D lattice is determined using 1D-FFT, giving a potential direct lattice vector $\boldsymbol{t} = \boldsymbol{N}/d^*$.

4.2.1 Implementation of *asdf* indexing algorithm

Starting from the reciprocal space coordinates of the observed diffraction spots, the algorithm consists of the following key steps [80]:

1. Generation of triplets - all possible groups of three of the observed reflections. Three nodes of the reciprocal lattice form a triangle. As follows from the definition of reciprocal lattice (Eqn. 2.14), the normal $\boldsymbol{N}$ to this triangle defines the direction of a direct lattice vector. If all other reciprocal lattice nodes are projected onto this normal they form a one-dimensional lattice with a period d^*.

2. Finding the period of one-dimensional lattice (Fig. 4.3). In *DirAx* the period is determined in an iterative procedure starting from the greatest distance between two consecutive projections, dividing it by an integer number and minimizing the sum of distances between projections and predicted 1D lattice points. In *asdf* the period is found using 1D-FFT instead, as it proved to be faster and more reliable.

3. When the period d^* is found, the vector $\boldsymbol{t} = \boldsymbol{N}/d^*$ is saved as a potential direct lattice vector together with the number of reflections n_t, which projections fit to the 1D lattice with the period d^* within a certain tolerance. The triplets are processed in random order. The search is stopped after 10000 triplets if the overall number of triplets is greater than that.

4. Unit cell search. The found possible direct lattice vectors $\boldsymbol{t}$ are sorted by decreasing number of fitting reflections n_t and increasing length. The highest occurring number n_t is set as acceptance level n_{min}. Starting from the shortest $\boldsymbol{t}$ vectors with the highest n_t, all combinations of three linearly independent vectors are tried as a candidate direct cell. If the number of reflections fitting to the candidate cell reaches n_{min} the search is finished, otherwise the search continues until the cell with the maximum number of fitting reflections is found.

* Here, in comparison to *DirAx*, when the expected lattice parameters are known only candidate unit cells with the volume within a certain tolerance from the expected primitive unit cell are tried by *asdf*. As will be shown later, it allows to significantly reduce the run time.

5. Unit cell reduction. The unit cell found by the algorithm is always primitive, but the resulting dimensions are usually not the shortest as they are built on randomly oriented triplets. Therefore, as a final step the primitive unit cell with three shortest dimensions is obtained from linear combinations of the found cell vectors, using a simple algorithm illustrated in Fig. 4.4. This gives consistent results which can then be easily compared to the expected unit cell.

4.2.2 Evaluation and comparison of *asdf* to *MOSFLM* and *DirAx*

To evaluate the accuracy and speed of *asdf* compared to other indexing methods available in *CrystFEL* at the time when *asdf* was first introduced, two datasets of simulated and experimental still diffraction patterns were indexed with *indexamajig* using *asdf*, *MOSFLM* and *DirAx*.

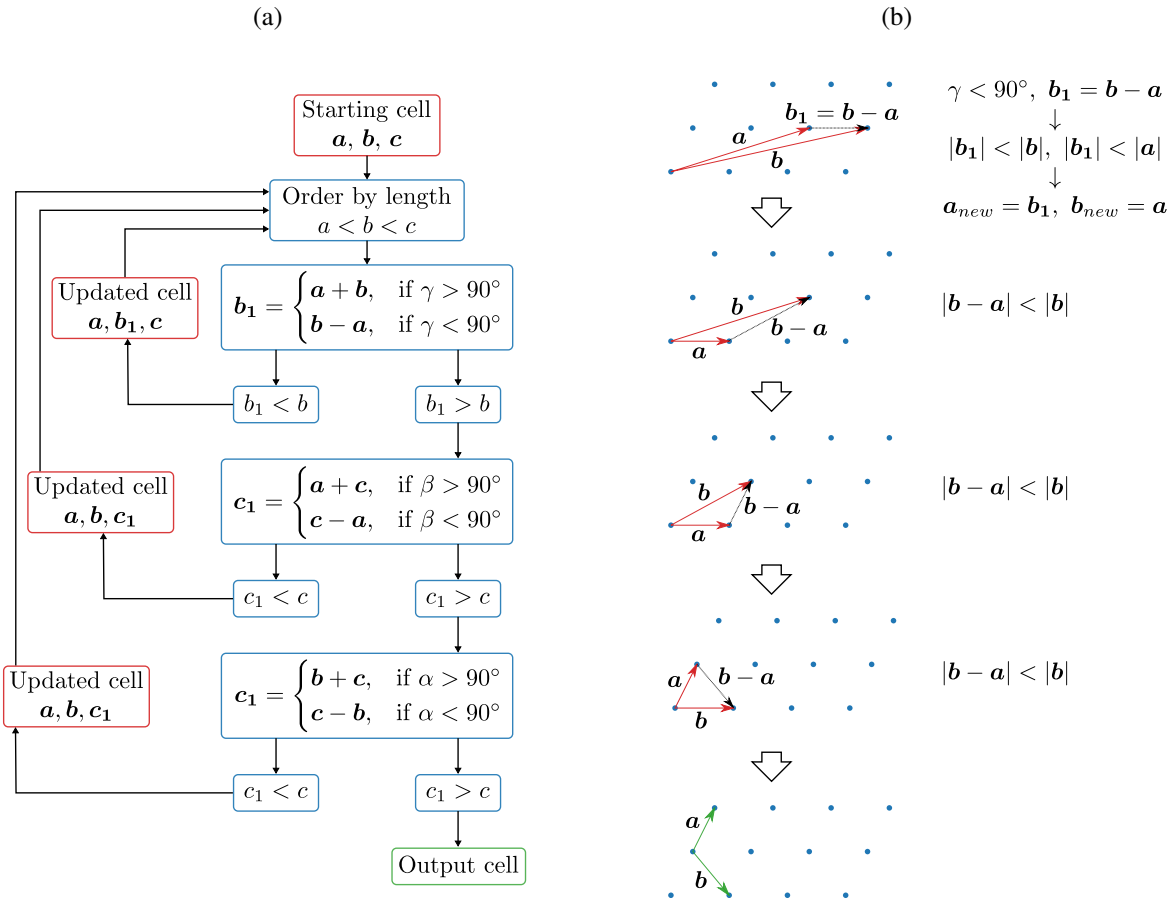


Figure 4.4: (a) Flow diagram of unit cell reduction algorithm in *asdf*. (b) Illustration of unit cell reduction in 2D case.

Indexing simulated data

To test the algorithm on simulated data, 1000 lysozyme diffraction patterns were generated with *CrystFEL* program *pattern_sim* with the number of peaks per pattern varying between 30 and 100. The data was indexed with and without target unit cell given as an input. All three indexers yielded 100% indexing fraction in both cases. The resulting execution times, shown in Table 4.1, demonstrate that both *asdf* and *DirAx* are noticeably faster than *MOSFLM*. Thanks to more efficient determination of 1D lattice period with FFT and usage of the prior knowledge of lattice parameters which allows to significantly speed up the unit cell search, *asdf* turned out to be about 25% faster than *DirAx* and more than two times faster compared to *MOSFLM* in case when the unit cell was provided.

To evaluate the accuracy of indexing results in the case of simulated data two metrics are used. Since both unit cell dimensions and lattice orientation are known exactly for this data they can be compared directly to the indexing results. The distributions of the obtained unit cell parameters and the misalignment angles between the simulated and found lattices are shown in Fig. 4.5 and 4.6, respectively. The mean values of unit cell parameters errors and misalignment angles are also given in Table 4.1. Both metrics showing similar results for all three indexing algorithms, although the distributions obtained by *MOSFLM* are slightly narrower and the errors are lower compared to *DirAx* and *asdf*. Nevertheless, the indexing solutions obtained by *asdf* are sufficiently accurate to correctly predict Bragg peak positions in the diffraction patterns and give a good estimations of the unit cell: the average parameters obtained by *asdf*

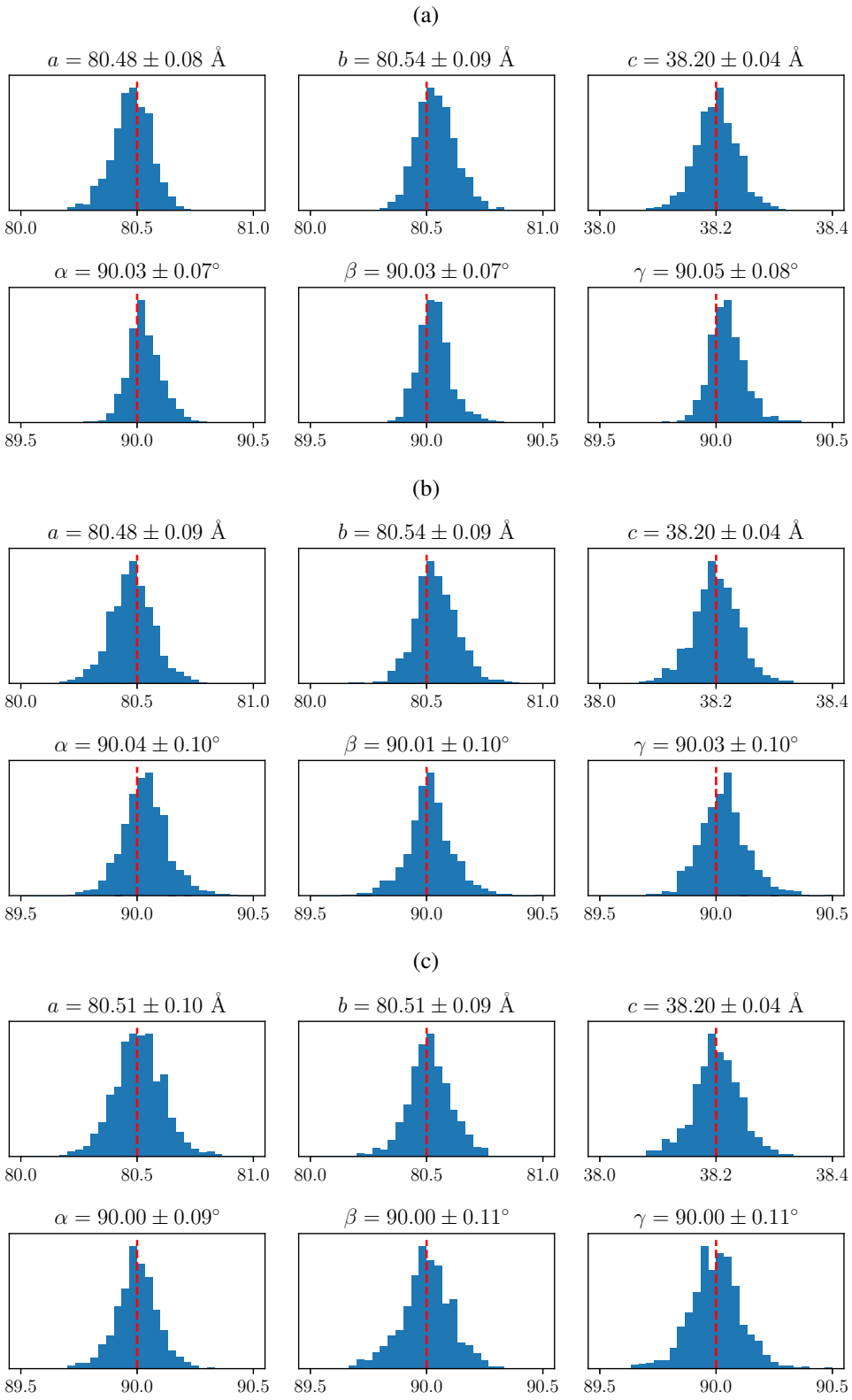


Figure 4.5: Distributions of the unit cell parameters obtained for simulated lysozyme data with (a) *MOSFLM*, (b) *DirAx* and (c) *asdf*. The numbers above each plot give the average ± standard deviation of the parameter. Dashed red lines show unit cell parameters used for simulation: $a = b = 80.5$ Å, $c = 38.2$ Å, $\alpha = \beta = \gamma = 90°$.

Indexing algorithm	*MOSFLM*	*DirAx*	*asdf*
Indexed patterns, %	100	100	100
Run time, min:sec			
without unit cell	14:08	09:00	08:22
with unit cell	13:25	07:21	05:26
Mean RMS error, %			
UC lengths $\langle \frac{\Delta a}{a}, \frac{\Delta b}{b}, \frac{\Delta c}{c} \rangle$	0.11	0.12	0.12
UC angles $\langle \frac{\Delta \alpha}{\alpha}, \frac{\Delta \beta}{\beta}, \frac{\Delta \gamma}{\gamma} \rangle$	0.09	0.11	0.11
Mean UC misalignment, °	0.08	0.09	0.09

Table 4.1: Comparison of execution time and indexing accuracy of three indexing algorithms processing simulated data.

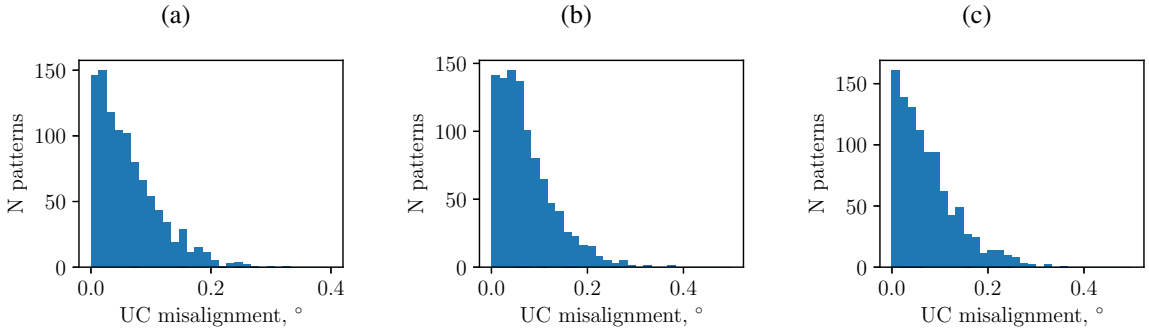


Figure 4.6: Distributions of the angles between the crystal lattices used in simulation and found by (a) *MOSFLM*, (b) *DirAx* and (c) *asdf*.

are all within 0.02% from the parameters used for simulation (Fig. 4.5c).

Indexing experimental data

A dataset consisting of 1000 diffraction patterns of *Cydia pomonella* granulovirus (GV) native occlusion bodies collected during the DFFN experiment at the CXI beamline at LCLS [53] was used to test indexing algorithm on experimental serial crystallographic data. Occlusion bodies are crystalline protein particles which form around the virus to protect it from the outside environment. GV crystals are very homogeneous in size and quality which makes them a good testing sample for serial crystallography. The crystals are very small, about 210×210×400 nm in size, and therefore produce relatively weak diffraction. Diffraction patterns used in this test contained on average only 35 observed peaks.

Diffraction patterns were again indexed with *indexamajig* using the three indexing methods. The expected unit cell parameters, $a = b = c = 103.3$ Å and $\alpha = \beta = \gamma = 90°$ were given as an input. The resulting indexing fractions and execution times are provided in Table 4.2. Although in this case *asdf* was not significantly faster compared to *DirAx*, it was still about 20% faster than *MOSFLM* and was able to index 4% more images than two other algorithms.

As the exact unit cell parameters and correct crystal orientations are not known for experimental data, different metrics are necessary here to evaluate the quality of indexing solutions. Firstly, the distributions of the obtained unit cell parameters are again plotted in Fig. 4.7. Both *asdf* and *DirAx* demonstrate broader distributions compared to *MOSFLM*, which likely means that *MOSFLM* solutions are more

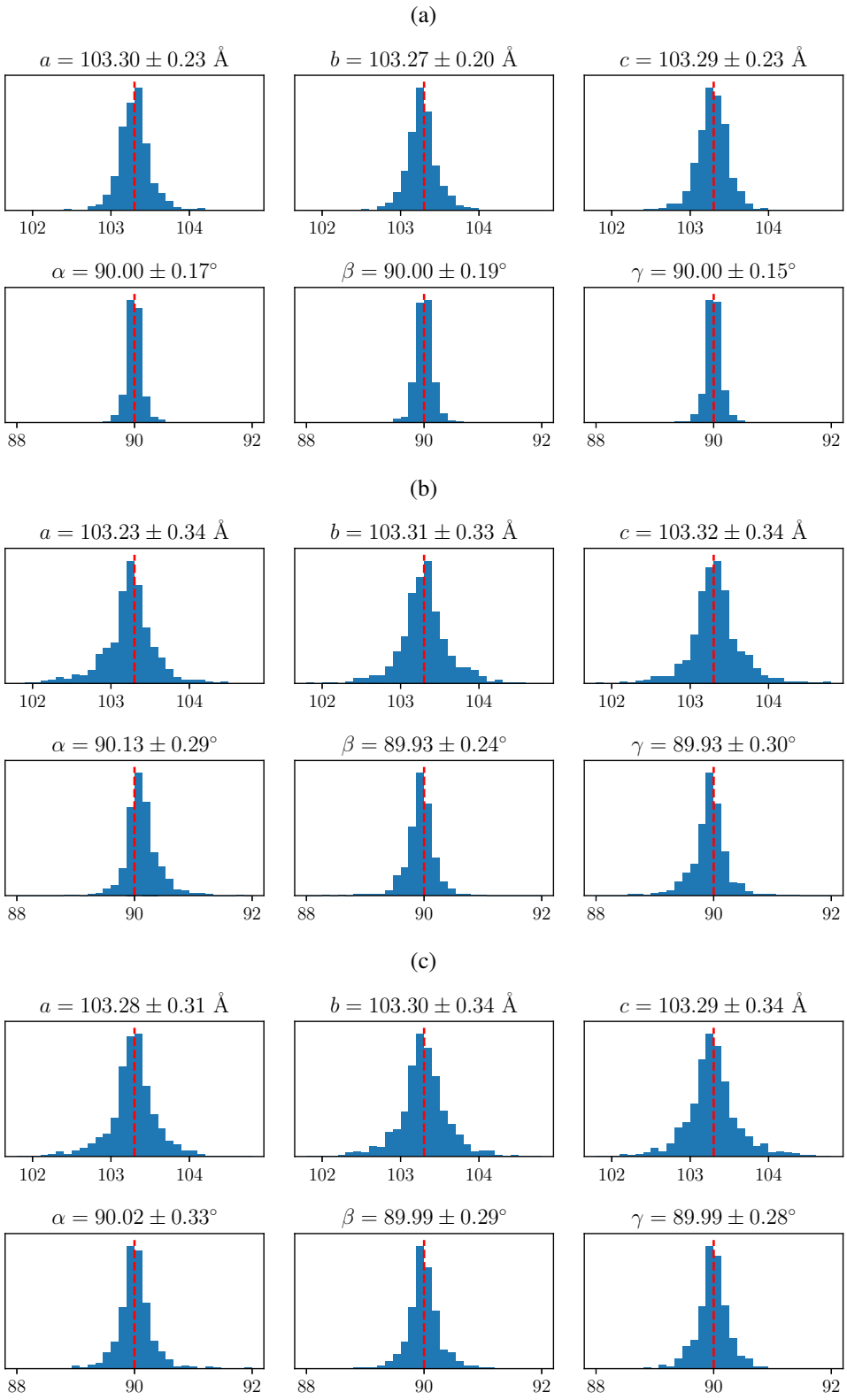


Figure 4.7: Distributions of the unit cell parameters obtained for experimental GV data with (a) *MOSFLM*, (b) *DirAx* and (c) *asdf*. The numbers above each plot give the average ± standard deviation of the parameter. Dashed red lines show the expected unit cell parameters: $a = b = c = 103.3$ Å and $\alpha = \beta = \gamma = 90°$.

Indexing algorithm	*MOSFLM*	*DirAx*	*asdf*
Indexed patterns, %	78.6	78.6	82.6
Run time, min:sec	27:50	23:42	22:44
Mean peak misalignment, pixels	0.45	0.47	0.48

Table 4.2: Comparison of execution time and indexing results of three algorithms indexing experimental GV data.

precise. However, *asdf*, similar to *MOSFLM*, gives more accurate estimations of the unit cell parameters compared to *DirAx*. This is suggested by the fact that the average values of a, b and c yielded by *asdf* are closer to each other and the average angles are closer to 90° - the result expected for the cubic lattice. Additionally, to compare the accuracy of the indexing results, the distances between the detected peaks and the corresponding predicted reflections were calculated for all indexed diffraction patterns. Their distributions for all three indexing algorithms are shown in Fig. 4.8 and the mean values are given in Table 4.2. As can be seen, *MOSFLM* on average predicts peaks slightly better than *asdf*. When compared to *DirAx*, *asdf* gives similar value of the mean misalignment, but due to higher number of indexed diffraction patterns predicts in total more peaks within approximately one pixel radius.

An alternative way to visualize the quality of both indexing and peak prediction is presented in Fig. 4.9. It shows the percentage of diffraction patterns indexed by each indexer, which contain a certain percentage of peaks predicted within the certain distance Δ_{max}. Despite the larger deviation in the unit cell parameters obtained by *asdf* and *DirAx*, Fig. 4.9 demonstrates that the accuracy of the peak prediction of all indexers is similar. Furthermore, because *asdf* was able to index more patterns, percentage of patterns which have the majority of peaks predicted with at least 2 pixels accuracy achieved with *asdf* is higher than with either of the other two indexers.

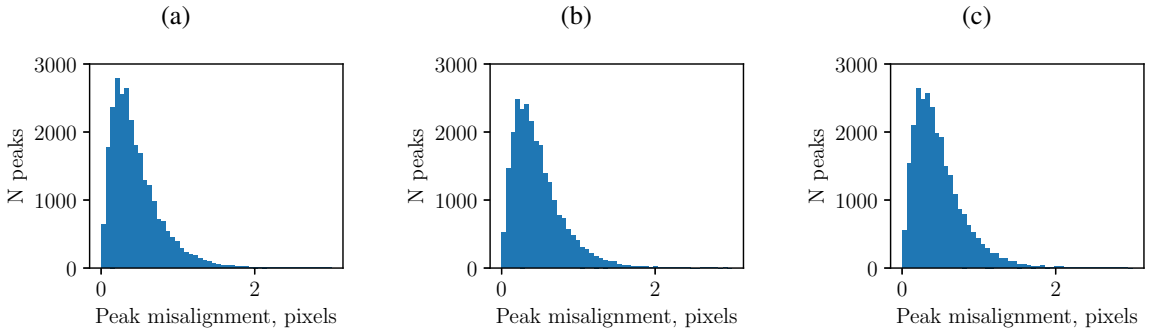


Figure 4.8: Distributions of distances between the observed peaks and corresponding reflections predicted using indexing solutions obtained with (a) *MOSFLM*, (b) *DirAx* and (c) *asdf* for experimental GV data.

4.2.3 Conclusion

Generally, the recommended practice is to use as many indexing algorithms as possible, in such way that if one fails the next one is attempted, achieving the maximum indexing fraction [88]. Since it was first added in *CrystFEL* version 0.6.1, *asdf* has been used in several serial crystallography experiments

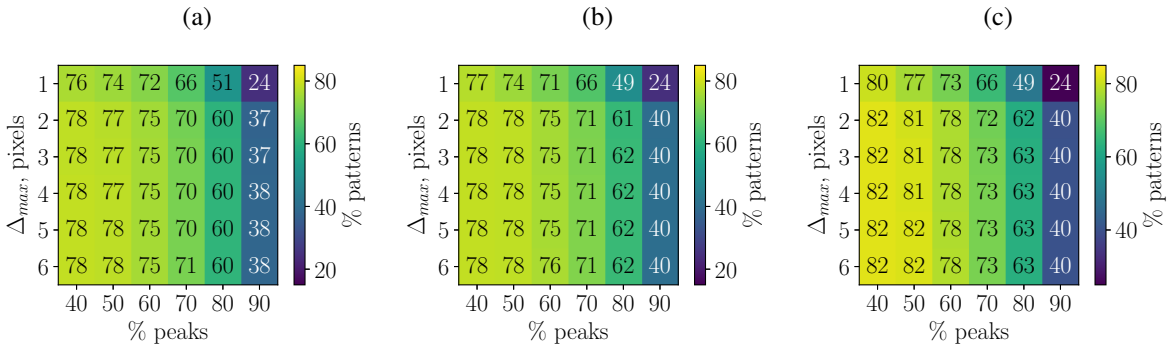


Figure 4.9: Percentage of GV diffraction patterns indexed with (a) *MOSFLM*, (b) *DirAx* and (c) *asdf*, where % peaks are predicted within the accuracy of Δ_{max} pixels.

in combination with other available indexers [53, 54, 90, 91]. It is currently the only built-in indexing algorithm in *CrystFEL* which can be used both with and without prior knowledge of the unit cell parameters. Additionally, due to its fast execution time, it has a potential to be used for real-time data processing to give fast feedback during the experiment.

CHAPTER 5

Analysis of FEL data

This chapter describes two serial crystallography datasets collected at the LCLS with the emphasis on the particular aspects of the data analysis which were used in addition to or expanding on the key processing steps explained in the previous chapter in order to improve the final data quality. My contributions to both experiments include the full analysis of SFX data from the raw images to the merged hkl-intensities.

5.1 Angiotensin II receptor AT_2R

The results presented in this section are published in "Structural basis for selectivity and diversity in angiotensin II receptors" by Zhang *et al.* 2017 [3]. The angiotensin II receptors AT_1R and AT_2R serve as key components of the renin–angiotensin–aldosterone system. AT_1R has a central role in the regulation of blood pressure, but the function of AT_2R is unclear and it has a variety of reported effects [93–95]. To identify the mechanisms that underlie the differences in function and ligand selectivity between these receptors, the structure of human AT_2R bound to an AT_2R-selective ligand was solved using SFX diffraction data, capturing the receptor in an active-like conformation. The results provide insights into the structural basis of the distinct functions of the angiotensin receptors, and may guide the design of new selective ligands.

5.1.1 Experiment at LCLS

Diffraction experiment was performed at the Coherent X-ray Imaging (CXI) instrument at LCLS. The LCLS was operated at a wavelength of 1.3 Å delivering individual X-ray pulses of 40 fs duration and approximately 10^{11} photons per pulse focused into a spot size of approximately 1.5 μm in diameter. Microcrystals of AT_2R with the average dimensions of $5\times2\times2$ μm^3 were delivered into the beam using the LCP injector with 50 μm nozzle. Diffraction patterns were recorded using the CSPAD detector. In total 2701530 images were collected and 175241 of them were identified as hits by *Cheetah*. The hits were then processed using *CrystFEL* (version 0.6.1+f5db71cc). Below I describe in detail three analysis steps which were essential to achieve the optimal data quality.

5.1.2 Refinement of detector geometry

As explained earlier, optimization of the detector geometry is extremely important as it is necessary for accurate peak prediction and integration. However, refinement of the relative positions and rotations of

individual detector segments with *geoptimiser* is not generally required for every experiment. Unless the arrangement of the detector panels was manipulated in some way between the experiments, previously determined detector geometry is usually sufficient. In such case the only parameters which have to be optimized are the beam center and sample-to-detector distance. The initial estimation of the detector center is obtained by constructing a virtual powder pattern from all detected peaks in a large set of collected diffraction patterns and manually aligning detector to the center of the resulting powder rings. The center is then more accurately determined during the prediction refinement procedure in *indexamajig*. As for the sample-to-detector distance, the best parameters to assess its accuracy are, for a crude estimate, indexing fraction and the agreement between the obtained and expected unit cell parameters and, for a more accurate estimate, the shape of the distributions of the obtained unit cell parameters.

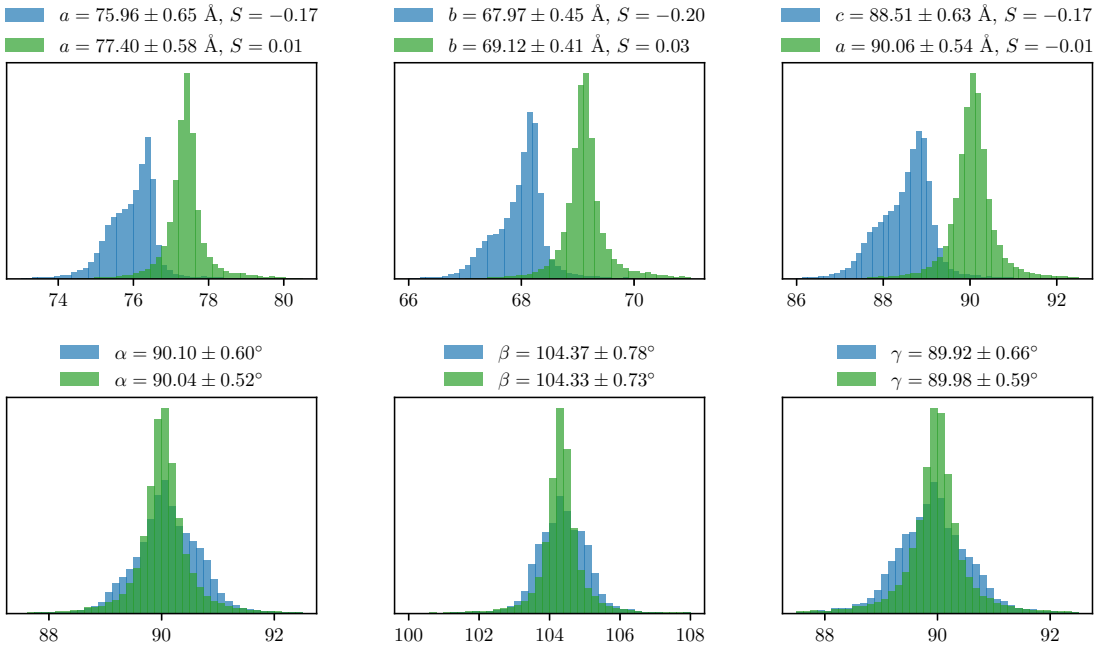


Figure 5.1: Distributions of the AT_2R unit cell parameters obtained before (blue) and after (green) detector distance optimization.

In this case well-optimized detector geometry was available from the previous experiment at the CXI station. Distributions of the unit cell parameters obtained during the initial indexing with *indexamajig* are shown in blue in Fig. 5.1. The unit cell parameters $a = 78.4$ Å, $b = 68.2$ Å, $c = 89.1$ Å, $\alpha = \gamma = 90°, \beta = 104.3°$, determined previously from the synchrotron diffraction data [3], were provided as an input in *indexamajig*. Despite the closeness of the obtained and expected unit cell, the resulting distributions are quite broad and asymmetric, which is usually an indication of the wrong detector distance. Thus, to optimize the detector distance, a sub-set of 2000 indexed images was randomly selected and indexed using different values of the detector distance varying in the range of ± 5 mm from the initial value of 151 mm provided by the facility. Then, to find the best estimate, two metrics were used: the width of the distributions or the standard deviation of the unit cell parameters and non-parametric skew as a measure of symmetry of the distributions defined as

$$S = \frac{\mu - m}{\sigma}, \tag{5.1}$$

where μ is the mean, m is the median and σ is the standard deviation of the parameter. The RMS of σ and S of the obtained lattice parameters a, b and c are plotted in Fig. 5.2 as functions of the detector distance. As can be seen, both of them reach their minimum around the same detector distance. Since the minimum in skewness is more sharp its value of 153.4 mm was used in further analysis. Resulting distributions of the unit cell parameters of the full dataset indexed with the optimized detector distance are shown in green Fig. 5.1, demonstrating clear improvement compared to the initial results.

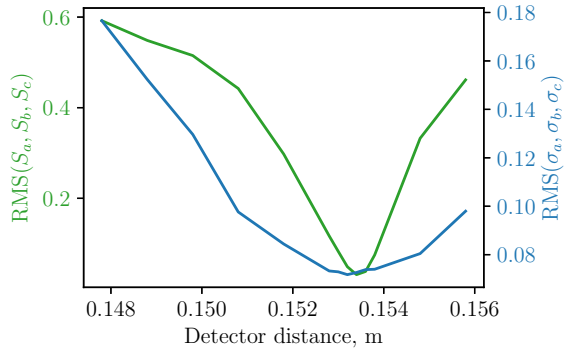


Figure 5.2: The root mean square of standard deviation and skew of the unit cell parameters a, b and c as a function of detector distance used for indexing.

5.1.3 Sorting of two crystal forms

Despite achieving a good agreement between the obtained and expected unit cell with the well-optimized detector geometry, using *MOSFLM*, *asdf*, *DirAx* and *XDS* indexing methods it was possible to index only about 23000 patterns out of 175000 hits. This indexing fraction is considerably lower compared to what was usually achieved in similar experiments. In order to investigate the reasons for such low indexing fraction, the data was indexed without providing *indexamajig* with the target unit cell, which revealed crystals with the orthorhombic lattice present in the same crystallization batch in addition to the previously known monoclinic lattice. The resulting unit cell parameters of two different crystal forms are given in Table 5.1. About 16000 diffraction patterns were indexed with the second orthorhombic unit cell.

Lattice	a, Å	b, Å	c, Å	α,°	β,°	γ,°
Monoclinic	77.4	69.1	90.1	90	104.3	90
Orthorhombic	70.3	78.8	93.4	90	90	90

Table 5.1: Unit cell parameters of two AT_2R crystal forms.

Since *MOSFLM* and *XDS* are the only two of the employed indexing methods which use the unit cell parameters provided as an input they can produce false-positive indexing solutions, e.g. incorrectly index the monoclinic crystal using the orthorhombic lattice parameters. Example of such pattern is shown in Fig. 5.3. To exclude these false-positive solutions, 715 diffraction patterns which were indexed as both monoclinic and orthorhombic crystals were isolated, processed separately without using the lattice information for indexing and, based on these indexing results, sorted back into the respective datasets (Fig. 5.4).

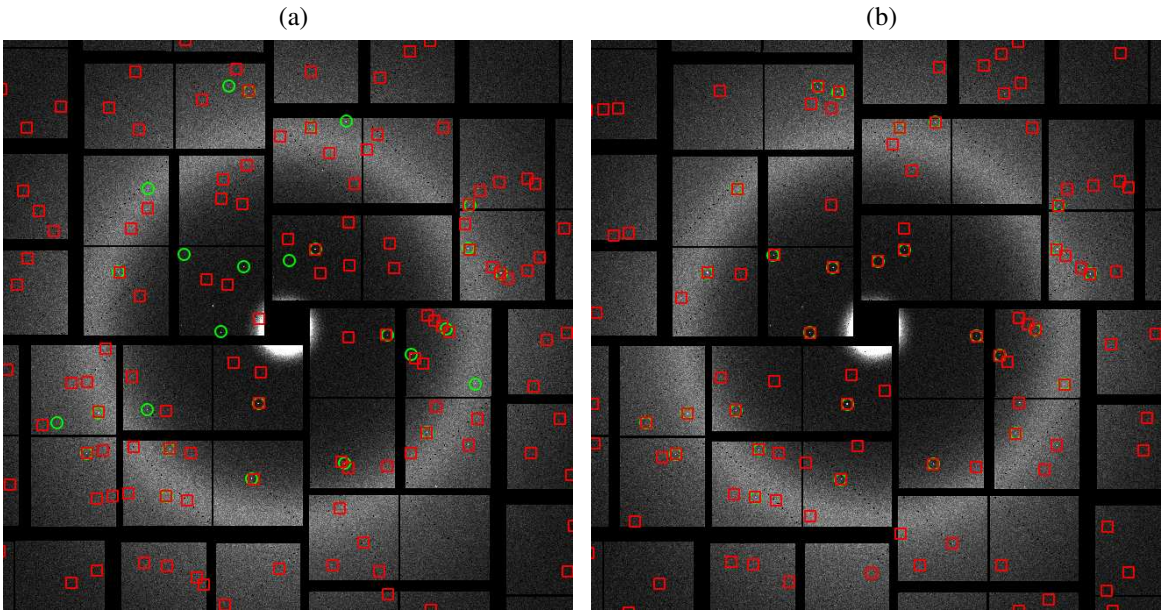


Figure 5.3: AT_2R diffraction pattern indexed by *MOSFLM* incorrectly using the orthorhombic unit cell (Fig. 5.3a) and correctly using the monoclinic unit cell (Fig. 5.3b). Found peaks and predicted reflections are shown as green circles and red squares respectively.

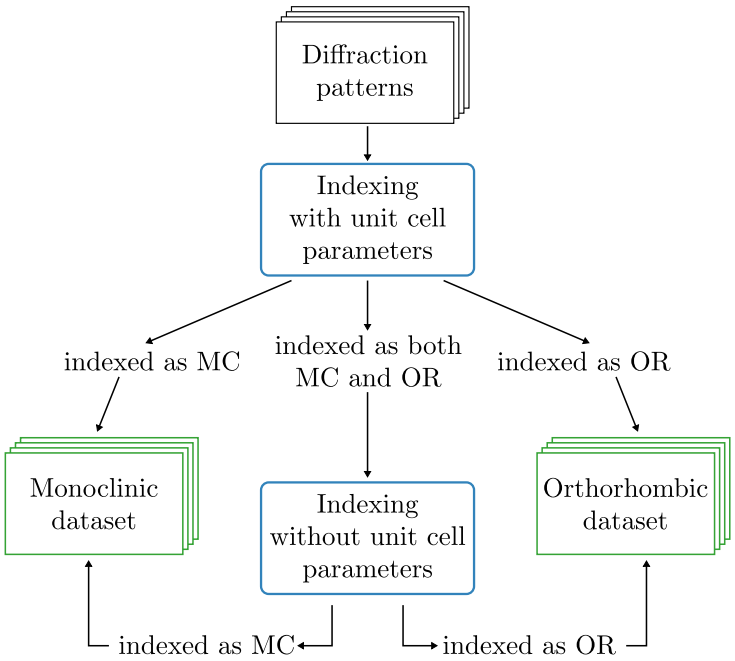


Figure 5.4: Sorting diffraction patterns of AT_2R crystals with monoclinic (MC) and orthorhombic (OR) unit cells into two respective datasets.

As a result, 38578 patterns were indexed: 22774 as monoclinic and 15804 as orthorhombic crystals, giving the total indexing fraction of 22%. This value is still relatively low, which is explained by the large fraction of weak hits with the low numbers of found peaks per pattern (Fig. 5.5a). Fig. 5.5b shows the dependence of the indexing rate on the number of found peaks for both crystal forms separately and combined: for the patterns with more than 30 found peaks indexing rate exceeds 80%, which is the expected value in such experiment.

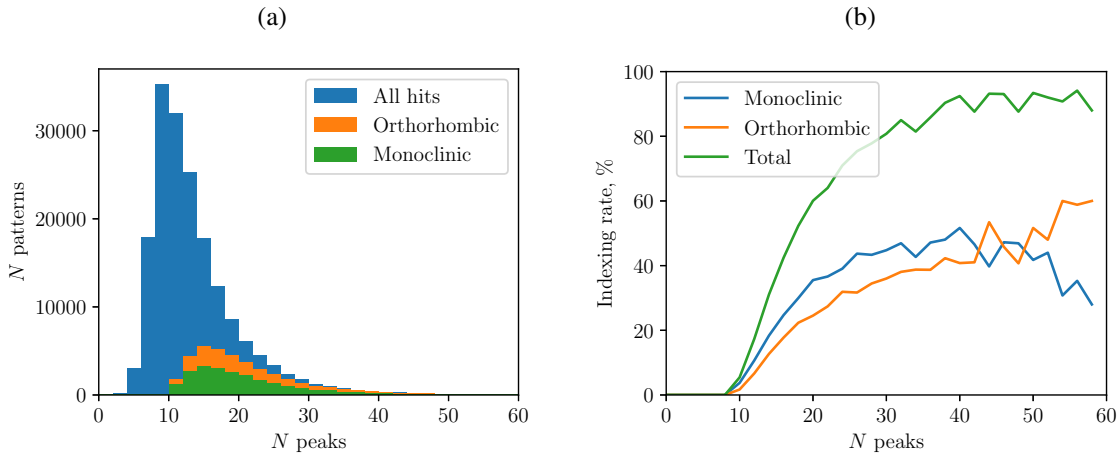


Figure 5.5: (a) Distribution of the number of peaks per pattern for all hits (blue), hits indexed as orthorhombic crystals (orange) and hits indexed as monoclinic crystals (green). (b) Indexing rate as a function of the number of peaks per pattern for monoclinic (blue), orthorhombic (orange) and both lattices combined (green).

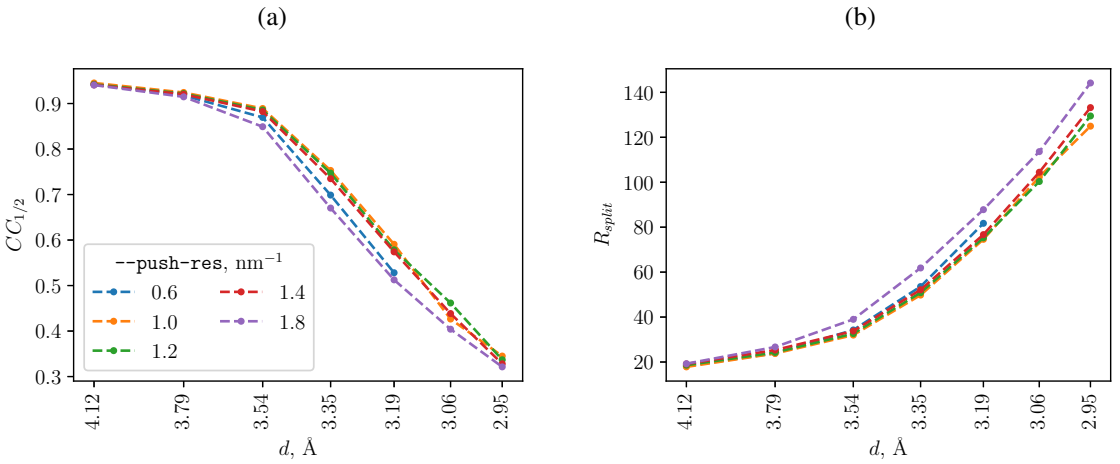


Figure 5.6: $CC_{1/2}$ and R_{split} as functions of resolution for different `--push-res` values.

5.1.4 Per-pattern resolution cut-off

Variations in the crystal size or illuminated crystal volume and crystal order as well as fluctuations in the FEL beam intensity lead to significant variations in the diffraction resolution of individual patterns. Accuracy of reflection prediction is limited by the resolution of the observed peaks as only their positions are used for indexing and refinement of the orientation matrix. Using the Monte Carlo approach, reflection intensity is integrated by averaging measurements from all patterns where the reflection was predicted. Including the measurements of the reflection intensities beyond the resolution limit of the individual crystals where prediction is inaccurate, i.e. adding noise to the calculation of average, reduces the final signal-to-noise ratio. On the other hand, prediction might still be valid in a certain resolution range above the resolution limit of the observed peaks. In this case applying strict cut-off and using only reflections below the resolution limit could lead to the loss of the meaningful intensities of the weak reflections not detected during the peak search. A compromise solution is to extend the resolution cut-off by a certain value above the estimated resolution of each diffraction pattern. This approach is implemented in

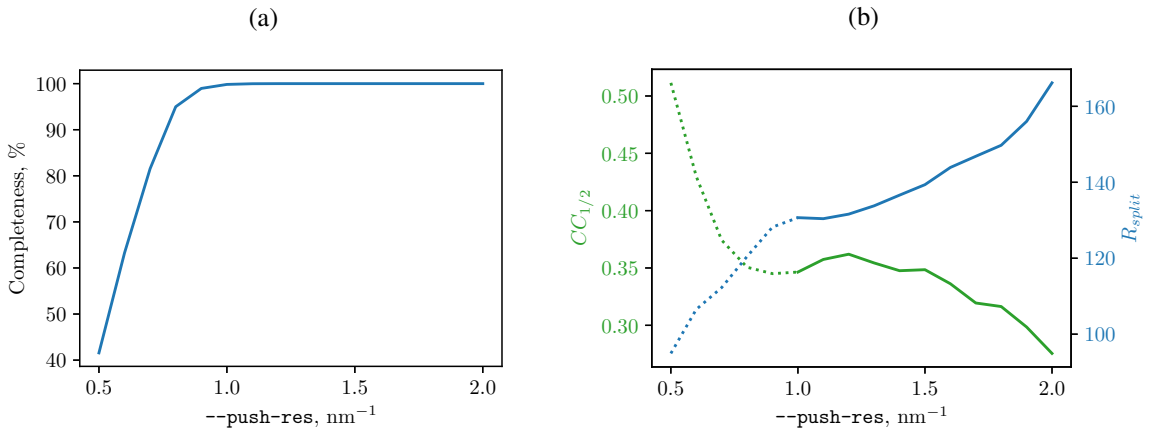


Figure 5.7: Completeness (a), $CC_{1/2}$ and R_{split} (b) in the highest resolution bin (2.8 - 3.0 Å) as functions of `--push-res`. Dotted and solid lines show the metrics where completeness is less and equal 100% respectively.

partialator as a `--push-res` option.

Fig. 5.6 shows the dependence of data quality metrics $CC_{1/2}$ and R_{split} on resolution for different `--push-res` values. Fig. 5.6 shows completeness, $CC_{1/2}$ and R_{split} in the highest resolution shell. As can be seen, after 100% completeness is reached in the highest resolution shell, both metrics reach their optimum values with the `--push-res` of 1.1 - 1.2 nm^{-1}, after which the data quality starts to drop. As said previously, this is an effect of adding noise to the average leading to lower signal-to-noise ratio and therefore worse data quality. Based on these results, resolution extension by 1.2 nm^{-1} was used to merge both datasets.

	Monoclinic	Orthorhombic
Space group	$P2_1$	$P2_122_1$
Unit cell parameters a, b, c, Å α, β, γ,°	77.4, 69.1, 90.1 90.0, 104.3, 90.0	70.3, 78.8, 93.4 90.0, 90.0, 90.0
N hits	175241	175241
N indexed	22774	15804
Resolution, Å	30-2.8 (2.9-2.8)	30-2.8 (2.9-2.8)
Completeness, %	100 (100)	100(100)
Multiplicity	61.6	85.5
$I/\sigma(I)$	4.06(0.8)	4.86(1.02)
CC^*	0.98(0.28)	0.99(0.42)
R_{split}	16.35(153)	14.8(115)
Wilson B-factor, Å^2	90.8	80.9
Reflections N_{work}/N_{free}	22906 / 1118	13269 / 691
R_{work}/R_{free}	0.227 / 0.256	0.241 / 0.262
RMS bonds, Å	0.010	0.009
RMS angles, °	0.90	0.92

Table 5.2: Data analysis and refinement parameters of the two resulting AT_2R datasets. Table from Zhang *et al.* 2017.

5.1.5 Results: two crystal structures of AT_2R

Resulting data analysis and structure refinement statistics of both crystal forms of AT_2R are presented in Table 5.2. Despite the overall lower number of merged patterns, due to the higher symmetry orthorhombic structure was more complete, and therefore was used for the description of the overall AT_2R structure by Zhang *et al.* 2017 [3]. Solution of the crystal structure of AT_2R in two different lattices provided important insights into its structure and function as they both showed non-canonical conformation of helix VIII (Fig. 5.8), suggesting that it is likely to be a genuine feature of AT_2R, rather than an artefact of crystallization.

Both angiotensin II receptors are important drug targets, since the blockade of AT_1R has anti-hypertensive effects, while the modulation of AT_2R could be useful for cardioprotection, neuropathic pain relief and the treatment of several other conditions. Designing molecules that selectively bind to a specific receptor type is often challenging, but can be crucial for different therapeutic purposes. Although the AT_1R and AT_2R ligands share common scaffolds, this study showed that ligand-binding pockets of these two receptors are markedly different, and these differences could be exploited for designing selective ligands. The AT_2R crystal structures determined in this study improve our understanding of the two types of the human angiotensin receptor and provide new insights into the structural basis for the binding and selectivity of small molecules of therapeutic significance.

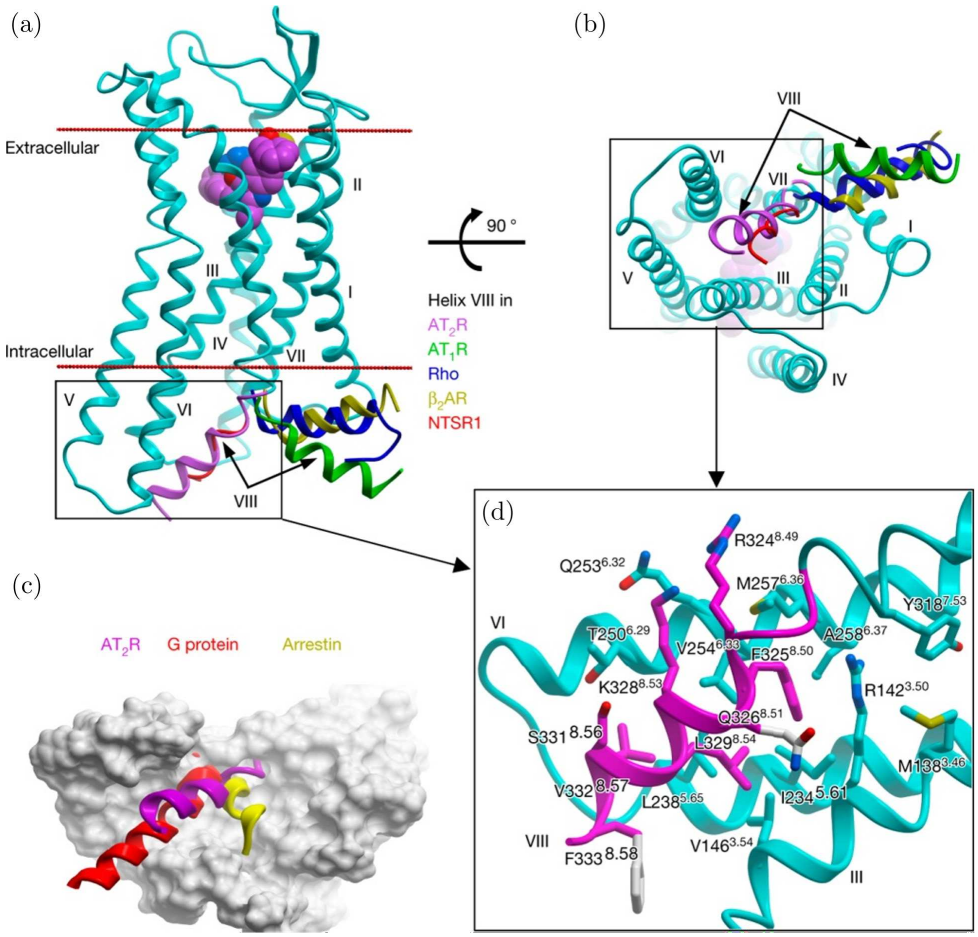


Figure 5.8: (a), (b) Varied positions of helix VIII in different GPCRs shown as cartoons. AT_2R is in cyan, with helix VIII coloured magenta, AT1R (PDB code 4YAY) in green, NTSR1 (PDB code 4GRV) in red, G-protein-bound βAR (PDB code 3SN6) in yellow, and arrestin-bound rhodopsin in blue (PDB code 4ZWJ), viewed from within the membrane (a) and from the intracellular side (b). (c) Shared interaction sites of AT2R 7TM domain (grey surface) for the helix VIII of AT_2R (magenta), the C terminus of G protein (red), and the finger loop of arrestin (yellow). (d) Interactions between helix VIII (magenta) and helices III, V and VI (cyan) in AT2R. Side chains of helix VIII not resolved in the crystal structure are shown by grey carbon atoms. Figure from Zhang *et al.* 2017.

5.2 Photosystem II

Photosynthesis, a process used by plants, algae and cyanobacteria to convert sunlight into chemical energy, takes place in two large protein complexes, photosystem I and II (PSI and PSII). PSI is responsible for production of high energy carriers ATP and NADPH using the light energy. PSII catalyses the light-driven water splitting process, providing the electrons for the photosynthesis to occur. Solving molecular structures of both complexes in all the states they go through under the influence of light is crucial for understanding of the whole photosynthesis process. This makes them one of the most important targets in serial crystallography. The structure of PSI was the first crystal structure solved with serial femtosecond crystallography [3]. Both PSI and PSII have since been studied extensively at the FELs, in particular in time-resolved serial crystallography experiments [96–100].

In addition to the particular importance of PSII structural dynamics, an interesting feature of PSII crystals is relatively strong diffuse scattering, i.e. scattering between the Bragg peaks caused by the predominantly translational disorder of large PSII molecules in the crystal. It has been shown that this diffuse scattering can be used to phase the diffraction patterns directly and obtain the structure of PSII molecule at the resolution beyond the limit of measurable Bragg peaks [101, 102]. The establishment of this method is another objective to do serial crystallography measurements of PSII.

This section describes the analysis of PSII diffraction data exploring in particular specific features introduced by the fixed-target sample delivery approach. The paper based on the results presented here is currently in preparation.

5.2.1 Fixed-target experiment at LCLS

Diffraction data was collected at the Macromolecular Femtosecond Crystallography (MFX) instrument at LCLS using the Roadrunner II goniometer for fixed-target sample delivery and the CSPAD detector. The Roadrunner II goniometer consists of a horizontal (x-axis) fast-scanning stage perpendicular to the incident beam (z-direction), mounted on a high-precision yz translation unit and a horizontal rotation axis. For data collection, a suspension of PSII crystals of 10-40 μm in size was pipetted onto the micro-patterned silicon chip mounted onto the scanning unit, and the excessive crystal growth solution was soaked off by blotting through the pores from the bottom side of the chip [62]. The chip was then scanned in a serpentine style by shooting the beam through the micropores in the chip at the LCLS operation rate of 120 Hz. Crystals larger than the pores tend to organize themselves according to the pore pattern during the blotting procedure, ensuring high hit rate [62, 65]. The effects of preferred orientation of the crystals on the chip are avoided by rotating the chip with respect to the beam. The Roadrunner II setup and the scanning procedure will be described in more detail in Section 6.4.

Roadrunner II chips used for the experiment had the dimensions of 33×12 mm and provided 67500 rectangular pores in a hexagonal pattern with a spacing of 50 μm. The pores were arranged into 15×5 compartments with a size of 1.5×1.5 mm and a membrane thickness of 10 μm, separated by a support frame with a width of 0.6 mm and a thickness of 300 μm. The pore size varied between the chips from 7 to 18 μm to match the size of the crystals.

During the data collection the chip was moved horizontally at a constant velocity of 6 mm/s to match the distance traveled between the subsequent X-ray pulses to the pore spacing. To ensure that X-rays hit the pores and to avoid hitting the support frame of the chip, the time structure of the LCLS pulses was

synchronized with the Roadrunner motion controller. The X-rays were blocked during the acceleration of the chip in the beginning of the line, passing through the support structure, deceleration in the end of the line and change of the scanning direction [65]. In the meantime, the detector was constantly saving frames at a rate of 120 Hz regardless of whether the X-ray pulse was actually blocked or not. As a result, the number of collected frames per chip was about 2 times higher than the number of pores. In total, diffraction data from 33 chips was collected amounting to 2,467,818 collected images, 598,886 of which were identified as hits by *Cheetah*, giving the hit fraction of 24% or, on average, 29 crystal hits per second.

As mentioned earlier, crystals placed on the membrane for fixed-target sample delivery have to be protected from dehydration. In our approach, to avoid degradation of crystals, the chip is kept in a chamber constantly flushed with humidified helium (Fig. 6.13). The advantage of this method is that it doesn't introduce any additional material into the beam resulting in lower background compared to the alternative approach, where the chip is sealed with Mylar or Kapton foil. On the other hand, since the humidity chamber is open from one side and the humidified gas flows from one side of the chamber to the other, the crystals on the chip may experience a gradient of humidity. Due to the large size and the irregular shape of the protein molecules and the absence of strong intermolecular forces substantial space between the protein molecules in the crystal, usually around 50%, is filled with solvent. Dehydration of protein crystals leads to the loss of water which is accompamied by shrinkage of the crystals [103]. As a consequence, PSII crystals in our experiment demonstrate significant variations in the unit cell parameters depending on their position on the chip.

I present here the detailed analysis of the deviations in the unit cell size and diffraction quality of PSII crystals, caused by the presence of humidity gradient in the measurement chamber, providing a comparison with the diffraction data collected from similarly prepared PSII crystals measured in their native solution in a liquid jet.

5.2.2 Variations in the unit cell parameters

Diffraction patterns were processed with *CrystFEL* (version 0.6.2). To adjust for the larger than usual variations of the unit cell the allowed tolerance between the expected and obtained unit cell parameters was set to 15%. As a result, 381,122 patterns (64% of hits) from 33 chips were successfully indexed.

When compared to PSII crystals prepared using the same protocol and measured in the native solution in a liquid jet [101], the unit cell parameters obtained in this experiment demonstrate similar mean values but much broader distributions (Fig. 5.9). Notably, the resulting distributions, in particular the distribution of parameter a, are clearly bimodal. The distribution of the unit cell volumes is, therefore, also bimodal and shows approximately 5 times larger standard deviation compared to the unit cell volumes of crystals measured in a liquid jet (Fig. 5.10). Since the chip stays within the few micrometre depth-of-focus of the in-line microscope during the data collection, the variations in the lattice constants cannot be attribited to the changes in the sample-to-detector distance.

5.2.3 Visualization of the unit cell distribution on the fixed-target chip

To confirm that the variations in the unit cell are indeed related to the humidity variations in the measurement chamber rather than being an inherent property of PSII crystals or an artifact of crystallization, the unit cell volume was mapped back to the position of the crystals on the chip. An example spatial distribution of the unit cell volume for one of the measured chips is shown in Fig. 5.11a. As can be seen,

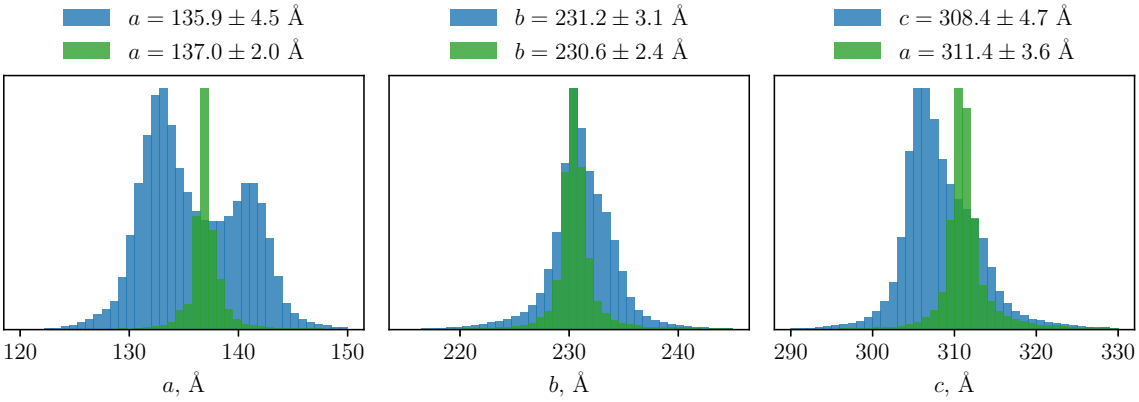


Figure 5.9: Unit cell parameters distributions of PSII crystals measured using Roadrunner II fixed-target setup (blue) and liquid jet [101] (green) for sample delivery.

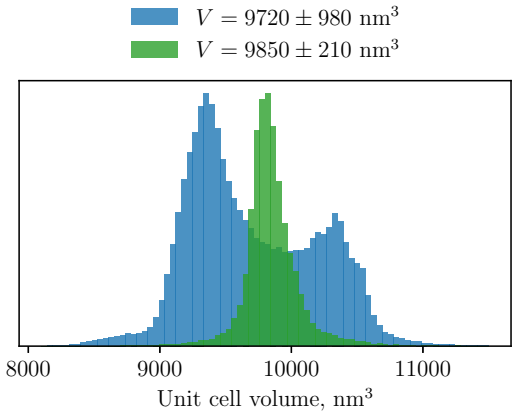


Figure 5.10: Unit cell volume distributions of PSII crystals measured using Roadrunner II fixed-target setup (blue) and liquid jet [101] for sample delivery.

the unit cell volume decreases from 10800 nm^3 in the top-left corner to 9000 nm^3 in the bottom-right corner of the chip. Fig. 6.13c and 6.13d, showing the start and end positions of the scan, illustrate how this variation is caused by an uneven humidity inside the chamber. Even though every crystal is measured in the same location within the chamber (the X-ray beam position), during the entire scan and during alignment the chip experiences a gradient of humidity due to the flow of humid air from one side of the chamber to the other.

Due to such large deviation in the unit cell, we also observe systematic changes in the diffraction data quality. These changes are demonstrated in Fig. 5.11b, which shows the distribution of the diffraction resolution limit of individual crystals as a function of the position on the chip. As described in Section 4.1.3, individual resolution limit for each crystal is estimated in *CrystFEL* by taking the 98th percentile of the scattering angles of all diffraction spots accounted for by the found lattice. It can be seen by comparing Fig. 5.11a and 5.11b, that the crystals in the middle area of the chip with the unit cell volume around 9500 nm^3 demonstrate the highest resolution of about 4 Å, while the crystal in the top-left and bottom-right corners with the larger and smaller unit cell volumes, respectively, demonstrate significantly lower

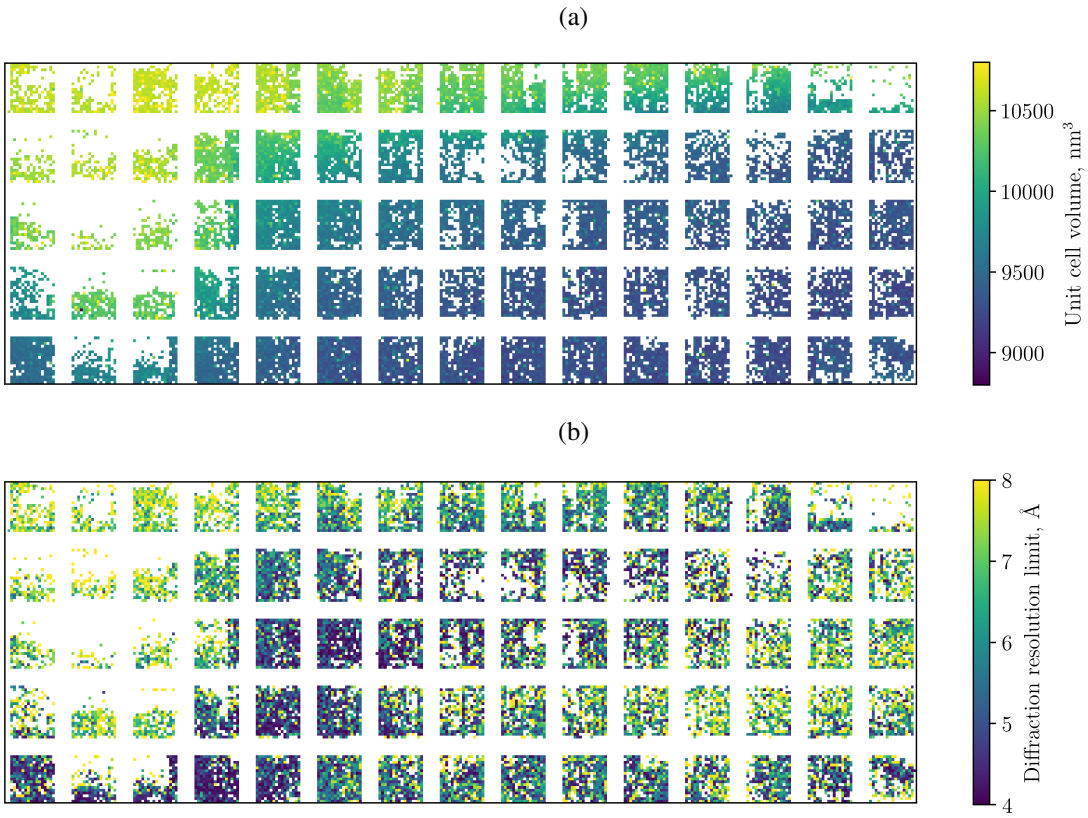


Figure 5.11: Spatial distribution of (a) unit cell volume and (b) diffraction resolution limit of PSII crystals, averaged in 2×2 bins, on the chip with 15×5 compartments, each 1.5×1.5 mm in size.

resolution, down to only 8 Å.

5.2.4 Influence of the humidity variation on the data quality

The relation between the unit cell size and the diffraction quality is further illustrated in Fig. 5.12, which shows the number of indexed diffraction patterns from all 33 chips as a function of the unit cell volume and diffraction resolution limit. The resulting distribution contains three distinct populations: one with the unit cell volume centered around 9400 nm^3 and the diffraction resolution of individual crystals ranged between 3.7 and 5.8 Å (marked as '1'), and two low-resolution ones, with the slightly smaller average unit cell volume of about 9300 nm^3 (marked as '2') and much larger average unit cell volume of 10300 nm^3 (marked as '3'). We assume that the low-resolution populations correspond, respectively, to severely dehydrated and over-hydrated (or swollen) crystals. It can also be noted, that the unit cell volume of the best diffracting crystals from population '1' is significantly smaller compared to crystals measured in the native solution (shown in green in Fig. 5.10), which suggests that the moderate dehydration of PSII crystals may actually improve diffraction quality. Individual diffraction resolution of the crystals in the liquid jet experiment varied between 4.7 and 20 Å and had the average value of 7.7 Å. However, the estimated resolution of individual crystals strongly depends on the peak-finding parameters and should not be compared directly between the two experiments.

The differences between the three observed populations of PSII crystals were investigated by selecting three sub-datasets, shown by red lines in Fig. 5.12. The subsets, each containing 40,000 indexed diffraction

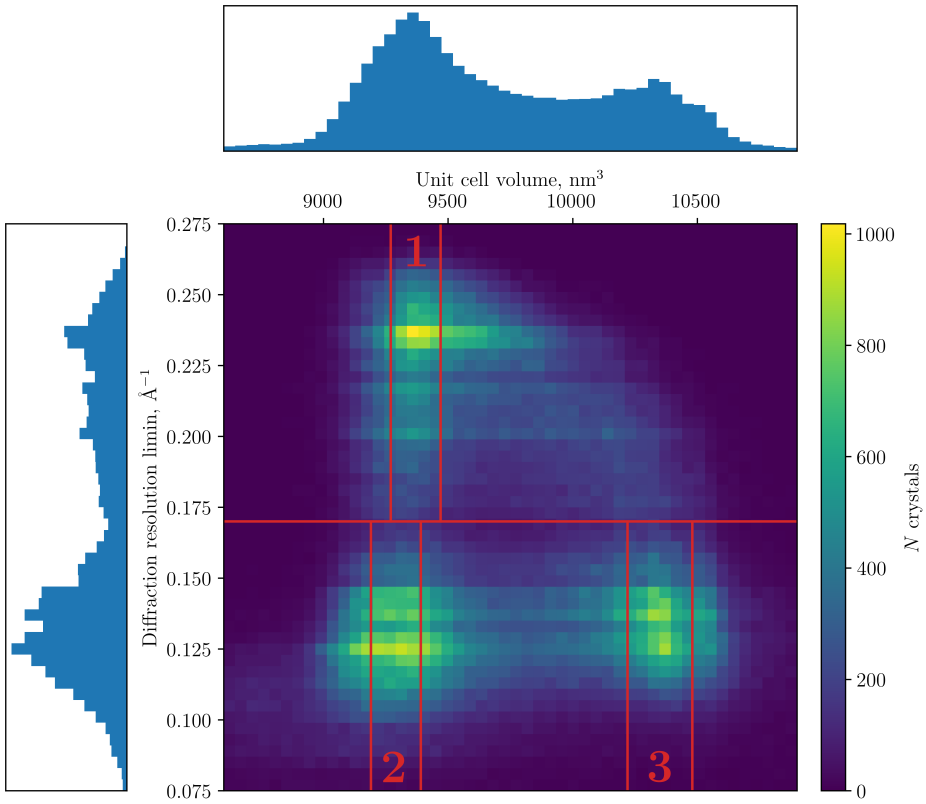


Figure 5.12: Number of indexed PSII crystals as a function of unit cell volume and diffraction resolution limit.

patterns around the maxima in the unit cell volume distributions of the corresponding population, were merged separately with *partialator*. Fig. 5.13 shows the dependence of the data consistency metrics $CC_{1/2}$ and R_{split} on resolution for the three subsets, confirming again that the crystals from population '1' demonstrate much higher diffraction quality.

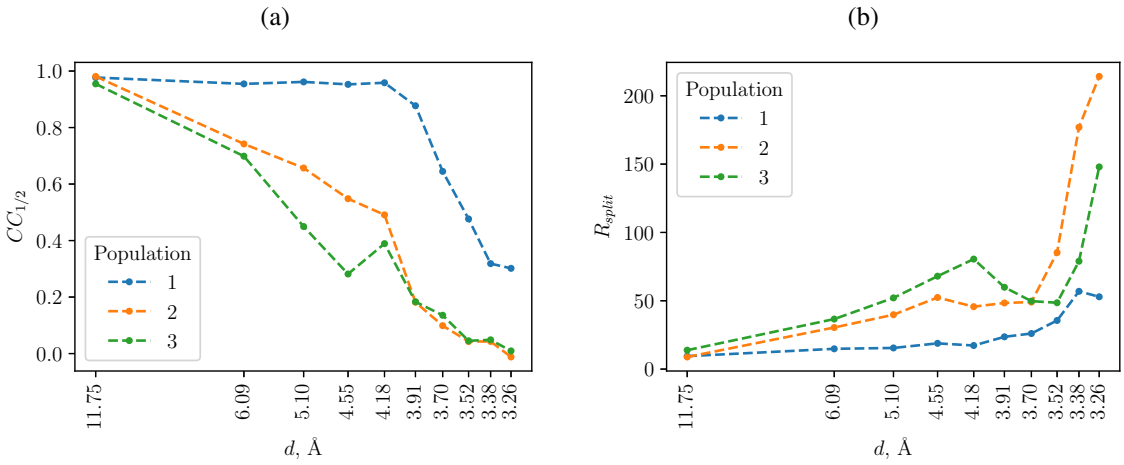


Figure 5.13: Dependence of (a) $CC_{1/2}$ and (b) R_{split} on the resolution for three sub-datasets consisting of 40,000 diffraction patterns from each population in Fig. 5.12.

To further investigate variations in the diffraction data quality and explore the influence of the humidity

variations on the molecular structures and packing, population '1' consisting of 143,187 best diffracting crystals was split into 14 sub-datasets binned by the unit cell volume, each containing 10,000 diffraction patterns, which were then merged individually. Fig. 5.14a shows the unit cell volume limits for each sub-dataset and Fig. 5.14b shows the dependence of the average diffraction resolution of the sub-dataset on the unit cell volume. The structures from these datasets were then solved by molecular replacement using the PDB structure 4PBU as a starting model [98]. The resulting refinement R-factors (Eqn. 2.34) are plotted in Fig. 5.14b, showing that the crystals with the unit cell volumes around 9100-9600 nm^3 give the best diffraction quality. The structures refined from these 14 sub-datasets were then used to analyze changes in packing and crystal contacts between the PSII molecules in different humidity conditions.

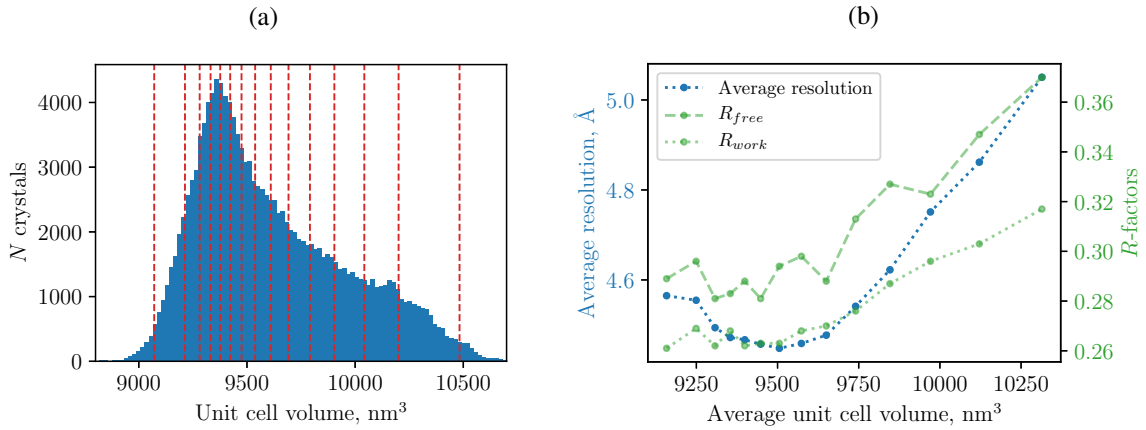


Figure 5.14: (a) Unit cell volume distribution of PSII crystals from population '1', with the diffraction resolution better than 5.8 Å. Dashed red lines show where the dataset was split into 14 sub-datasets. (b) Dependence of the average diffraction resolution (blue) and the refinement R-factors (green) on the average unit cell volume of the resulting sub-datasets.

Finally, to compare the quality of the structures obtained from PSII crystals in their native solution in the liquid jet to the structure of dehydrated PSII crystals on the fixed-target chip, 25,585 diffraction patterns (the exact same number as in the liquid jet experiment) from population '1' were merged and used to refine the structure using 3.5 Å resolution cut-off. The resulting statistics and refinement parameters are compared in Table 5.3. Despite that the structure from the fixed-target data was refined to 1 Å higher resolution the refinement R-factors in both cases are similar: R_{work}/R_{free} = 24.8 / 27.2 in the case of the liquid jet and 24.9 / 27.0 in the case of fixed-target experiment. This observation confirms that moderate dehydration of PSII crystals which can be achieved using Roadrunner fixed-target setup for sample delivery can indeed improve the quality of the resulting structure.

5.2.5 Discussion

Although the large variations in the unit cell parameters observed in this experiment is obviously a drawback of the fixed-target method which has since been addressed by improving the design of the humidity chamber, the analysis above also reveals several advantages of the method and opens new opportunities.

First, it shows that given the possibility to adjust the humidity in the chamber surrounding the fixed-target chip it should be possible to controllably dehydrate the crystals and by that obtain higher resolution structure, which can not be achieved using liquid jets. Second, solving the structure from data with

Experiment	Liquid jet [101]	Fixed-target
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters a, b, c, Å α, β, γ,°	137.0, 230.6, 311.4 90, 90, 90	133.1 230.3 305.8 90, 90, 90
Average unit cell volume, nm^3	9840	9370
N indexed	25,585	25,585[a]
Resolution, Å	30.0 - 4.5 (4.62 - 4.5)	40.7 - 3.5 (3.625 - 3.5)
Completeness, %	99.9	99.9
$I/\sigma(I)$	5.01 (1.49)	5.21(3.02)
CC^*	0.99 (0.94)	0.99 (0.75)
R_{split}	11.48 (46.16)	18.3 (38.7)
Wilson B-factor, Å^2	191.6	77.5
N reflections	55,609	118,803
R_{work}/R_{free}	24.8 / 27.2 (34.9 / 36.2)	24.9 / 27.0 (34.9 / 39.8)
RMS bonds, Å	0.005	0.016
RMS angles, °	1.40	1.25

[a] Selected around the maximum in the unit cell volume distribution in population '1' to match the total number of patterns to the liquid jet experiment.

Table 5.3: Data analysis and refinement parameters of two PSII datasets collected using liquid jet [101] and fixed-target sample delivery.

variable unit cells can give information about flexibility of different regions of the protein which might provide important insights into its function. Additionally, the positional differences between the molecules in the 14 datasets described above hint on regions of packing flexibility, which may help better model the type of disorder described by Ayyer *et al.* [101]. That in turn might enable *ab initio* phasing using the diffuse scattering [102].

Finally, even Bragg data from several unit cell variants might be used for solving the phase problem *ab initio*. Since Bragg diffraction data almost always under-samples the molecular transform [104], additional information is needed for uniquely determining the phases. For this purpose, the most common macromolecular phasing methods utilise additional information from other, but similar molecules or from isomorphic crystals of molecular derivates (Section 2.2). Yet, non-isomorphic crystal variants of the very same molecule, like the data described here, can also provide the necessary additional information, at least in principle. Already in 1952 Bragg and Perutz noticed that hemoglobin crystals with different solvent contents might be employed for phasing [105]. Even though their attempts were not successful, using modern computational methods it may be possible to implement the approach of phasing via multiple crystal forms.

5.3 Conclusion

Serial crystallography is a relatively new technique and almost every experiment is different in some way from the previous ones. New experiments often require development of new data analysis approaches. In this chapter I described the analysis of two different SFX experiments highlighting their individual aspects which required implementation of several new analysis steps. Some of them were only applied in this

particular case: for example, sorting of two crystal forms of AT_2R, as having two different crystal forms in one crystallization batch is relatively rare. The others, such as detector distance optimization, are a part of common practice and essential for every experiment. The tools developed here for the visualization and assessment of the unit cell parameter variations on the chip were particularly useful in all subsequent fixed-target experiments using Roadrunner setup, including the experiment described in the next chapter. Following the analysis presented here, the design of the humidity chamber was significantly improved to allow for a more stable humidity environment as will be shown in Section 6.4.

CHAPTER 6

Pink-beam serial crystallography

6.1 Motivation

Conventional X-ray protein crystallography diffraction experiments are performed with monochromatic radiation and exposure times of 10 ms and longer using photon-counting detectors, such as PILATUS [41] or EIGER [106]. Protein crystallography beamlines at 3rd generation synchrotrons provide about 10^{13} monochromatic photons per second with an energy of 12 keV and an energy spread of $\Delta E/E = 10^{-4}$, which corresponds to 10^{11} photons incident to the sample per image and exposure period. Assuming a beamsize of 10×10 μm^2 and a protein crystal with dimensions of about 50 μm this corresponds to a dose of ∼500 kGy delivered to the crystal per 10 ms exposure. With a dose limit of 30 MGy for data collection at cryogenic temperature this translates into 60 diffraction images, which can be recorded before the crystal is destroyed by the X-ray radiation. In case of room temperature data collection with a much smaller dose limit around 300 kGy only about 1 image with an attenuated beam can be collected per crystal.

By removing the requirement to obtain high signal-to-noise data from individual crystals, and by reducing dose by spreading the exposure over many crystals, serial crystallography also becomes a compelling method at synchrotron sources. Using available beamlines, several such experiments have been conducted using monochromatic radiation [10, 40, 107]. To obtain sufficient diffraction signal from small crystals, exposure times used in these experiments are typically in the few milliseconds range. Using a small bandwidth of only 0.01%, serial experiments with monochromatic synchrotron radiation require at least ten thousand of still diffraction images to obtain a complete high-quality dataset [40].

Crystallography with a polychromatic X-ray beam, also called pink beam, has been for a long time a standard technique for time-resolved measurements (Section 3.4) [27, 31, 32, 35, 108, 109]. The usage of a polychromatic X-ray beam allows collecting data much faster, as the number of photons is about 2 orders of magnitude higher compared to monochromatic radiation and reaches about 10^{15} ph/sec. By using the full intensity of a certain undulator harmonic with an energy spread of $\Delta E/E \simeq 0.05$ it is possible to realize much shorter exposure times down to 100 ps. Besides, the large number of photons spread over a large energy range leads to many more Bragg reflections from a crystal in a certain orientation, and also solves the partiality problem, which is the major source of errors in serial crystallography (Section 3.3.1). As a result, the measurement of much fewer crystal orientations is required compared to the monochromatic radiation to obtain a complete dataset. For serial crystallography experiments with the data collection times ranging from 10 minutes in a few FEL experiments [53, 65] to several hours for

the synchrotron experiments [10, 40, 107], usage of the pink beam should allow such experiments to be performed much faster, with much smaller amounts of sample and in a more competitive way compared to conventional single crystal X-ray diffraction experiments where data collection typically takes below 2 minutes for a complete dataset [22].

Polychromatic X-ray beams are available at a few synchrotron beamlines, including beamline ID09 at the European synchrotron radiation facility (ESRF) in France and the BioCARS instrument at the Advanced Photon Source (APS) in the US.

6.2 Challenges

Compared to experiments with monochromatic X-rays, to successfully perform serial crystallography experiment with a pink beam several challenges must be addressed.

Experimental challenges

The high X-ray intensities require dedicated hardware with active cooling of some components. This is particularly severe for the sections with a focused X-ray beam around the sample where special care has to be taken not to damage the equipment including the detector. A high heat load chopper can significantly reduce the thermal load on the downstream components.

Due to the large number of photons in a Bragg spot arriving at the detector in a very short time (down to 100 ps) current photon-counting detectors are not applicable. CCD detectors with their much lower performance regarding noise levels and readout times, have remained standard detectors at current polychromatic synchrotron beamlines. Recent developments of integrating detectors based on direct conversion for experiments at FELs, such as CSPAD [42], AGIPD [43] and JUNGFRAU [44], have also opened new opportunities for experiments using polychromatic radiation as they can handle the high photon flux occurring in combination with very high framing rates.

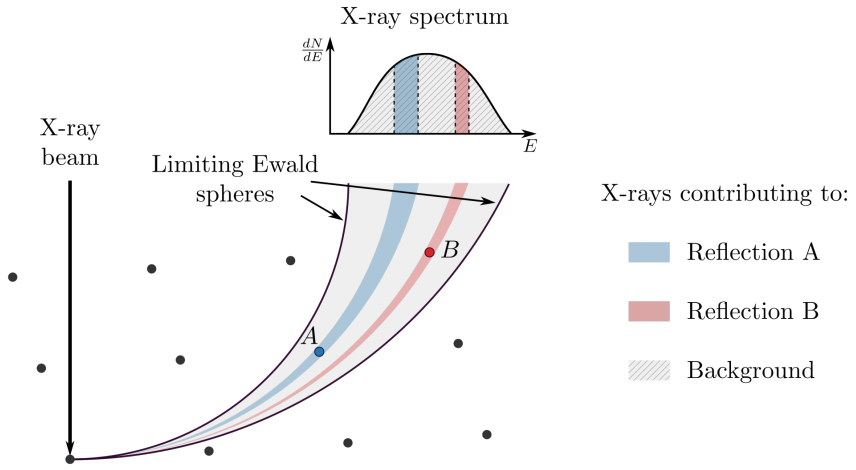


Figure 6.1: In crystallography experiment with polychromatic X-rays only fraction of the X-ray spectrum contributes to each Bragg peak while the whole spectrum contributes to the background, leading to lower signal-to-background ratio compared to monochromatic experiments.

While the larger bandwidth increases the number of Bragg reflections in a single still diffraction pattern, only the small fraction of the spectrum contributes to each reflection (Fig. 6.1). The background scattering however is produced by the whole spectral bandwidth. As a consequence, diffraction experiments with polychromatic X-rays result in a larger number of much weaker Bragg reflections compared to when using monochromatic radiation. This makes the success of the method extremely sensitive to the achievable background levels, which needs to be carefully minimized, in particular when dealing with very small crystals.

Recently we demonstrated pink-beam serial crystallography measurements at the BioCARS beamline at APS using a fixed-target setup optimized for achieving extremely low background levels [2]. Using a full harmonic of the undulator spectrum with a bandwidth of 5.7% (FWHM) and exposure times of 100 ps only, still diffraction patterns from about 50 crystals were sufficient to obtain a high-quality dataset.

Data analysis challenges

In contrast to previous serial crystallography experiments using radiation of narrower bandwidth, processing of this data turned out to be challenging, with only 13% of patterns being amenable to indexing using available algorithms. This difficulty was attributed to the low energy tail of the polychromatic beam, which extends to photon energies 20% lower than the peak energy, giving rise to the large number of Bragg spots observed in a diffraction pattern. Automatic indexing algorithms used for monochromatic radiation are in general not suitable for processing such patterns as it is impossible to distinguish which wavelength of the incident polychromatic beam diffracted into a particular spot and calculate its reciprocal-space coordinates (Section 4.1.3). Available pink-beam indexing algorithms, such as currently most widely used software Precognition [110], rely on finding ellipses in the diffraction spots patterns and often fail when small or weakly-diffracting crystals are used, which is typically the case in serial crystallography experiments. Furthermore, they require prior knowledge of the unit cell parameters, which in cases when they are unknown have to be determined using monochromatic radiation.

Large bandwidth also leads to a significant fraction of spatially overlapping reflections, in particular in the case of crystals with high mosaicity [111, 112]. This effect is illustrated in Fig. 6.2. Deconvolution of intensities of the overlapping reflections, although possible [113, 114], is not included in the available pink-beam diffraction processing software. Besides, existing implementations would not be applicable in the case of serial crystallography, as they rely on the diffraction measurements from the same crystal in different orientations defined by controlled rotation. As a result, the dataset of structure factors obtained by Laue diffraction is never full, with a typical completeness of 60-70% [2].

The algorithms, such as intensity integration based on peak profile fitting, have been developed for image plate and CCD detectors with large point spread functions and are also not compatible with handling new formats from pixelated detectors. Data processing of polychromatic diffraction still requires substantial manual input and cannot be run in an automated fashion [111].

Due to all these limitations most of the polychromatic diffraction experiments have been performed by collecting data from a small number of large crystals, which were typically individually mounted in glass capillaries. Even though it was possible to use traditional data analysis software in the case described by Meents *et al.* 2017 [2], in order to fully establish the method of pink-beam serial crystallography,

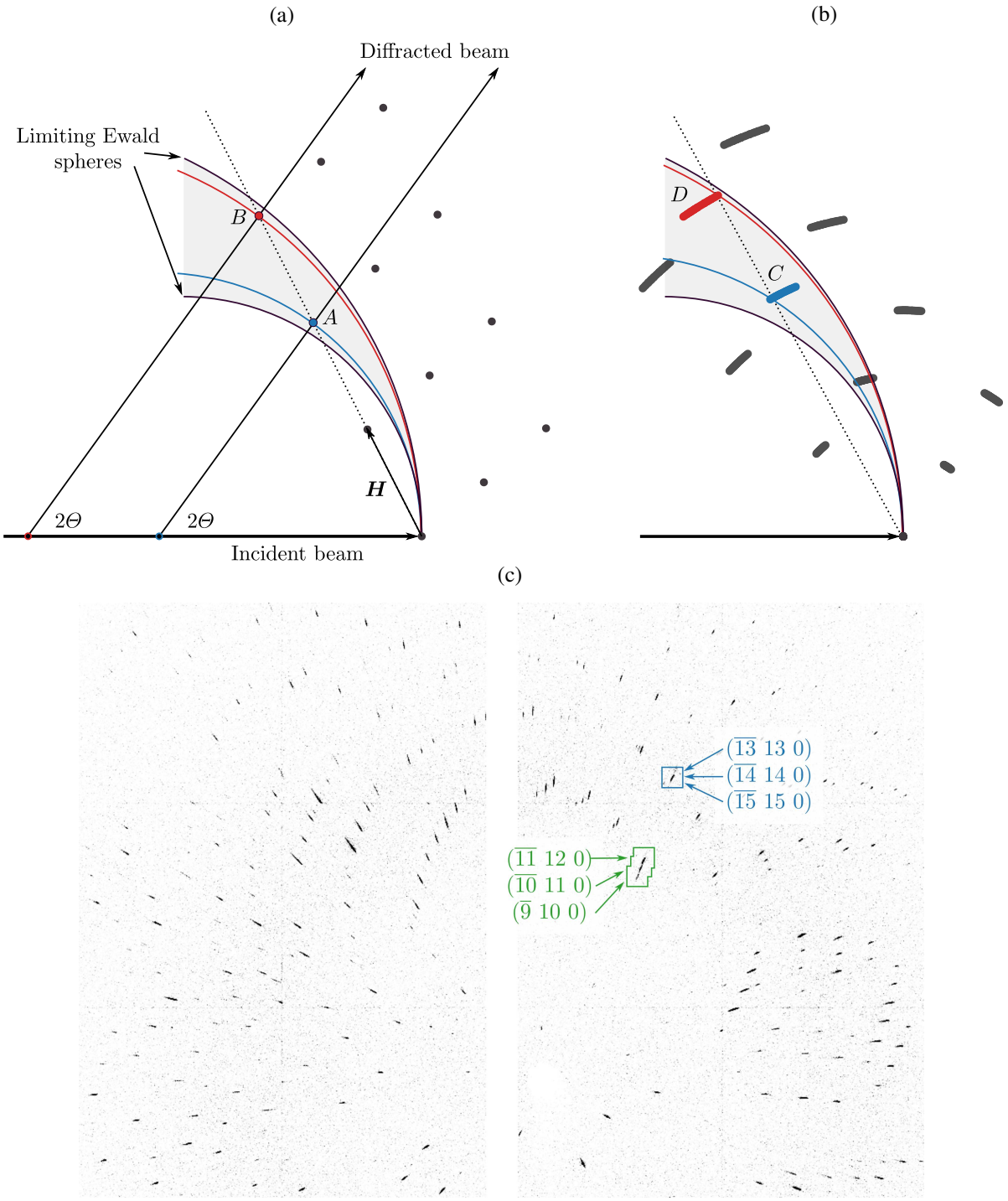


Figure 6.2: Illustration of overlapping reflections in pink-beam crystallography. Reflections lying on the line going through the origin of the reciprocal space diffract in the same angle.(a) Reflections A and B which belong to the same reciprocal lattice direction $\{n\boldsymbol{H},\ n \in \mathbb{Z}^+\}$ and lie between the limiting Ewald spheres will both diffract in the same angle 2Θ and therefore appear on the same spot on the detector. (b) In the case of mosaic crystal, reciprocal lattice points spread tangentially which leads to radial elongation of the diffraction spots. As a result, even reflections which don't belong to same reciprocal lattice direction, such as C and D, may intersect the line going through the origin of reciprocal space and overlap on the detector. (c) Example polychromatic diffraction pattern from a mosaic crystal featuring overlapping diffraction peaks.

the dedicated automatic data processing software needs to be developed, capable of handling pink-beam diffraction data collected from small crystals with the fast modern detectors.

It has been recently reported by Martin-Garcia *et al.* 2019 that usable intensities can be obtained from sparse pink beam diffraction data using the standard pipeline for monochromatic data processing with *CrystFEL* [115]. In the described case, only 10% of the diffraction patterns could be indexed and the valuable high-resolution diffraction spots could not be indexed and were not included in the structure refinement. The limitations of this approach are examined in detail in the following section. One way to overcome them is to narrow the spectrum to approximately 3%, cutting the low energy tail. This can be achieved at an undulator beamline using the multilayer monochromator (Section 3.2). This choice of bandwidth also produces a reasonably high proportion of Bragg peaks that are fully integrated, so accurate structure factors can be estimated from far fewer diffraction patterns than needed when using narrower bandwidth radiation, where most peaks are only partial reflections.

For example beamline ID09 at ESRF provides up to 10^8 photons in a single 100 ps pulse when using a multilayer monochromator. Higher photon fluxes of more than 10^9 photons can be achieved with microsecond exposure times. This appears to be an optimal compromise between flux and time resolution for investigating many biological processes. Using a multilayer X-ray beam at beamline ID09 in combination with the newly developed charge integrating JUNGFRAU detector we have applied the method of fixed-target serial crystallography to collect data at a rate of 1 kHz with microsecond exposure times. With these experimental conditions only about 3000 diffraction patterns were required to obtain a complete high-quality diffraction dataset, which corresponds to 30 seconds of data collection time. The modification which were implemented in *CrystFEL* to process this diffraction data and the results of the experiment are presented in Section 6.4.

Building up on these modifications and making use of the novel indexing algorithm for pink-beam diffraction data developed by Gevorkov *et al.* 2019 [5], the data processing pipeline for pink-beam serial crystallography is proposed in the next chapter.

6.3 Using monochromatic software for polychromatic data processing

The approach of using monochromatic serial crystallography software for pink-beam data processing described by Martin-Garcia *et al.* 2019 [115] is based on the assumption that the majority of the detected Bragg peaks in a weak sparse diffraction pattern are sampled by the strongest part of the spectrum, which is assumed to be narrow enough to be considered monochromatic. If that were the case, the monochromatic auto-indexing algorithms should be able to index diffraction patterns using the wavelength corresponding to the peak in the energy spectrum to find the reciprocal space coordinates of all found peaks.

This assumption, however, is generally not valid. Even though a certain reflection with an arbitrary structure factor amplitude is more likely to be detected if it is sampled by the stronger part of the spectrum, a significant fraction of peaks in a diffraction pattern will correspond to stronger reflections sampled by weaker parts of the spectrum. This fact is illustrated in Fig. 6.3. Fig. 6.3a shows lysozyme diffraction pattern simulated using typical undulator spectrum with the mode energy $E_m = 15.2$ keV, bandwidth $\Delta E/E_m = 5\%$ (FWHM) and 15% low energy tail. 30 strongest Bragg peaks are circled, the numbers mark how far the central X-ray energy contributing to each reflection is from the mode energy. Fig. 6.3b

shows the distributions of X-ray energies contributing to 500, 30 and 10 strongest peaks. As can be clearly seen from the distributions, in the case of 30 strongest peaks, which would be considered as a sparse pattern but still easily indexable by most auto-indexing algorithms, the X-rays with the energies ranging from 13.3 to 15.25 keV are contributing. This corresponds to almost 13% of the mode energy. In the case of 10 strongest peaks, when even using monochromatic radiation auto-indexing algorithms would most likely fail because of too few peaks, the peaks are still sampled by the X-ray energy range $\Delta E/E_m > 6\%$, which is in no way can be assumed monochromatic. This conclusion is further confirmed in Section 7, where the real pink-beam diffraction data is shown to exhibit behavior identical to Fig. 6.3b.

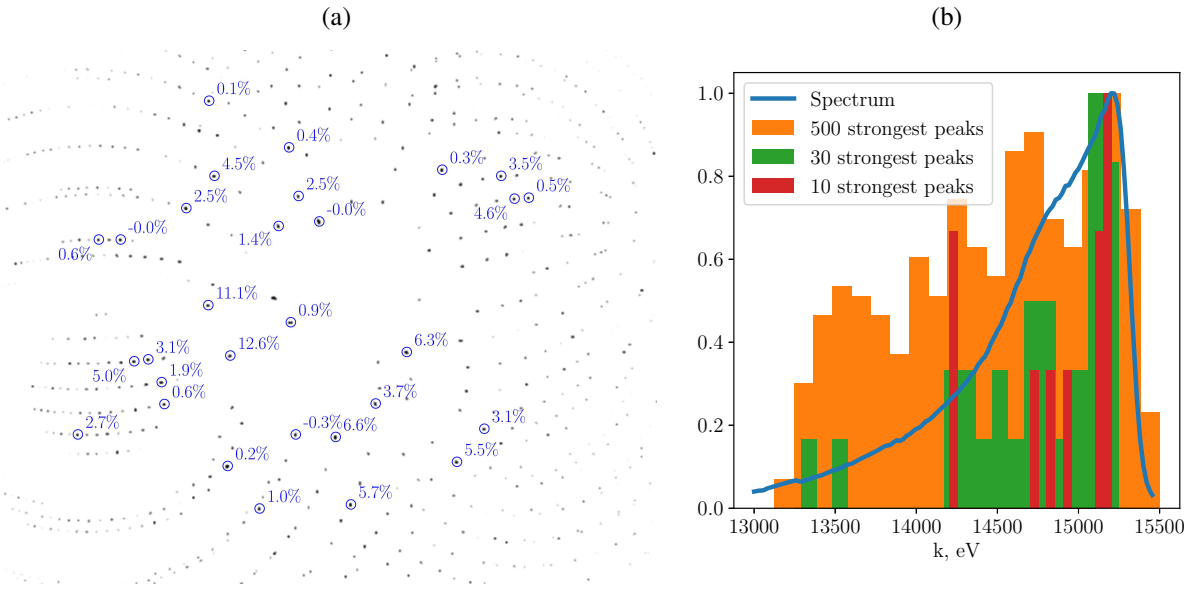


Figure 6.3: (a) Simulated polychromatic lysozyme diffraction pattern, 30 strongest peaks are circled with the numbers pointing out the relative difference between the central X-ray energy contributing to the peak and 15.2 keV. (b) shows the spectrum used for the simulation and the distributions of X-ray energies sampling 500, 30 and 10 strongest peaks.

Despite the fact that the peaks in the pink-beam diffraction pattern are produced by the wide range of X-ray energies, it has been clearly demonstrated by Martin-Garcia *et al.* 2019 that monochromatic indexing algorithms can sometimes recover crystal orientation with the accuracy sufficient to successfully integrate reflection intensities and solve the crystal structure. However, using this approach it was only possible to recover low resolution structure because all high resolution data had to be discarded during the processing. This can be explained as following: most indexing algorithms are based on finding the periodicity of the reciprocal lattice, which starts with projecting diffraction peaks onto the Ewald sphere to determine their reciprocal space coordinates. The peak detected at the angle 2Θ in the polychromatic diffraction pattern can be produced by the X-rays of any wavelength between λ_{min} and λ_{max}. It can therefore originate from any reciprocal lattice spot found on the interval between the limiting Ewald spheres shown in red on Fig. 6.4. When the monochromatic indexing algorithm is used to index polychromatic diffraction pattern, the inaccuracy of the determined reciprocal space position of each peak depends on the distance between the Ewald spheres: the positions of low resolution reflections are determined much more accurately. Given the sufficient number of low resolution reflections and omitting the high resolution reflections, as has been done by Martin-Garcia *et al.* 2019, the monochromatic indexing algorithms can indeed be successful in determining crystal orientations from pink-beam diffraction data.

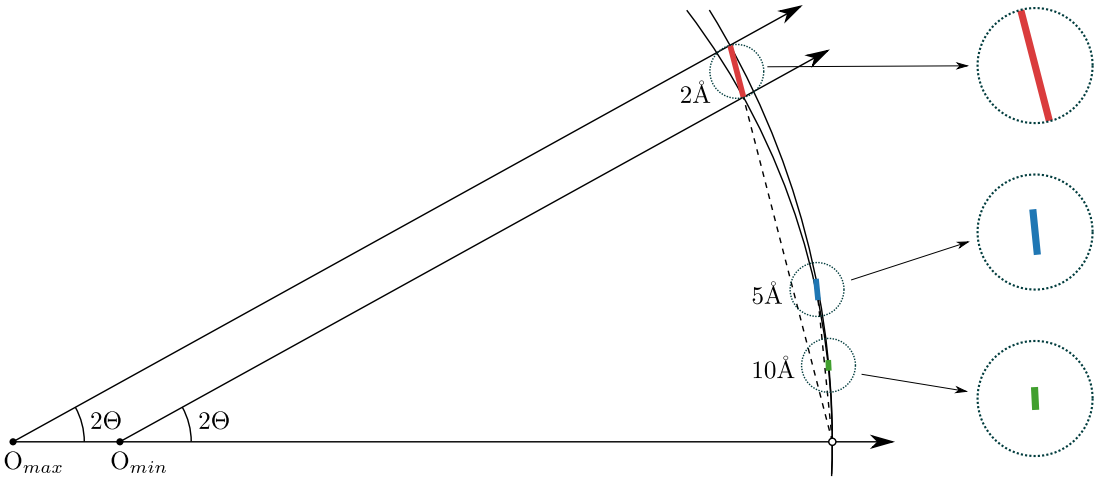


Figure 6.4: To-scale illustration of Ewald sphere construction in case of polychromatic beam with λ = 1Å, $\Delta\lambda$ = 0.15Å (15% bandwidth). Red, blue and green intervals show regions of the reciprocal space, where 2Å, 5Å and 10Å diffraction peaks can originate from.

In order to investigate the feasibility of this approach to polychromatic data processing, two lysozyme diffraction datasets were simulated using the typical undulator spectrum of 4.8% bandwidth (FWHM) with 15% low-energy tail and the Gaussian shaped spectrum with 3% bandwidth (FWHM), close to what can be achieved using multilayer monochromator (Fig. 6.5).

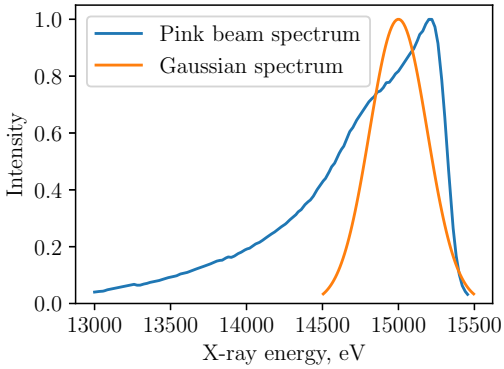


Figure 6.5: X-ray spectra used for simulation: pink beam spectrum (blue) with the peak at 15.2 keV, 4.8% FWHM and 15% low-energy tail, and gaussian-shaped spectrum (yellow), centered at 15 keV with 3% FWHM.

Both datasets were generated using the same set of 100 random crystal orientations. Diffraction patterns were indexed with *indexamajig* program in *CrystFEL*, using *MOSFLM*, *asdf* and *DirAx* monochromatic indexing algorithms. Found unit cell parameters and crystal orientations were then compared to the ones used for simulation. If the found parameters were less than 5% different from the expected ones and the misalignment angle between the found and the target unit cell was less than 5°, the pattern was considered to be indexed correctly. By careful optimization of the peak-finding parameters, i.e. the number of peaks and the high-resolution cut-off, and using the `--retry` option in *indexamajig* [1], it was possible to index 79% of diffraction patterns simulated using the undulator spectrum. In the case of the

[1] `--retry` option successively rejects the weakest reflections and attempts indexing until the correct solution is found.

gaussian spectrum, no optimization was required as *indexamajig* yielded ≥98% indexing fractions as long as at least 20 peaks were found in each diffraction pattern.

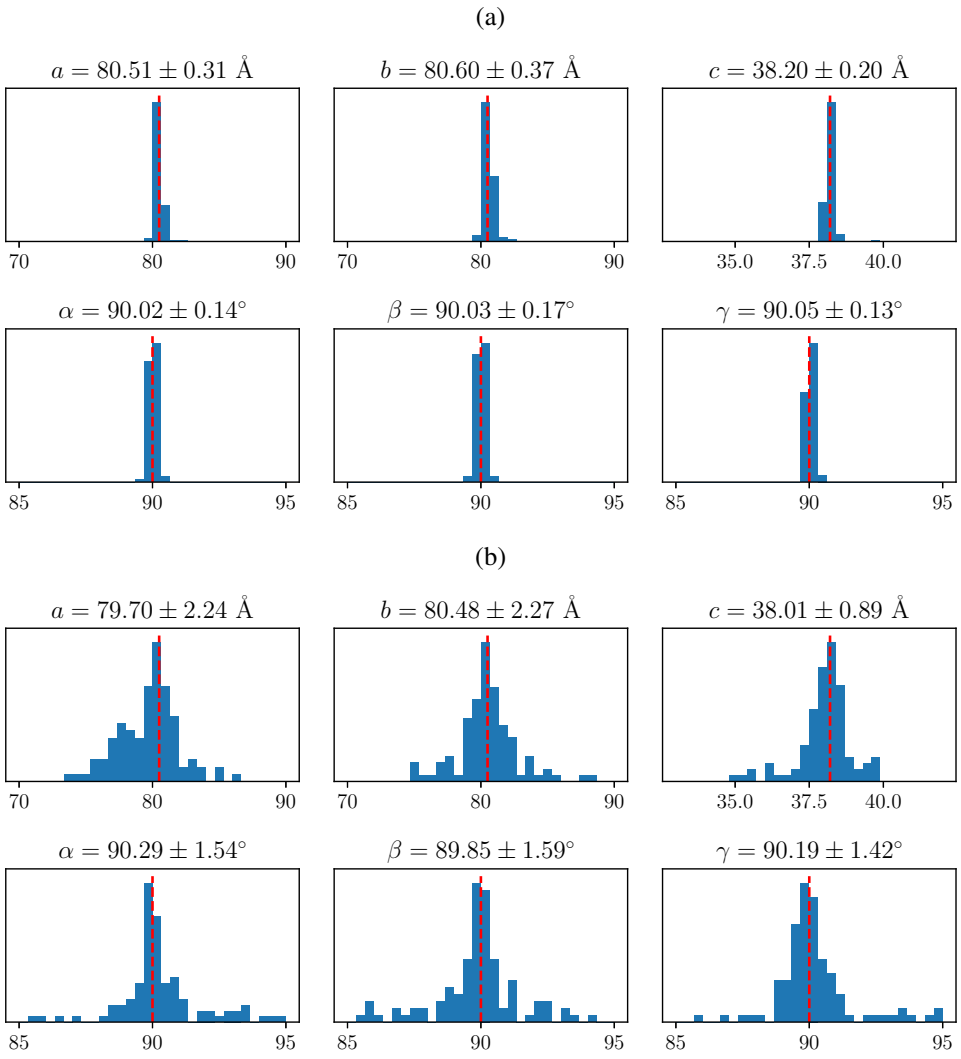


Figure 6.6: Distributions of the unit cell parameters obtained by indexing simulated polychromatic diffraction data using monochromatic indexing software in the case of (a) gaussian-shaped spectrum and (b) pink-beam spectrum. Dashed red lines show the target unit cell used for simulation: $a = b = 80.5$Å, $c = 38.2$Å, $\alpha = \beta = \gamma = 90°$.

The resulting distributions of the unit cell parameters and the misalignment angles are shown in Fig. 6.6 and 6.7 respectively. The indexing solutions demonstrate much higher discrepancy between the found and the target unit cell parameters in the case of the 4.8% BW pink-beam spectrum compared to 3% BW gaussian spectrum. Apart from the overall ∼7 times larger standard deviation of each parameter, the mean values of a and b, which should both be equal 80.5Å, differ by about 1% for pink-beam data while in the case of the gaussian spectrum the difference is only 0.1%. The root-mean-square unit cell misalignment angle in the case of pink-beam is as well about 7 times larger than in the case of the gaussian beam. As a consequence, although the majority of the simulated pink-beam diffraction patterns can be indexed with the monochromatic algorithms, the discrepancy in the found unit cell leads to inaccurate Bragg spot prediction and, therefore, incorrect integrated intensities, as it is illustrated in Fig. 6.8. While in the case of the gaussian-shaped spectrum (Fig. 6.8a) the predicted Bragg spot positions correspond

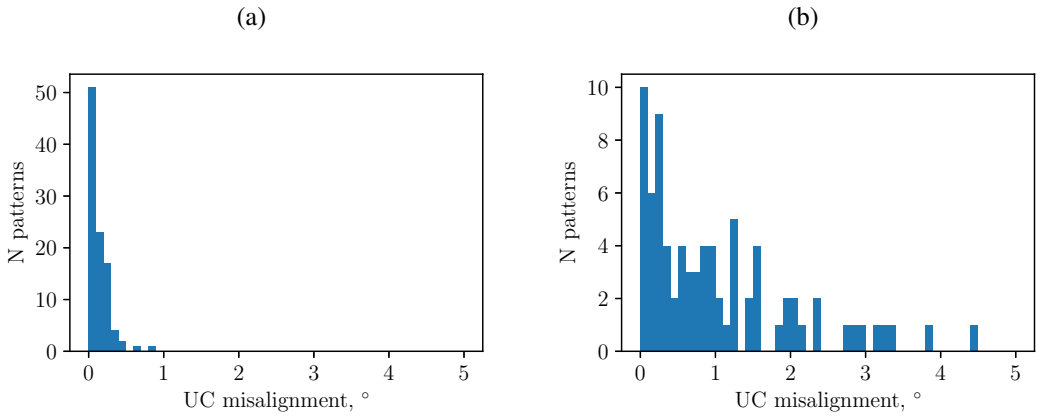


Figure 6.7: Distributions of misalignment angle between the simulated and found unit cells in case of (a) gaussian-shaped spectrum and (b) pink-beam spectrum. The root-mean-square deviation from zero is 0.19° in (a) and 1.44° in (b).

well with the found diffraction peaks, in the pink-beam diffraction pattern (Fig. 6.8b) the predicted and found spots are clearly misaligned. This misalignment is further quantified in Fig. 6.9, which shows how many diffraction patterns in each dataset contain a certain percentage of peaks predicted within the certain distance from the correct positions. Fig. 6.9b makes it clear that although 79% of pink-beam diffraction patterns could be indexed, far from all of them would provide valuable information. For example, given the peak size of 4-5 pixels, if the reflection intensities were integrated within 5 pixel radius around the predicted spot position, only 28% of patterns would have more than 90% of reflections correctly integrated in the case of the pink-beam as compared to 97% in the case the of gaussian beam.

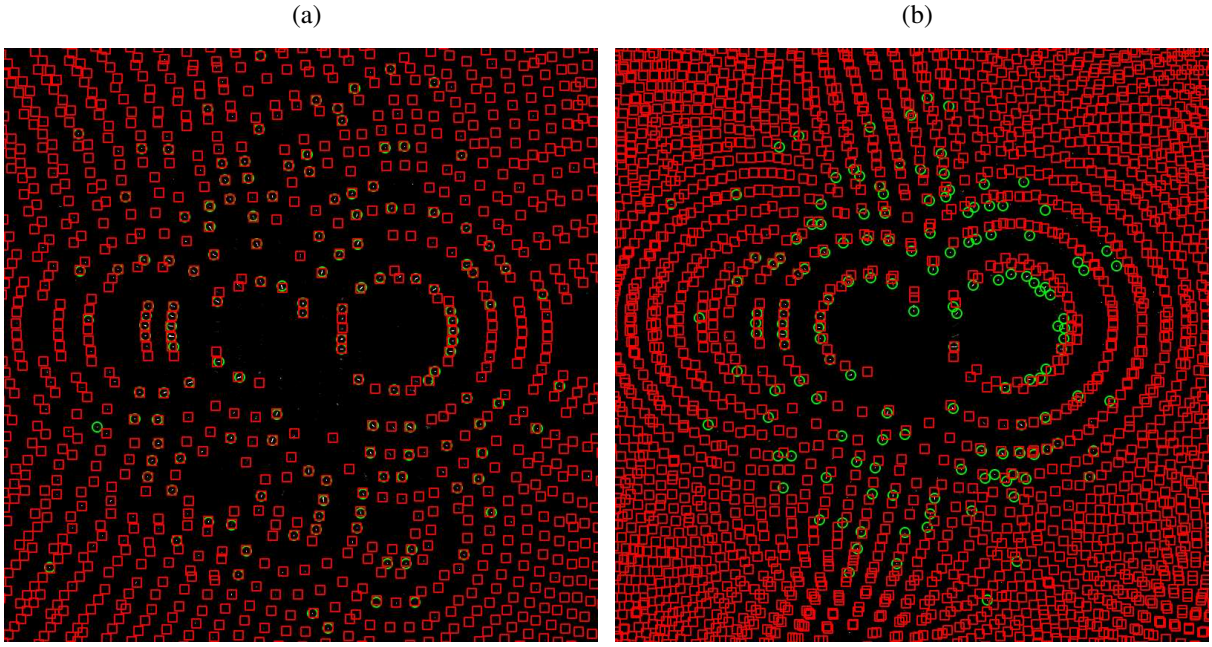


Figure 6.8: Lysozyme diffraction patterns simulated using (a) gaussian-shaped spectrum and (b) pink-beam spectrum in the same crystal orientation, both indexed by *MOSFLM*. Green circles show the peaks used for indexing, red squares show peak positions predicted using the found indexing solution.

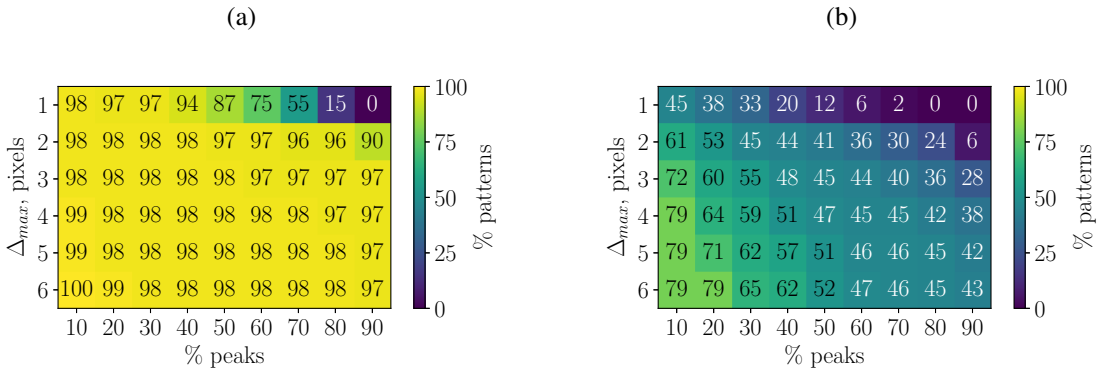


Figure 6.9: Percentage of diffraction patterns, where % peaks are predicted within the accuracy of Δ_{max} pixels in case of (a) gaussian-shaped spectrum and (b) pink-beam spectrum.

The analysis above leads to a conclusion that the approach of using monochromatic indexing algorithms for pink-beam diffraction data is rather limited: in order to obtain usable intensities from such indexing solutions one would have to either exclude the majority of patterns where the spots don't overlap with the prediction, as it was done by Martin-Garcia *et al.* 2019 [115] and resulted in only 10% of the collected diffraction patterns being used, or develop a new prediction refinement algorithm as the current ones are only suitable for monochromatic diffraction. On the other hand, the method was proved to be perfectly suitable for the analysis of polychromatic diffraction data simulated using the gaussian-shaped spectrum with 3% bandwidth, which suggests it can be applied to process serial crystallographic data collected using the multilayer monochromator.

6.4 Serial crystallography with 2.5% X-ray bandwidth

This section describes experimental implementation of the serial crystallography method using polychromatic X-ray beam with 2.5% bandwidth, which was published in Tolstikova *et al.* 2019 in *IUCrJ* [1]. This is a proof-of-principle study which demonstrates the feasibility of the method for fast and efficient data collection at a rate of 1 kHz at a synchrotron, using a JUNGFRAU detector and a Roadrunner II goniometer for high-speed sample delivery [65].

6.4.1 Experiment at beamline ID09 at ESRF

Diffraction experiments were performed at beamline ID09 at ESRF using the multilayer monochromator installed at the instrument. The resulting energy spectrum of the X-rays used for the experiment is shown in Fig. 6.10 with an energy spread of 2.5% (FWHM) centered at a photon energy of 15.2 keV. The measured beam size at the sample position was 60×60 μm^2.

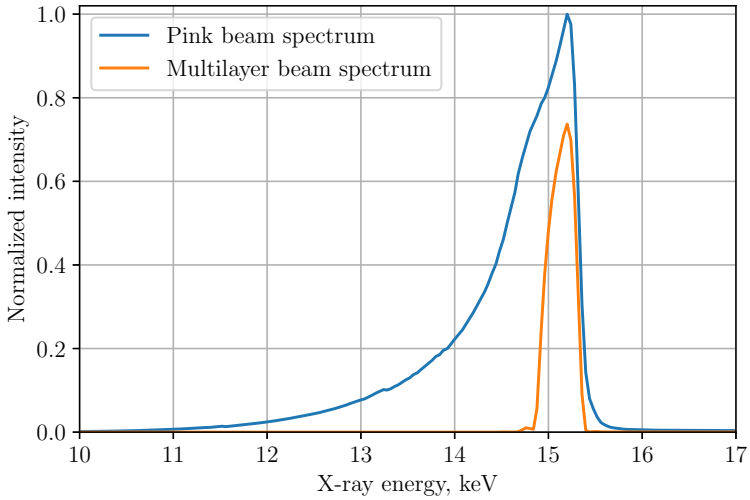


Figure 6.10: Measured X-ray energy spectrum at beamline ID09 with and without multilayer monochromator.

Diffraction patterns were recorded on a JUNGFRAU 1Mpixel detector. The detector consisted of two individual 500Kpixel JUNGFRAU modules mounted on top of each other with a vertical gap of 2.8 mm in between them. Both modules were in the same plane perpendicular to the X-ray beam. With the given detector area of 77 mm×80 mm and a minimum detector distance of 100 mm limited by geometrical restrictions at the instrument, we decided to offset the detector center horizontally by 26 mm with respect to the incident X-ray beam to allow collection of reflections up to 1.4 Å at the edge of the detector (Fig. 6.11). The detector provides single photon sensitivity in combination with a dynamic range of about 8000 photons at the 15.2 keV energy used in our experiment (Fig. 6.14). This high dynamic range is possible thanks to the automatic gain capability of each individual pixel: each pixel adjusts (lowers) its gain to cope to the incoming signal [44].

Crystals of the two model compounds, lysozyme and proteinase K, were directly grown on micro-patterned silicon chips as described in more detail by Lieske *et al.* 2019 [117]. Crystals of both proteins all had dimensions of about 50×50×50 μm^3. A technical drawing of the silicon chips used for the

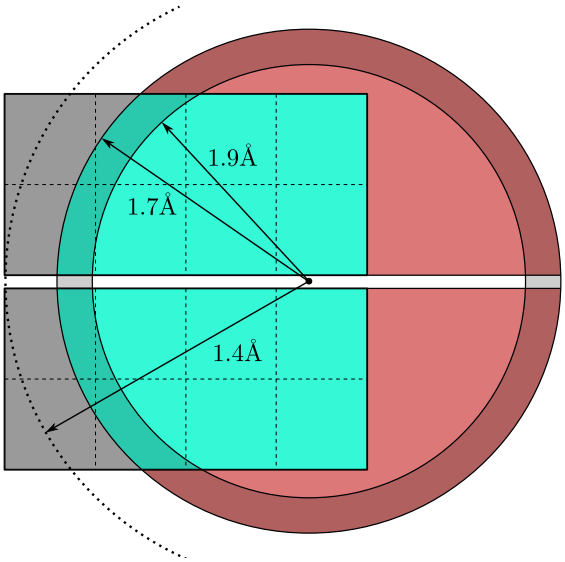


Figure 6.11: Illustration of the detector geometry used in experiment at ID09: JUNGFRAU 1M detector consisting of two panels with the total area of 77 mm×80 mm moved horizontally by 28 mm. The green area shows where the data of up to 1.7 Å resolution is recorded, the red area shows where the data is lost.

experiment, also referred to as Roadrunner II chips, and the corresponding pore-pattern are shown in Fig. 6.12. For data collection, a Roadrunner chip with crystals was taken out of its crystallization chamber and the crystal growth solution was removed through the pores by blotting the underside of the chip with filter paper [17]. After blotting the chips were protected with a cover and transferred to the beamline where they were inserted into the Roadrunner II measurement chamber whilst the protective cover was retracted from the chip area.

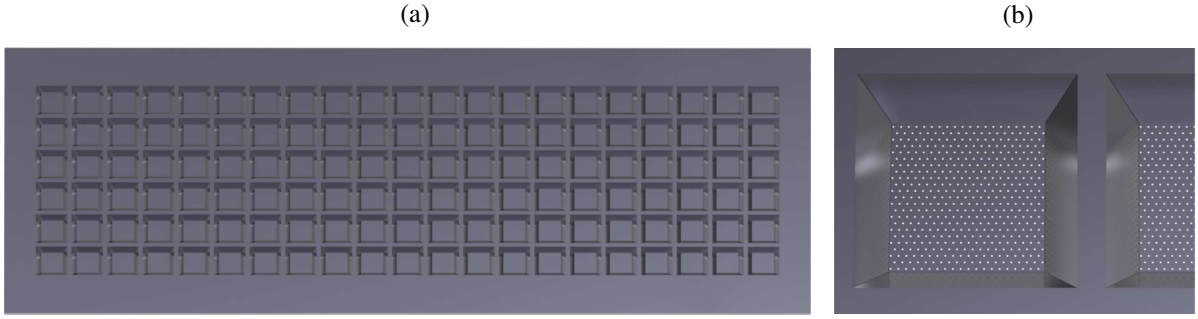


Figure 6.12: (a) Roadrunner II chip with dimensions of 33×12 mm^2 (h×v). The chip provides 21×6 compartments each with a size of 1.0×1.0 mm separated by a support frame structure with a width of 600 µm and a thickness of 300 µm. (b) The membrane thickness of the 126 individual compartments is 10 µm and the membranes are equipped with hexagonal patterns of micro-pores with diameters of 20 µm and a spacing of 50 µm between the pores. Due to the horizontal beamsize of 60 µm used for the experiments, which is larger than the pore separation, it was decided to expose at intervals of twice the pore-spacing in order to avoid double exposure of the same crystal.

In contrast to instrumentation usually available at crystallography endstations, the Roadrunner II goniometer is equipped with a high-speed horizontal scanning stage (x-axis) capable of scanning at speeds of up to 100 mm/s. This fast scanning axis is mounted on a y, z-translation stage allowing for it to be

positioned vertically (the y-direction) and along the X-ray beam direction (the z-direction). This whole scanning unit can be rotated (by an angle ω) around the x-axis, using a high-precision air-bearing. A technical drawing of the Roadrunner II goniometer as used for the experiment is shown in Fig. 6.13.

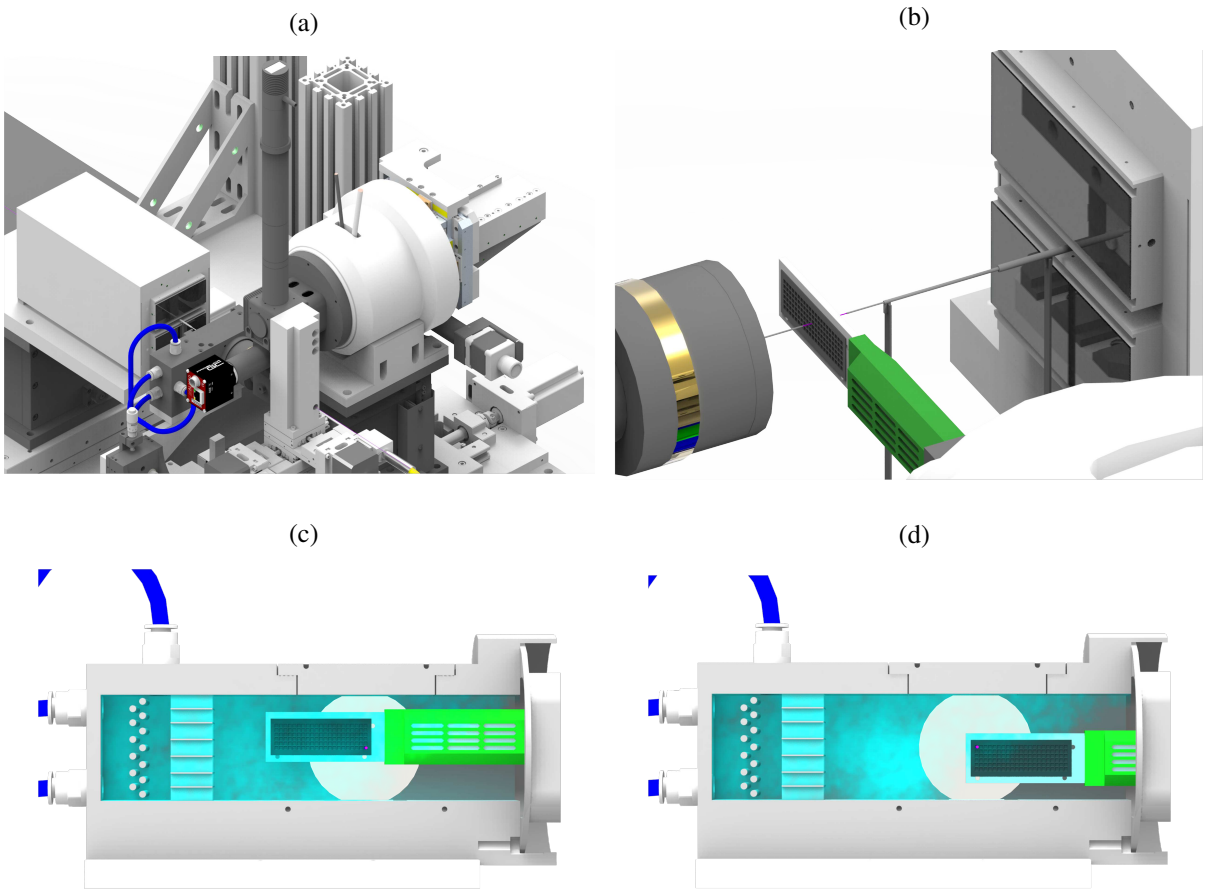


Figure 6.13: (a) Technical drawing of Roadrunner II goniometer together with the JUNGFRAU 1 M detector installed at beamline ID09 at the ESRF. (b) Close-up of the interaction region showing the inline sample-viewing microspcope with the collimator (left), the chip with the retracted humidor (in green), the capillary beamstop enclosing the direct beam shortly after the chip and the JUNGFRAU 1M detector (right). For better visibility, the humidity chamber is not shown here. The X-ray beam bath is highlighted in pink. (c) Roadrunner II chip inside the measurement chamber. The observed humidity gradient from the top left to the lower right side as observed in the chamber is indicated in light blue, with areas of higher humidity being brighter. In the 'in-position' at the start of a scan the whole chip area is in a region of high relative humidity. (d) 'out-position' of a chip at the end of a measurement. In particular the lower right side of the chip is in an area of lower relative humidity.

Once mounted onto the scanning unit, each chip was aligned with respect to the X-ray beam with an inline sample-viewing microscope and the scanning grid was defined using the Roadrunner software. For subsequent data collection the chip was continuously scanned through the X-ray beam in the horizontal direction with a constant velocity of 100 mm/s. With an X-ray pulse frequency of 1 kHz, generated by an X-ray chopper, this corresponds to a spatial separation of 100 m between two shots, which is about twice the beamsize at the sample position. During an X-ray exposure of 1 µs, the crystal moves only by 100 nm, which is insignificant compared with the crystal and beam sizes. The scans started at the bottom right corner of every chip. After a horizontal line scan was finished, the chip was moved down vertically by

100 µm to the next line, rotated by a small ω increment, and then scanned along x in the reverse direction. This procedure was repeated for the whole chip.

In order to exploit the high repetition rate of 1 kHz achievable by the Roadrunner II gomiometer and the JUNGFRAU detector, it was necessary to select X-ray pulses at the same repetition rate and to vary their duration to vary the X-ray dose to the crystals. This was achieved by the use of two choppers and a fast shutter. The first chopper (heat load chopper) was located upstream of the ID09 beamline and selected 80 µs duration X-ray pulses at 1 kHz. The second chopper (high speed chopper) was operated in the so-called "tunnel-less" mode [118] and used to select X-ray pulses with variable duration. The chopper slit width gradually increases with its horizontal position, so a horizontal translation of the whole device allows adjusting the exposure time in the 1-25 µs range.

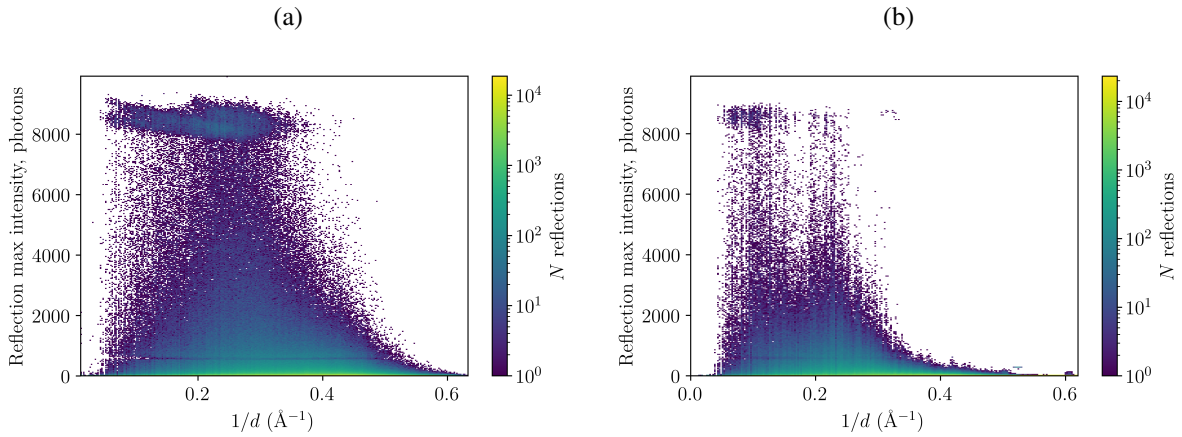


Figure 6.14: Distribution of the highest pixel values in each reflection as a function of resolution for lysozyme diffraction data collected with a JUNGFRAU detector with (a) 80 µs and (b) 5 µs exposures. The high-density cloud around 8000 photons represents saturated reflections.

After the initial test measurements of lysozyme crystals with 80 µs exposure time it became clear that at such conditions the large portion of Bragg peaks saturate the detector (Fig. 6.14a). The optimal exposure time was found to be around 5 µs as it gave a small fraction of saturated peaks while still allowing to successfully utilize the whole dynamic range of JUNGFRAU detector (Fig. 6.14b). To demonstrate feasibility of the method at even shorter exposure times, lysozyme and proteinase K data was also collected with 1 µs exposure. With the beam parameters mentioned above and 3.5×10^9 photons per 5 µs exposure and 7×10^8 photons per 1 µs exposure, these exposure times correspond to X-ray doses of 500 Gy and 100 Gy, respectively. At these doses data should not be affected by radiation damage or sample heating effects, even without cryogenic cooling (Section 2.3) [14, 65]. This was also confirmed experimentally by measuring multiple diffraction patterns at the same position of the chip and comparing the Bragg peaks at high resolution.

Two example diffraction patterns of a lysozyme crystal and a proteinase K crystal are shown in Fig. 6.15. As can be seen in Fig. 6.15, the background scattering levels obtained in our measurements are very low, with the vast majority of the pixels having zero counts. This is a result of the low-background experimental setup [2] in combination with the single photon sensitivity of the JUNGFRAU detector. This leads to an improved signal to-noise-level of the data compared to that obtained with contemporary CCD detectors, and thereby a higher overall data quality than usually achievable in conventional crystallographic

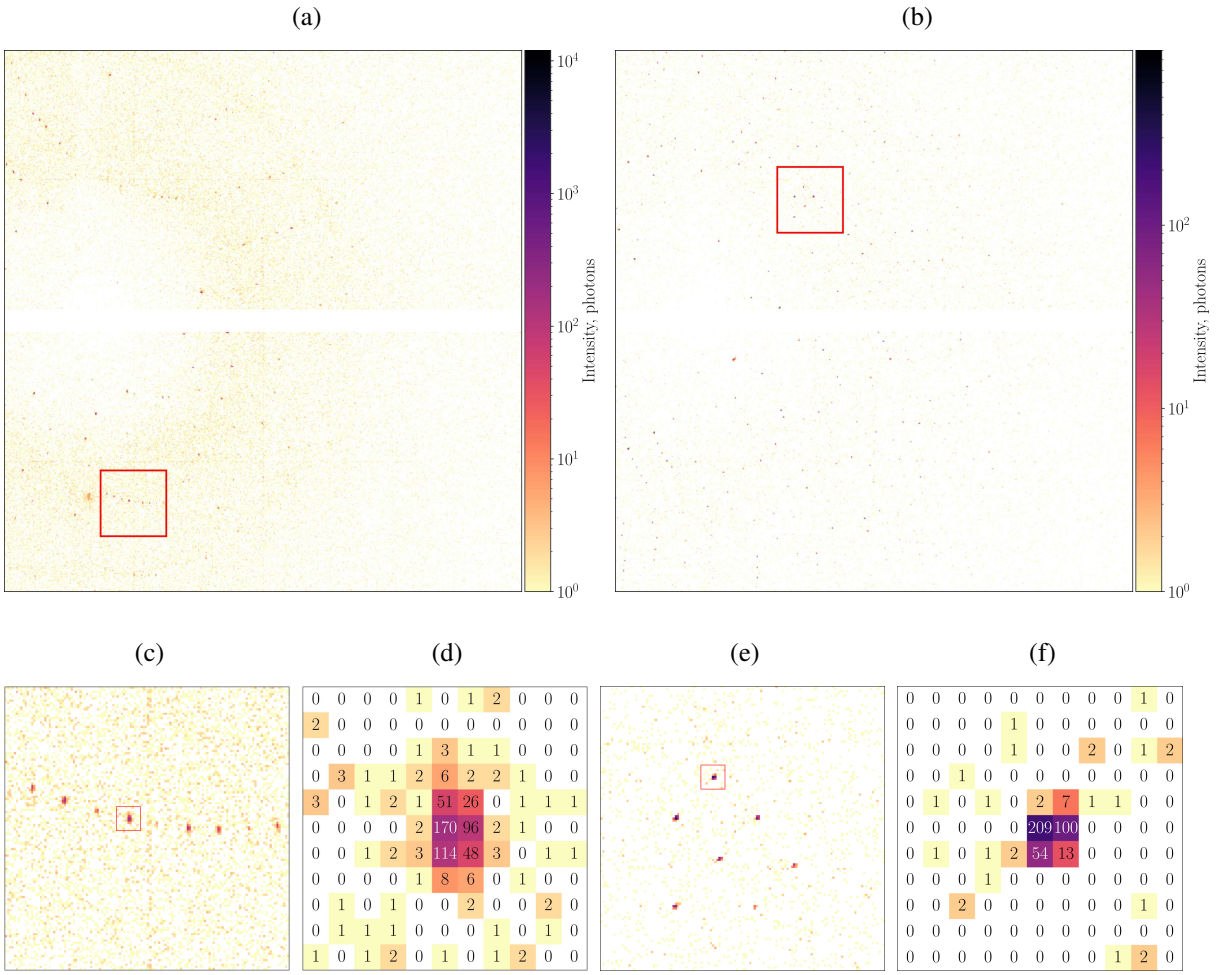


Figure 6.15: Polychromatic diffraction patterns of a lysozyme crystal (a) from chip lys08 recorded at beamline ID09 with an 1M JUNGFRAU detector with an exposure time of 5 µs and a proteinase K crystal (b) from chip protk04 with an exposure time of 1 µs. Magnified areas of the diffraction images indicated by a red square in images (a) and (b) are shown in figures (c) and (e). Figures (d) and (f) show even higher magnifications of the areas indicated in (c) and (e) and highlight the achievable low background scattering levels around the Bragg reflections at 3.1 Å in case of lysozyme (d) and 3.6 Å in case of proteinase K crystals (f).

experiments, especially for high-resolution reflections.

In total we collected diffraction data from 10 chips. Scanning and data-collection parameters for every chip are provided in Table 6.1. On average 36000 diffraction patterns were collected per chip with a scanning time of about 150 s for an entire chip. This is longer than the 36 s of data collection time, due to the overhead of changing direction at the end of the scan. The hit fraction depends on the crystal growth conditions. In the case of lysozyme crystals with 5 µs exposure, the average hit fraction was 30% and of these patterns 76% could be indexed, corresponding to an effective data collection rate of 55 indexed patterns per second. For the lysozyme and proteinase K crystals measured with 1 µs exposure, the effective data collection rate was lower with 28 and 9 indexed patterns per second, respectively, which is probably a result of a lower crystal density on the chip.

Chip name	lys08	lys09	lys10	lys11	lys12	lys13	lys14	lys15	protK3	protK4
Exposure time, μs	5	5	5	1	1	1	1	1	1	1
Number of horizontal scan points	331[a]	331[a]	331[a]	333[a]	331[a]	151[b]	163[b]	156[b]	310[a]	156[b]
Number of vertical scan points	116	105	104	116	105	24[b]	116	105	114	105
Total number of scan points	38396	34752	34423	38628	34754	3580	18907	16365	35340	16379
Number of hits	12209	12512	7489	5376	4621	937	4707	3443	2538	640
Number of indexed and merged hits	9238	8813	6293	4448	3885	762	3386	2312	1366	219
Total scanning time, s	158	143	142	158	143	28	139	125	153	125
Hits per second	77	87	53	34	32	33	34	28	17	5.1
Indexed patterns per second	58	62	44	28	27	27	24	18	8.9	1.8
Effective scanning rate, frames/s	243	243	243	244	243	130	137	131	231	132

[a] 15 scan points at the beginning and end of every line were used for acceleration and deceleration of the linear axis, so the total horizontal scanning range slightly exceeds the chip lengths.
[b] These chips were only partially scanned.

Table 6.1: Chip scanning parameters for 1 kHz fixed target data collection with the Roadrunner II goniometer. All chips were scanned with a horizontal scanning speed of 100 mm/s.

6.4.2 Data analysis

Dark pedestal images were collected, for all the detector gains, at the beginning of every measurement run, adding few tens of seconds to each runtime. Using the pedestal images and per-pixel gain factors, determined in a laboratory-based dedicated calibration procedure [119], the raw data was converted into photon counts and saved to multi-event HDF5 files. Hit finding was performed within *CrystFEL* using *peakfinder8* for peak detection and an option to skip over patterns with the number of peaks lower than a certain threshold. The images with more than 20 peaks were classified as hits. The initial estimate of the detector distance as well as the relative positions of two detector modules were determined from alanin powder diffraction images collected at two different detector distances.

Similar to the analysis of simulated data presented in the previous section, diffraction patterns were indexed with *indexamajig* using monochromatic indexing algorithms *MOSFLM*, *asdf* and *DirAx* with '–retry' option. Bragg spot prediction and integration however were modified to account for non-monochromaticity. The indexing results were then used to further refine detector geometry with *geoptimiser* [86]. All diffraction patterns from multiple crystals (those with less than 80% of the found peaks correctly predicted) and patterns from crystals with high mosaicity (elongated peaks, for example

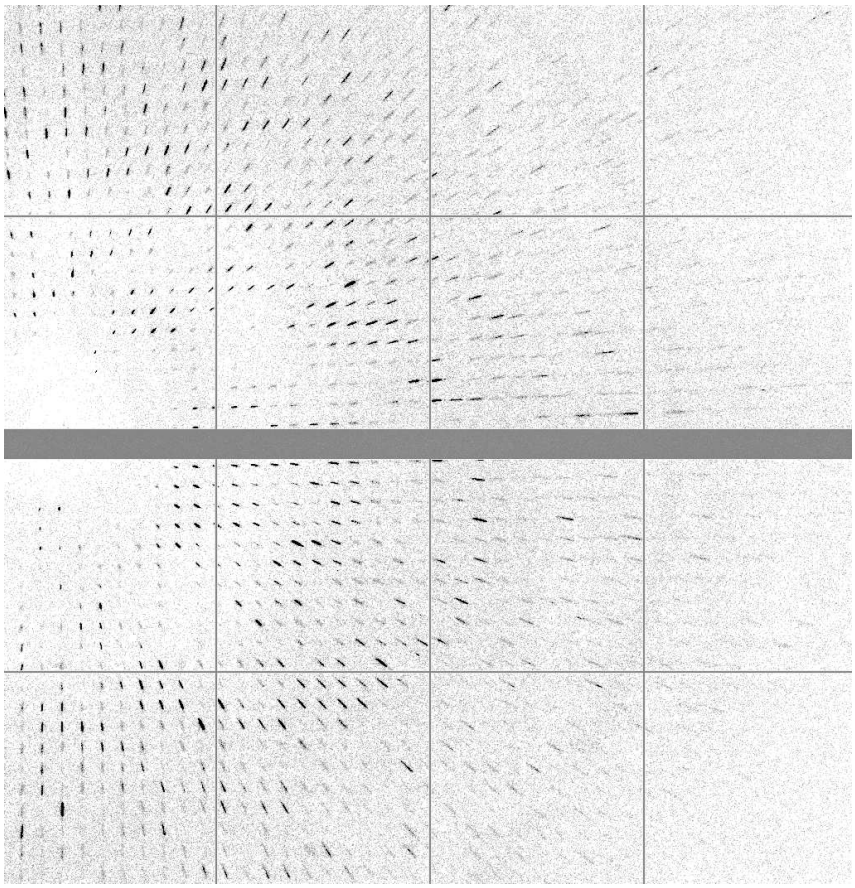


Figure 6.16: Example diffraction pattern of a highly mosaic proteinase K crystal measured with 2.5% X-ray bandwidth using multilayer monochromator at the beamline ID09 at ESRF.

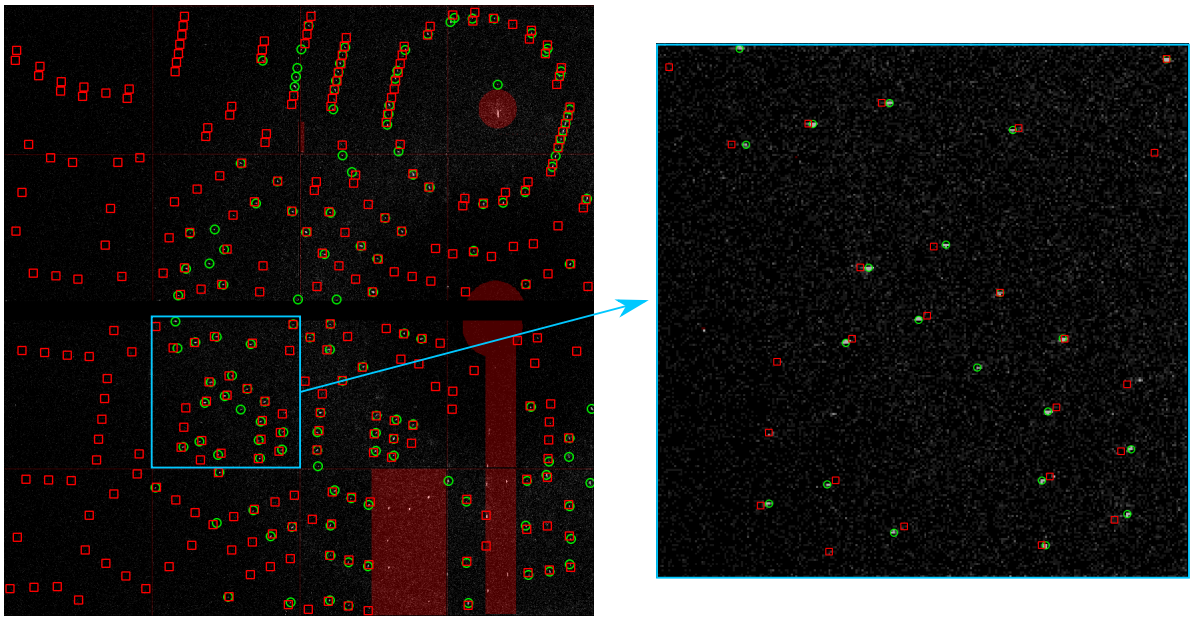
Fig. 6.16) were discarded before merging. In total 95% of lysozyme hits and 75% of proteinase K hits were indexed, 20% and 34% of them were discarded.

6.4.2.1 Bragg peak prediction

After the initial attempts to use *indexamajig* for indexing and integration of both simulated and experimental polychromatic data, it became apparent that the default Bragg peak prediction algorithm, developed for monochromatic data, was not suitable in this case. The example lysozyme diffraction pattern shown in Fig. 6.17a perfectly illustrates the problem: although the pattern is clearly indexed correctly thus the overall patterns of observed and predicted peaks are matching, the close-up view shows that predicted spot positions don't match with positions of the found peaks. If the integration radius of 4 pixels, which is a reasonable value considered the peaks are rather small (see Fig. 6.15), was used, the majority of the peaks of the enhanced region in Fig. 6.17a would not have been integrated. Therefore, modification of the spot prediction was the first necessary step in the adaptation of the processing pipeline to polychromatic data.

First, *indexamajig* was modified to accept the X-ray energy spectrum as an input parameter. Spectrum was sampled with a step size chosen in such way that the spot positions predicted for the same reflection at the highest possible resolution using two consecutive X-ray energies would differ by less than one pixel. In our case, with 1.4 Å resolution at the edge of the detector, it corresponds to the step size of 12 eV. To predict reflections originating from the whole X-ray energy range the following procedure was implemented:

(a)



(b)

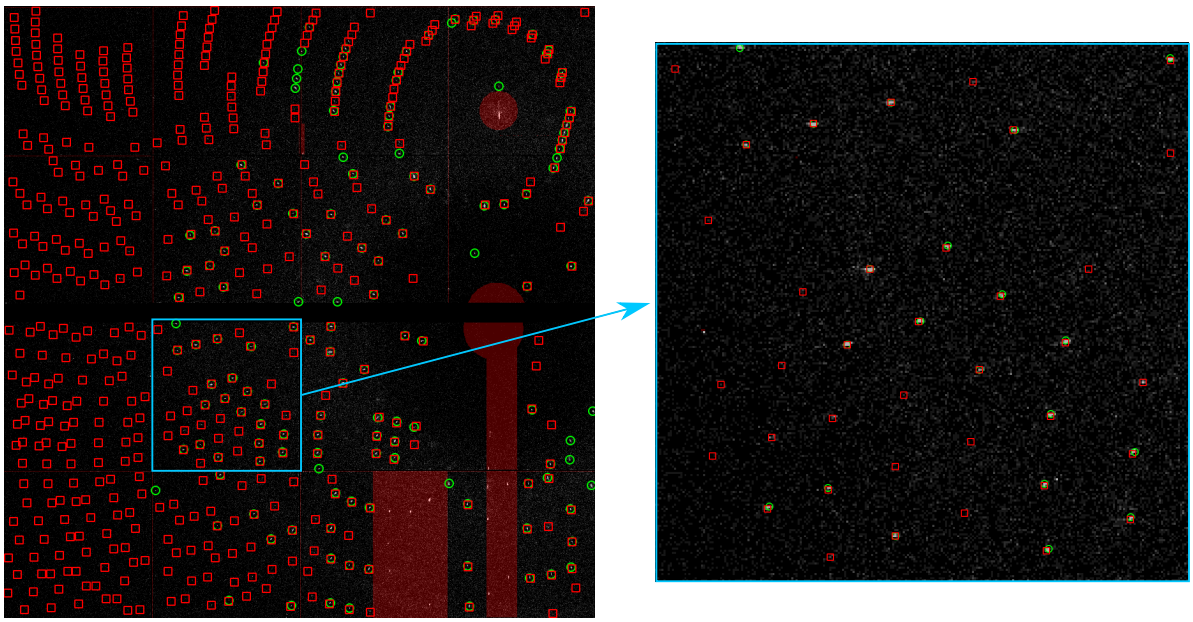


Figure 6.17: Lysozyme diffraction pattern indexed with CrystFEL without modifications (a) and with modifications to take into account 2.5% bandwidth (b). Found and predicted Bragg spot positions are shown as green circles and red squares respectively.

- the standard monochromatic prediction algorithm is invoked for each energy in the spectrum,
- the range of contributing X-ray energies is determined for each reflection, predicted at least once,
- detector position of each predicted reflection is calculated by averaging the contributing energies,
- the spectrum intensity of the average X-ray energy is saved for further scaling.

Fig. 6.17b demonstrates clear improvement of the Bragg spot prediction arising from the modified

algorithm. It shows the same lysozyme diffraction pattern and exactly same indexing solution as Fig. 6.17a, but the predicted reflections and the found peaks overlap significantly better.

It has to be noted here, that the automatic determination of the peak profile radius in *indexamajig* proved to be not applicable to the polychromatic data: as it is based on finding the distance between the reciprocal lattice points and an infinitely thin Ewald sphere, it tends to largely overestimate the profile radius for pink-beam data. While similar procedure can in theory be applied in case of the polychromatic beam with the sharply cut off energy spectrum, for the tailed spectrum a different approach is required. Currently, peak profile radius has to be set and optimized manually to a value which best predicts the low-resolution reflections.

6.4.2.2 Integration of intensities

Reflection intensities were integrated using the three-rings integration method in *indexamajig*(Section 4.1.4). The integrated intensity of each reflection was then divided by the normalized spectral weight of the central X-ray energy contributing to the reflection. The obtained intensities from single crystal diffraction patterns were then merged with *partialator* with one round of scaling without partiality correction (Section 4.1.4.2).

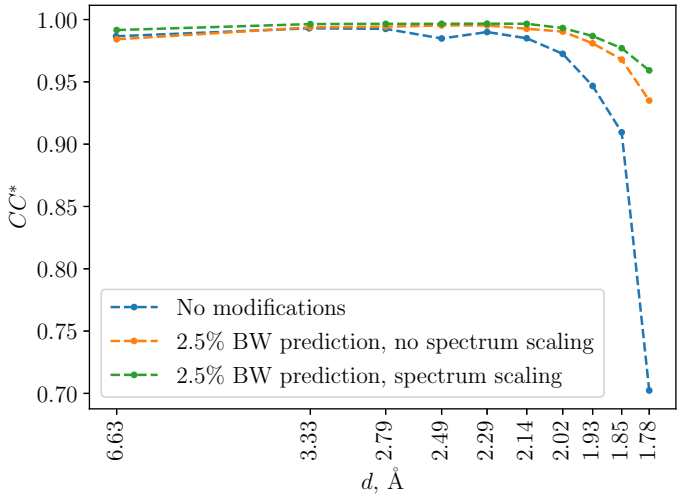


Figure 6.18: Comparison of CC^* as function of resolution for the dataset consisting of 8813 indexed diffraction patterns collected from the chip lys09, processed with *CrystFEL* without modifications (blue), with modifications to take into account 2.5% bandwidth (yellow) and scaling the Bragg intensities according to the intensities in the spectrum (green).

CC^* of the dataset merged from one lysozyme chip processed using three different versions of *indexamajig* (default version without modifications, version with modified peak prediction and version with modified peak prediction + spectrum normalization) is plotted as a function of resolution on Fig. 6.18. It reveals a significant improvement arising from the implemented spot prediction, especially at high resolution. A further slight improvement arises from scaling of the Bragg intensities according to the X-ray spectrum.

6.4.3 Results

In total, 24,344 of 5 μs exposure and 14,793 of 1 μs exposure lysozyme diffraction patterns were merged and used for structure refinement. In the case of proteinase K with 1 μs exposure, 1585 patterns were merged. Structure refinements for all generated datasets were carried out using PHENIX [120]. PDB structures 6FTR and 5KXV served as starting models for lysozyme and proteinase K [54, 121]. Table 6.2 shows data collection and structure refinement parameters of the resulting datasets. The lysozyme data collected with 1 μs shows only slightly worse metrics compared to 5 μs dataset with exactly same number of patterns, implying that the method is indeed feasible for such short exposure times. Lack of significant increase of the B-factor between 1 μs and 5 μs suggests that there is very little radiation damage even at 5 μs exposure. Proteinase K structure, although refined from only 1585 patterns, is of reasonable quality, with R_{work}/R_{free} = 0.17/0.23, which again proves that serial crystallography with the polychromatic beam requires far fewer patterns compared to monochromatic case.

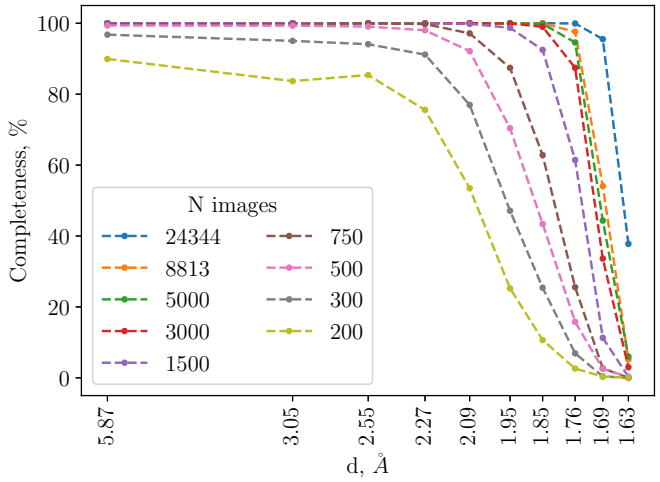
6.4.3.1 Dependence of data quality on number of merged patterns

The lysozyme diffraction datasets recorded with 5 μs exposure time (lys08, lys09, lys10) were further analyzed to determine the dependence of analysis metrics on the number of diffraction patterns collected. These three chips provided a total of 24344 indexed diffraction patterns. I created sixteen subsets from this group, consisting of 200 to 15000 randomly selected lysozyme diffraction patterns, plus the full dataset, which were all individually merged. The resulting data completeness, percentage of reflections with $I/\sigma(I) > 2$ and CC^* as function of resolution are shown in Fig. 6.19. From all datasets, structure refinements were then carried out with PHENIX [120], and the resulting refinement R-factor, R_{free}, is plotted in Fig. 6.20.

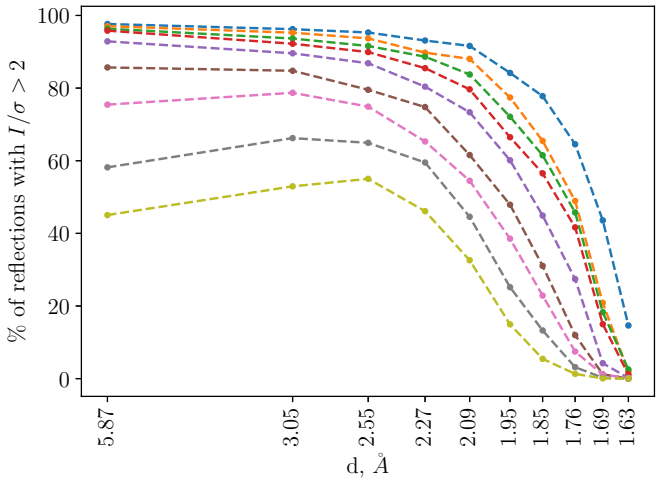
As seen in Fig. 6.19, all datasets containing more than 500 diffraction patterns show almost 100% completeness up to a resolution of 2.3 Å. Here, completeness is defined as a fraction of reflections in the resolution shell that have been integrated at least once regardless of their intensity. With increasing numbers of merged patterns this metric extends to higher resolution. The completeness of the dataset containing all 24 344 frames remains close to 100% for a resolution of up to 1.7 Å. The dependence of both fraction of reflections with $I/\sigma(I) > 2$ and CC^* on the number of patterns exhibit a different behavior. Only the datasets containing more than 1500 patterns show CC^* values larger than 0.95, which then falls off at resolutions higher than 2.3 Å. Again, the dataset containing all patterns shows the highest CC^*, which also extends to the highest resolution. Interestingly, the datasets consisting of a smaller number of diffraction patterns show lower metrics not only for the high-resolution but also for the low-resolution reflections.

As seen in Fig. 6.20, R_{free} first decreases rapidly with the number of patterns, from 0.35 for 200 patterns to about 0.21 for 1500 merged patterns. Beyond this number of patterns there is little improvement in this metric which decreases only slightly to 0.175 when all 24000 patterns are included. Both the low CC^* values for the low-resolution reflections and the relatively high R_{free} values for datasets consisting of less than 1500 diffraction patterns can be better understood with reference to the Ewald construction of diffraction as illustrated in Fig. 6.21. The limiting spheres of the minimum and maximum wavelengths of the polychromatic radiation define a volume of reciprocal space where reflections occur. To first order, this volume can be thought of as a wedge. At low resolution, the wedge is thinner than the peak width, giving mainly partial reflections that contribute to a variance in their measured intensities. This situation

(a)



(b)



(c)

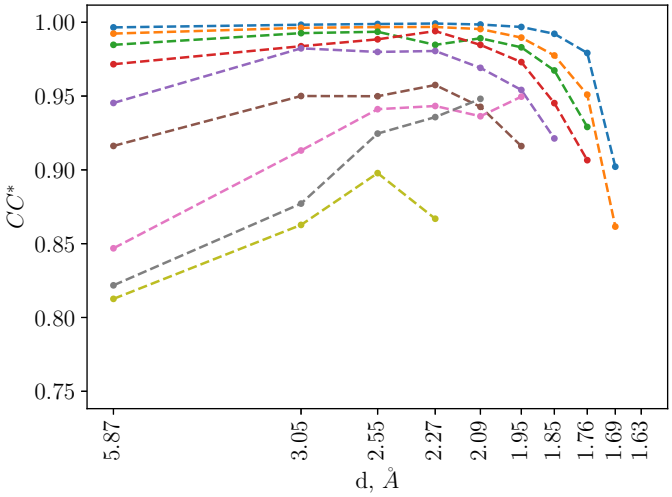


Figure 6.19: Completeness (a), percentage of reflections with $I/\sigma \geq 2$ (b) and CC^* (c) as a function of resolution for different numbers of merged patterns from the lysozyme datasets with 5 μs exposure time.

Sample	Lysozyme								Proteinase K
Dataset	24k, 5 µs	1 chip, 5 µs	3k, 5 µs	1.5k, 5 µs	750, 5 µs	15k, 5 µs	15k, 1 µs	1 chip, 1 µs	2 chips
Space group	$P4_32_12$								$P4_32_12$
Unit cell parameters a, b, c (Å) α, β, γ (°)	79.8(0.2) 79.9(0.2) 38.0(0.1) 90.0(0.1) 90.0(0.1) 90.0(0.1)								68.6(0.2) 68.6(0.2) 104.5(0.5) 90(0.2) 90(0.1) 90(0.2)
Exposure time, µs	5	5	5	5	5	5	1	1	1
Number of merged images	24344	8813	3000	1500	750	14793	14793	4448	1585
Multiplicity	315.5	115.2	39.7	20.9	11.4	192.1	162.8	63.8	23.1
$I/\sigma(I)$	13.66	8.24	6.06	4.2	3.84	10.72	9.12	6.78	4.6
CC^*	0.9973	0.9936	0.9823	0.9668	0.9427	0.9966	0.9973	0.9908	0.9647
R_{split} (%)	6.01	9.71	18.1	24.8	34.8	7.57	7.55	12.67	24.19
Wilson B-factor	19.48	20.18	19.5	19.92	19.56	19.57	20.15	25.28	22.79
Resolution range	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.7 (1.761 - 1.7)	19.37 - 1.95 (2.02 - 1.95)	21.7 - 1.94 (2.009 - 1.94)
Unique reflections	14032 (1354)	13953 (1283)	13698 (1064)	13042 (642)	11637 (221)	14024 (1348)	11142 (1088)	9435 (47)	18492 (1402)
Completeness (%)	99.89 (99.85)	99.27 (93.74)	96.53 (69.42)	91.36 (37.55)	81.27 (11.66)	99.81 (99.26)	95.21 (58.03)	97.94 (81.89)	96.25 (73.54)
Reflections used in refinement	14032 (1354)	13941 (1274)	13676 (1048)	12997 (622)	11569 (204)	14021 (1345)	11140 (1086)	9415 (45)	18430 (1381)
Reflections used for R_{free}	1382 (133)	1378 (130)	1353 (108)	1290 (62)	1138 (17)	1381 (132)	1097 (107)	921 (5)	1767 (133)
R_{work}	0.1486 (0.1943)	0.1615 (0.2591)	0.1654 (0.2826)	0.1863 (0.3089)	0.2152 (0.3066)	0.1510 (0.2043)	0.1560 (0.2028)	0.1687 (0.2871)	0.1721 (0.2345)
R_{free}	0.1742 (0.2193)	0.1859 (0.2811)	0.1944 (0.3302)	0.2062 (0.3405)	0.2417 (0.3074)	0.1796 (0.2260)	0.1830 (0.2429)	0.2022 (0.4103)	0.2293 (0.2844)
RMS(bonds)	0.01	0.008	0.009	0.004	0.003	0.009	0.007	0.007	0.01
RMS(angles)	1.05	0.99	0.99	0.69	0.55	1.28	0.81	0.83	1.31
Ramachandran favored (%)	99	99	99	99	99	99	99	99	97.47
Ramachandran allowed (%)	0.72	0.72	0.72	1.4	1.4	0.72	0.74	0.74	2.53
Ramachandran outliers (%)	0	0	0	0	0	0	0	0	0
Average B-factor	22.28	21.95	22.06	22.24	23.25	21.7	21.57	26.02	24.68

Table 6.2: Data collection and structure refinement parameters for the 9 resulting lysozyme and proteinase K datasets.

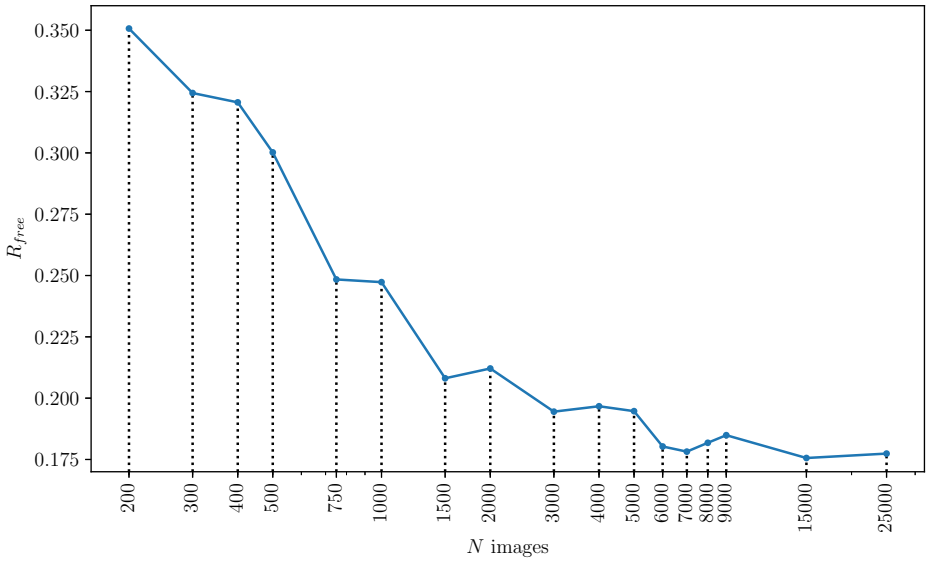


Figure 6.20: R_{free} as a function of the number of merged patterns from the lysozyme datasets with 5 μs exposure time.

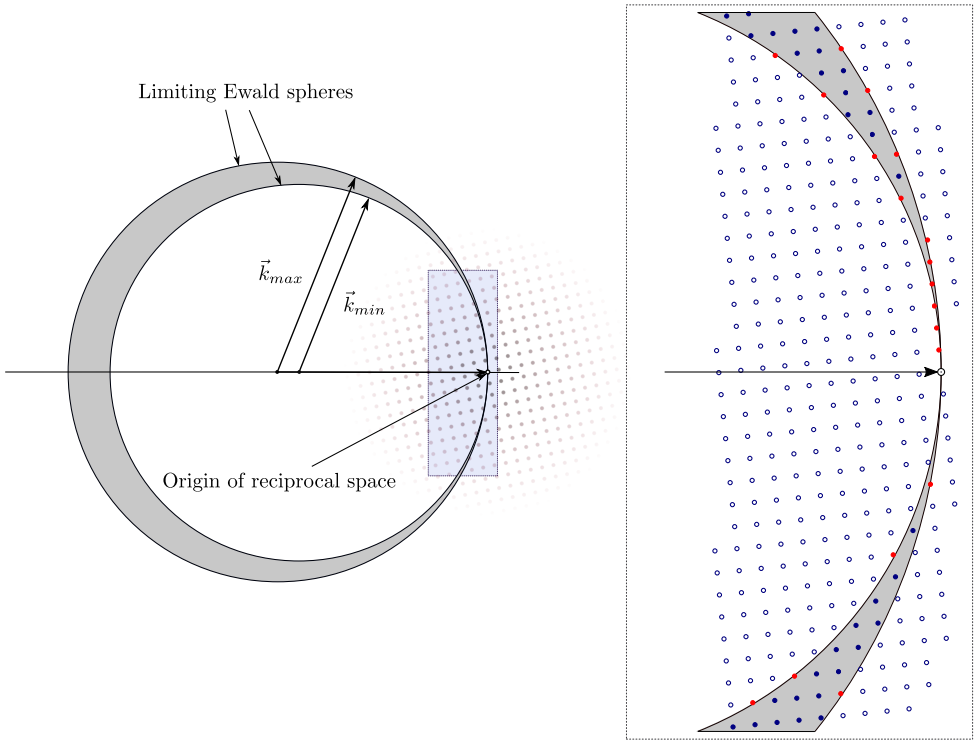


Figure 6.21: In case of the polychromatic X-rays with a wavelengths spread between λ_{min} and λ_{max}, the Ewald sphere becomes a shell (shown in gray) between two limiting spheres with radii of $1/\lambda_{min}$ and $1/\lambda_{max}$. In this case all the reflections lying fully within the shell (shown as filled blue circles) will be fully integrated. The reflections intersecting the edge of the shell and the reflections at lower resolution, where the distance between the limiting spheres is smaller than the reflection diameter (shown as red circles), will be only partially integrated. As the distance between the limiting spheres becomes larger at higher resolution the fraction of fully integrated reflections as well as the total number of diffracted reflections also increase.

is similar to the case of monochromatic radiation. As the resolution increases, so too does the width of the wedge, which eventually is broader than the peak width. At this resolution and higher, reflections are predominantly fully recorded, giving measurements with less variance. These reflections therefore need measurements from fewer patterns to achieve a given confidence.

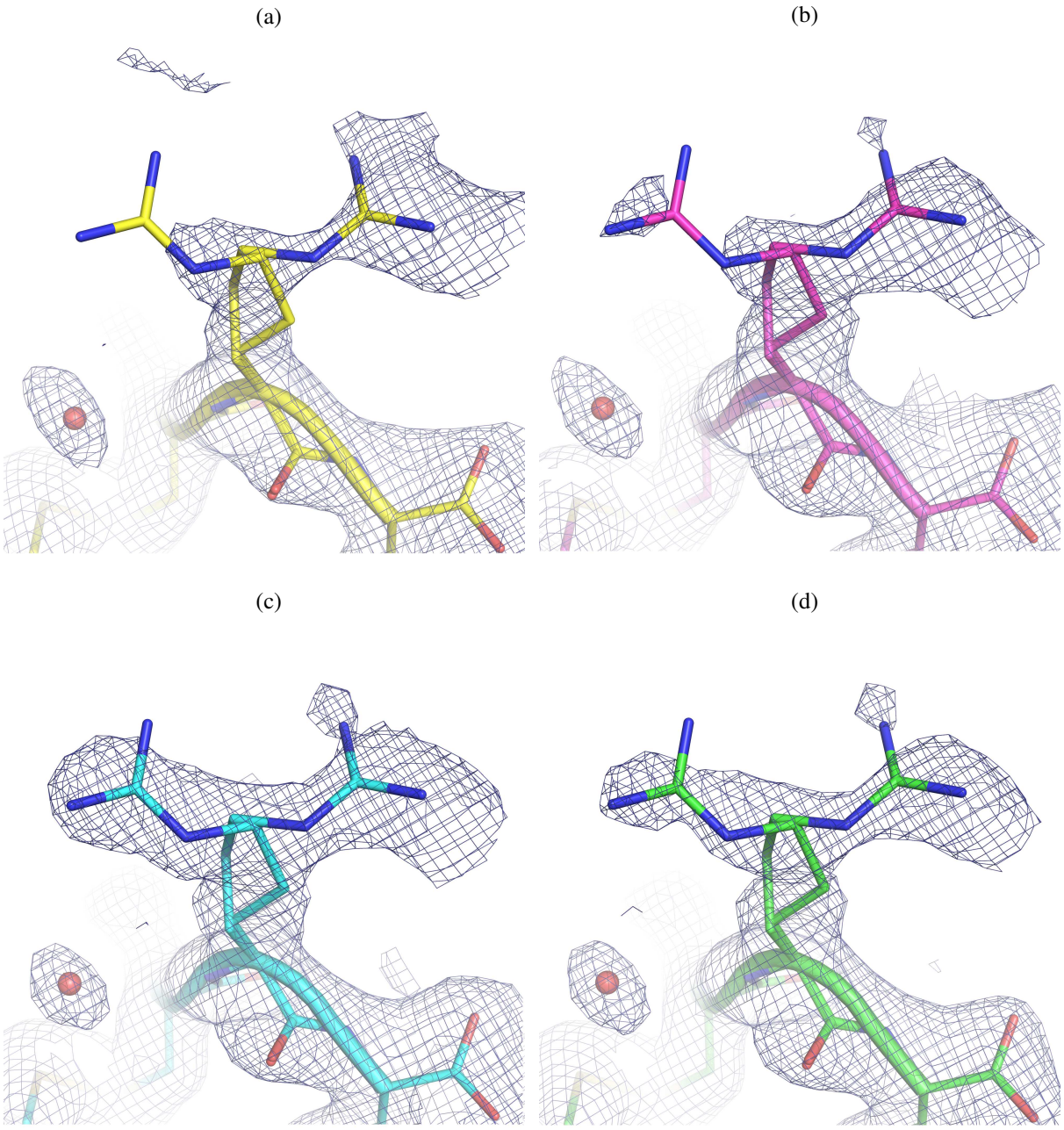


Figure 6.22: 2mFo-DFc electron density maps at 0.7 sigma level showing the poorly defined surface residue arginine 128 generated from datasets consisting of different numbers of merged diffraction patterns: (a) in yellow: 750 patterns, (b) in pink: 1500 patterns, (c) in cyan: 3000 patterns, (d) in green: all patterns (24344). Whereas in (a) and (b) the electron density is ambiguous, the merged datasets from 3000 (c) and all 24344 patterns (d) clearly reveal a second conformation of Arg128.

Example electron density distributions around the Arg 128 residue for different subsets of the 5 µs lysozyme measurements are shown in Fig. 6.22. Whereas for the densities determined from 750 and 1500

merged diffraction patterns only one conformation is visible, merges from 3000 and all 24344 patterns clearly reveal the occupation of a second conformation of residue 128. For other electron density regions of the lysozyme structure a similar trend of additional conformations appearing can be observed. This is consistent with the observation of a relatively strong decrease of the R_{free} values for merges from 250 to 3000 patterns and only a moderate further decrease when more diffraction patterns are considered.

The analysis shows that 3000 single crystal diffraction patterns collected with 2.5% bandwidth X-ray beam are sufficient to obtain a high-quality structure. One important consideration needs to be made here: as described earlier due to the limitations of the experimental setup the detector was moved to the side to be able to record high resolution reflections. Fig. 6.11 illustrates the detector geometry used in the experiment: the structures were refined up to 1.7 Å resolution, the green and red areas demonstrate where the diffraction data of up to 1.7 Å resolution was recorded or lost, respectively. The areas of red and green are almost equal, which means that due to not ideal experimental geometry half of the data is effectively lost. This proportion is even worse at high resolution: in the range between 1.9 and 1.7 Å (shaded area) less than 30% of the diffracted X-rays fall on the detector. Therefore the estimation of the number of required diffraction patterns can be adjusted accordingly: if a bigger detector were available or it were possible to move it closer rather than to the side so that the whole area is covered up to 1.7 Å, the same results could have been obtained with at least 2 times fewer patterns, i.e. 1500 patterns instead of 3000. Given that only one third of the high resolution data is recorded, it is possible that the number would have been even smaller.

6.4.3.2 Unit cell volume variations on the chip

In the method of fixed-target serial crystallography which uses humidified gas stream to prevent the crystals from drying out, the unit cell parameters of the room-temperature crystals may vary depending on their positions on the chip. As shown in Section 5.2, this effect may be very severe leading to large variations in the data quality. To evaluate the effect in this experiment we first compare the variations of the unit cell volume of lysozyme crystals on the chip lys09 to the unit cell volumes of similarly prepared lysozyme crystals also measured at room temperature but in solution, in a liquid jet. For this we use serial crystallographic data collected at the European XFEL by Wiedorn *et al.* [54].

The unit cell volumes of all indexed lysozyme crystals from chip lys09 are found to be normally distributed, with a mean value of 242400 Å^3 and a standard deviation of 1400 Å^3 (Fig. 6.23). This is about 1.5 times the standard deviation obtained from the measurements at the European XFEL, which yielded unit cell volumes that were normally distributed with a mean of 237100 Å^3 and standard deviation of 900 Å^3.

The spatial distribution of the unit-cell volume of crystals on the lys09 chip is plotted in Figure 8. The unit-cell volume varies between 241000 Å^3 and 245000 Å^3, a relative change of about 1.6%, diminishing from the top left corner to the bottom right corner of the chip (Fig. 6.24). Another effect that can be seen in Fig. 6.24b is an oscillation of the unit cell volume in the y direction with a periodicity of two rows and a magnitude of 0.5% of the average unit-cell volume. This magnitude matches the overall gradient experienced in longitudinal direction. Since the total horizontal line scan takes only about 0.3 s but the deceleration at the end of the line, vertical movement and acceleration at the beginning of the next line takes about 1 s, the crystals might have enough time to shrink while the chip stays in the right part of the chamber, where humidity is lower, and then partially recover when the chip is again in the left part, where

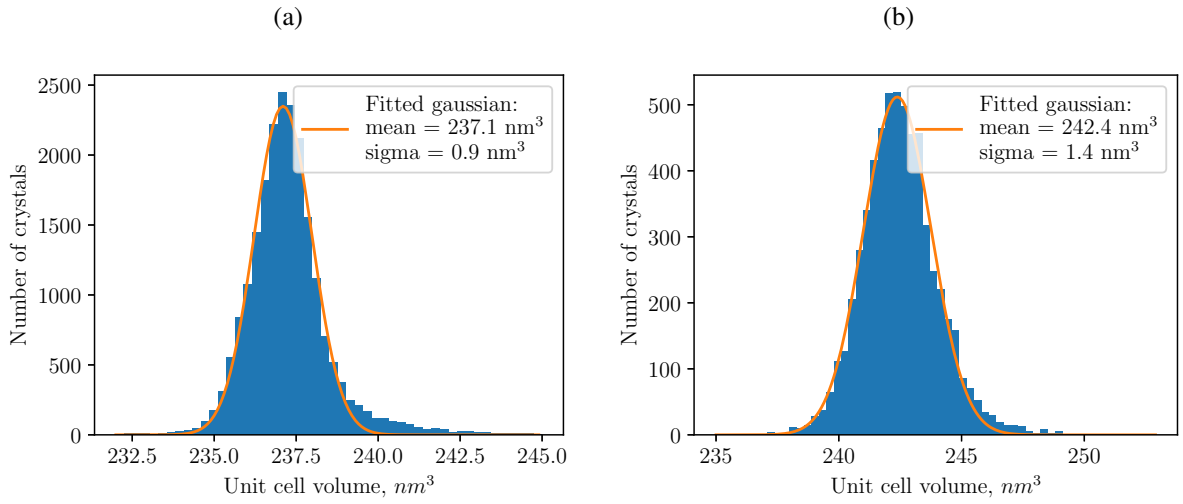


Figure 6.23: Unit cell volume distribution of lysozyme crystals measured in (a) European XFEL serial crystallography experiment using liquid jet [54] and in (b) experiment at beamline ID09 using fixed-target Roadrunner goniometer.

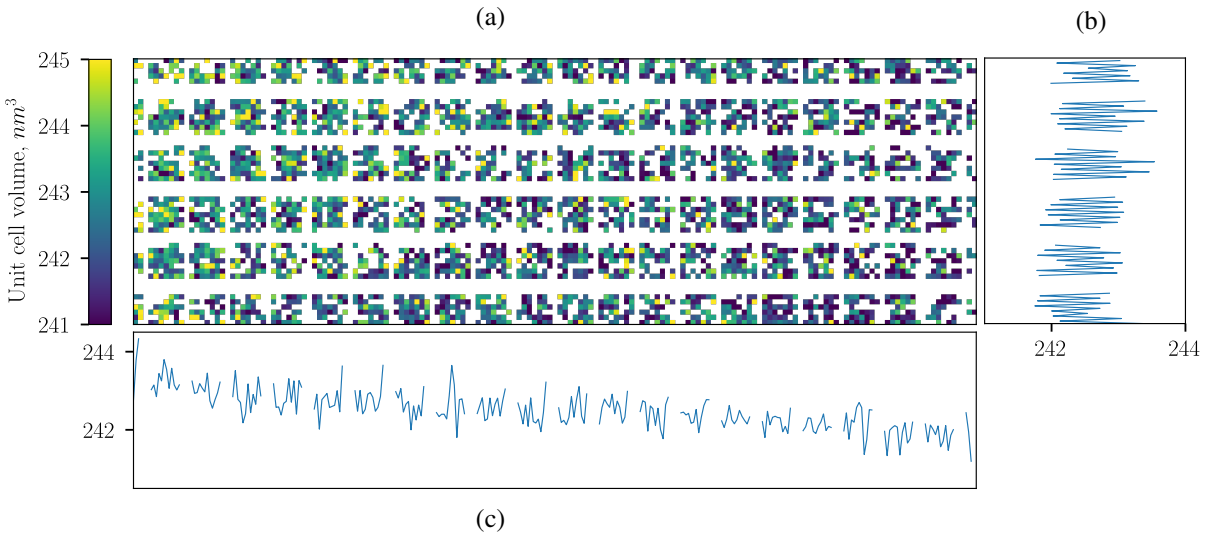


Figure 6.24: (a) Spatial distribution of the unit cell volume of lysozyme crystals on the chip lys09 with dimension of 33 x 12 mm2, averaged in 2x2 bins. (b) Horizontally averaged unit cell volume as function of the vertical position on the chip. (c) Vertically averaged unit cell volume as function of the horizontal position on the chip.

humidity is higher. Another explanation for this variation could be a systematic shift of the stage in the z direction depending on scan direction. However, we verified that this was not the case since the chip stays within the few micrometer depth of focus of the in-line microscope throughout the scan.

The overall relative change of 1.6% is much smaller compared to 18% relative change observed for PSII crystals (Section 5.2), which is likely caused by the improvements in the design of the humidity chamber between the two experiments and by the fact that lysozyme crystals are much less sensitive to variations in the humidity. As the variations are relatively small, unlike the PSII case, we did not observe any systematic changes in the diffraction data quality, and structure refinements carried out with the lysozyme data collected from different areas of the chip did not reveal any significant structural changes.

Despite the systematic changes in the unit cell volume in a scan, the structures of lysozyme determined with the method of fixed target serial crystallography with microsecond exposure times at a synchrotron are of similar quality as the recent structure determination carried out at European XFEL [54, 55] using femtosecond exposure times. It should be noted here, that with conventional crystallography of large lysozyme single crystals much higher resolutions of up to 0.94 Å have been achieved [122].

6.4.4 Discussion

Using the approach of high-speed fixed-target serial crystallography with the polychromatic beam in combination with the new JUNGFRAU integrating pixel detector we were able to collect a complete high-quality diffraction dataset consisting of 3000 patterns in about 30 seconds. The total time for preparation and measurement of one chip, including blotting, mounting of the chip on the goniometer, definition of the scan grid, and data collection was about ten minutes. After aligning the setup and establishing the data collection procedure we were able to measure 10 chips in one hour, which directly translates into at least 10 structure determinations.

In contrast to other sample delivery methods, here the entire membrane area is systematically scanned through the X-ray beam guaranteeing that most of the material on the chip is exposed and contributes to the dataset. With about 10000 crystals per chip with average dimension of 50 μm this corresponds to a total amount of 1.6 mg of sample per chip, which is a fairly large amount of protein for a structural biology project. The main reason for using crystals of this size was to match the X-ray beam size of about 60 μm at ID09 at that time. The applied X-ray doses were as low as 100 Gy for the 1 μs exposure times, which is only about 16 times the LD50 dose for human beings of 6 Gy and more than 5 orders of magnitude less than typical doses of 50 MGy in cryo-crystallography [15, 123]. This highlights the potential of the method for investigations of the undamaged structure of redox-sensitive metallo-proteins at unprecedented low dose levels [124–126].

Using soon available smaller polychromatic X-ray beams will allow a tremendous reduction of the amount of sample required for a pink beam structure determination. Reducing the beam area from 60×60 μm^2 used here to 10×10 μm^2 while retaining the same number of photons per pulse will on one hand increase the dose by a factor of 36 from 100 Gy to 3.6 kGy for a 1 μs exposure, but on the other hand should allow the collection of datasets of similar quality using only 1.6 mg / 36 = 44 μg of sample. X-ray doses of 20 kGy are still well below the room-temperature dose limit of about 300 kGy, and an amount of 44 μg corresponds to one single crystal with dimensions of 330 μm, a crystal size typical for structure determination with X-ray tubes in the lab.

With an achievable time resolution in the microsecond range, and even down to below a nanosecond with single bunches, the method is ideally suited for ligand binding studies and laser pump probe experiments. With crystals of size of several micrometers, which will be measurable at beamlines with smaller beams and higher fluence to what was demonstrated here, diffusion times and hence the achievable time-resolution in such experiments should be in the few millisecond range [29]. Many serial crystallography experiments currently performed at FEL sources tend to use crystals that are large enough to give measurable diffraction signals at high-intensity synchrotron beamlines. This method represents an attractive alternative for such experiments.

In comparison to diffraction experiments with polychromatic X-rays using the full 5% bandwidth of an undulator harmonic, the spectrum produced by the multilayer monochromator is approximately

symmetric and confined within less than 5% from the peak energy, which tremendously facilitates data processing. The problem of the reflections overlap is much less pronounced in the absence of 20% low energy tail of the undulator spectrum. The data can be indexed without prior knowledge of the unit cell parameters using standard monochromatic indexing algorithms with sufficient accuracy to detect relatively small systematic deviations of the unit cell volume due to varying humidity. The pink beam spot prediction and integration procedures established here significantly improve quality of the merged data, allowing to fully exploit the advantages provided by the usage of polychromatic radiation. As they are not inherently limited to the multilayer monochromator spectrum, they can as well be used for the data measured with the full undulator beam (Chapter 7).

CHAPTER 7

Serial crystallography using the full undulator bandwidth

With the peak prediction and integration procedures adapted for broad bandwidth, indexing becomes the main bottleneck of using *CrystFEL* for automatic processing of pink-beam serial crystallographic data (Section 6.3). Recently a new algorithm, *pinkIndexer*, has been developed and integrated in *CrystFEL* by Gevorkov *et al.* [5], capable to reliably index diffraction data acquired with polychromatic X-rays of arbitrary bandwidth. This chapter presents the full data processing pipeline for pink-beam serial crystallography with *CrystFEL* using this new indexing algorithm.

The first part gives an overview of all processing steps starting from individual diffraction patterns to a merged set of reflection intensities and provides a comparison with Precognition [110], using as an example the data collected in the first pink-beam serial crystallography experiment at BioCARS instrument at APS [2]. The other two parts show the pipeline applied to more difficult cases: the data collected at the beamline ID09 at ESRF using the full undulator bandwidth and the same fixed-target setup described in the previous chapter (Section 6.4), and the sparse diffraction data collected from small crystals in the liquid jet, both using JUNGFRAU detector.

7.1 Data processing pipeline for pink-beam serial crystallography with *CrystFEL*

In the first pink beam serial crystallography experiment, described by Meents *et al.* 2017 [2], we collected diffraction data from proteinase K crystals of 10-20 μm in size on a CCD detector using single pulse X-ray exposures of 100 ps with the pink beam of 5% bandwidth (FWHM) with approximately 12.5% low energy tail (Fig. 7.1). Thanks to using the low-background fixed-target setup and the high quality crystals with a small degree of mosaicity, the collected diffraction patterns contained on average more than 300 diffraction spots with only small elongation in the radial direction. As a result it was possible to process the data with Precognition software [110]. From 1011 diffraction patterns classified as hits, 140 were indexed and 59 were used for the structure refinement. In this section the same proteinase K data is processed with *CrystFEL* using *pinkIndexer* algorithm. The new analysis reveals a second crystal conformation with a slightly different unit cell parameters and $\sim$0.2 Å higher resolution, which was

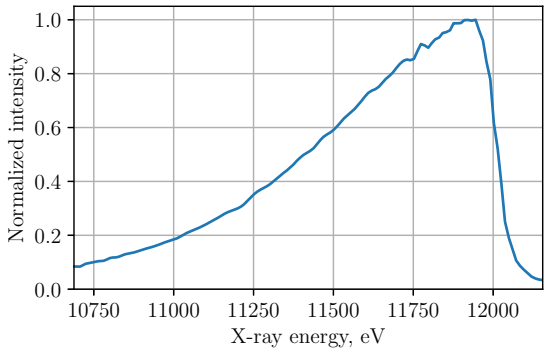


Figure 7.1: Measured X-ray spectrum at BioCARS beamline.

previously overlooked due to Precognition requirement of the prior knowledge of the unit cell.

7.1.1 Indexing and unit cell scaling

Similar to Precognition, *pinkIndexer* requires the knowledge of the unit cell parameters for indexing. However, they do not need to be defined precisely, as *pinkIndexer* also performs a refinement procedure. As was shown in Section 6.3, monochromatic indexing algorithms in *CrystFEL* can index up to 80% of polychromatic diffraction patterns obtained using the full undulator bandwidth. Although the found indexing solutions are not accurate enough to use for integration, they should be sufficient to provide the first estimation of the unit cell parameters for *pinkIndexer*. To test this hypothesis, the proteinase K data was indexed using *MOSFLM*, *asdf* and *DirAx*. As the data was collected using polychromatic beam but the indexers assume it was monochromatic, the indexing results would depend on the X-ray energy set for indexing. In this case it was set to 11940 eV - the mode energy in the spectrum. Choosing other energies within 5% of the mode energy produced similar unit cell parameter distributions and indexing rates. The resulting cell parameter distributions are shown in Fig. 7.2a. Two things need to be noted here. First, the distributions of 330 unit cells (which is approximately 25% of the found hits) are narrow enough to give a good estimation of the parameters. Second, the distribution of the c parameter clearly shows two separate sharp peaks which implies the crystals with two different unit cells are present in the dataset. The parameters of these two unit cells are given in Table 7.1.

Dataset	a, Å	b, Å	c, Å	α,°	β,°	γ,°
1	67.8	67.8	103.7	90	90	90
2	67.7	67.7	107.1	90	90	90

Table 7.1: Unit cell parameters of two crystal types of proteinase K obtained by indexing polychromatic diffraction data with monochromatic indexing algorithms.

Once the unit cell parameters were determined, the data was indexed using *pinkIndexer*. The average parameters of the two obtained unit cells ($a = b = 67.75$Å, $c = 105.45$Å, $\alpha = \beta = \gamma = 90°$) were given as an input to *pinkIndexer*. From 1424 patterns 1015 were successfully indexed: 571 as single, 267 as double and 177 as triple or more crystal hits. Fig. 7.2b shows the resulting unit cell parameter distributions. The separation in the c axis although still clearly visible, became slightly less distinct. This can be explained by the fact that *pinkIndexer* does not use the shape of the X-ray energy spectrum but

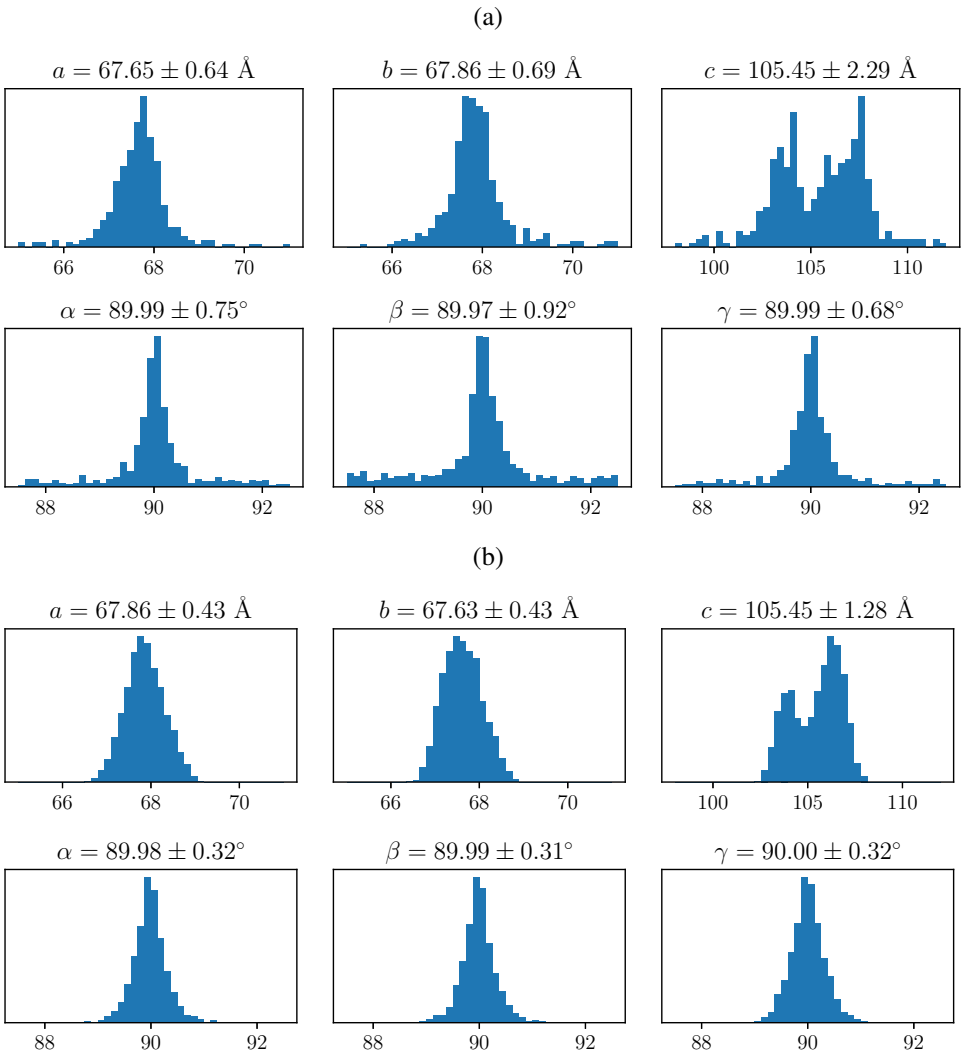


Figure 7.2: Distributions of the unit cell parameters obtained by indexing proteinase K polychromatic diffraction data using (a) monochromatic indexing algorithms and (b) *pinkIndexer*.

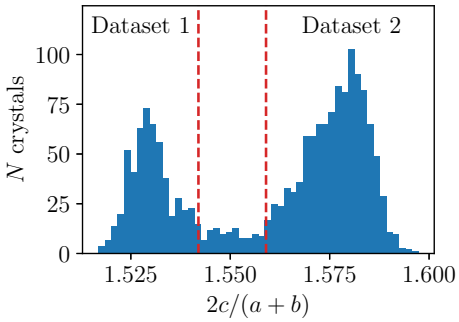


Figure 7.3: Distribution of $2c/(a+b)$ ratio, dashed red lines show where the dataset was split in two. The crystals between the lines were discarded.

only its range, i.e. the minimum and maximum energies. Since the peaks in the diffraction pattern can be produced by any interval inside the whole energy range, during indexing and refinement of the unit cell *pinkIndexer* cannot determine the absolute values of the unit cell lengths but only their ratios. When the

real unit cell is slightly different from the given unit cell, it refines the ratios of c/a and c/b to be close to the real ones while keeping the absolute values of all three parameters close to what was given as an input. Therefore, in order to distinguish two crystal types, the ratio $2c/(a+b)$ is plotted on Fig. 7.3. Based on this distribution, which shows much clearer separated peaks corresponding to each population, the dataset was split in two. The crystals from the middle region between the dashed red lines were discarded from the subsequent processing.

After the same-type crystals were sorted into corresponding datasets, it is necessary to determine the absolute values of the unit cell parameters. This can be achieved by exploiting the fact that the probability to detect a Bragg peak produced by the X-rays of certain energy is proportional to its spectral intensity. When the peak corresponding to the reflection hkl is detected at the scattering angle 2Θ, the X-ray energy which produced this peak can be calculated using the Bragg's law (Eqn. 2.18) as following:

$$k_0 = \frac{|h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*|}{2\sin\Theta} \tag{7.1}$$

where $\boldsymbol{a}^*$, $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$ are the reciprocal lattice basis vectors.

Using the *pinkIndexer* indexing solutions and the spot prediction algorithm described previously (Section 6.4.2.1), for each detected peak the matching predicted reflection is found. Then, using the peak detector position and the hkl indices of the corresponding reflection, the central X-ray energy contributing to the peak is calculated. The distributions of the X-ray energies contributing to all detected peaks in all diffraction images in datasets 1 and 2 are plotted on Fig. 7.4a and 7.4b respectively. It can be seen that the distributions are similar in shape to the measured X-ray spectrum $I_s(k)$, but shifted along k by a different factor. By fitting the measured spectrum, $I_s(sk)$, to the peak energy distribution, the unit cell scaling factor s can be found. Fig. 7.4c and 7.4d show the resulting unit cell distributions for both datasets after the scaling is applied.

Since Precognition requires more precisely defined unit cell parameters and the values used for indexing were $a = b = 68.3$ Å, $c = 108.3$ Å, which are very close to the parameters of the dataset 2 obtained here, the only proteinase K structure refined in Meents *et al.* 2017 [2] was the crystal structure of dataset 2. The second unit cell was not found, which means that at least half of the information acquired in the experiment was overlooked. This was probably one of the reasons why only 140 from 1011 found hits were successfully indexed. Another reason was the large portion of multiple hits, which is a common feature of fixed-target serial crystallography experiments (Section 3.5.1). Even when Precognition was able to index one lattice in a multiple crystal diffraction pattern, the pattern was discarded during the manual inspection. *pinkIndexer* on the other hand was successful in determining up to 5 lattices in one pattern. As a result, instead of one dataset consisting of 59 single crystal diffraction patterns, by using *pinkIndexer* in combination with the prior unit cell determination with monochromatic indexers it was possible to retrieve two different datasets, of 616 and 816 crystals respectively, from the same diffraction data.

7.1.2 Integration of reflection intensities

Once the diffraction pattern is indexed, reflection intensities have to be integrated and scaled according to the spectrum, i.e. divided by the spectral weight of the contributing X-ray energy as described in Section 6.4.2.2. Accurate integration using three-rings integration method (Section 4.1.4) requires accurate peak prediction. Accurate scaling requires accurate determination of the X-ray energy contributing to the

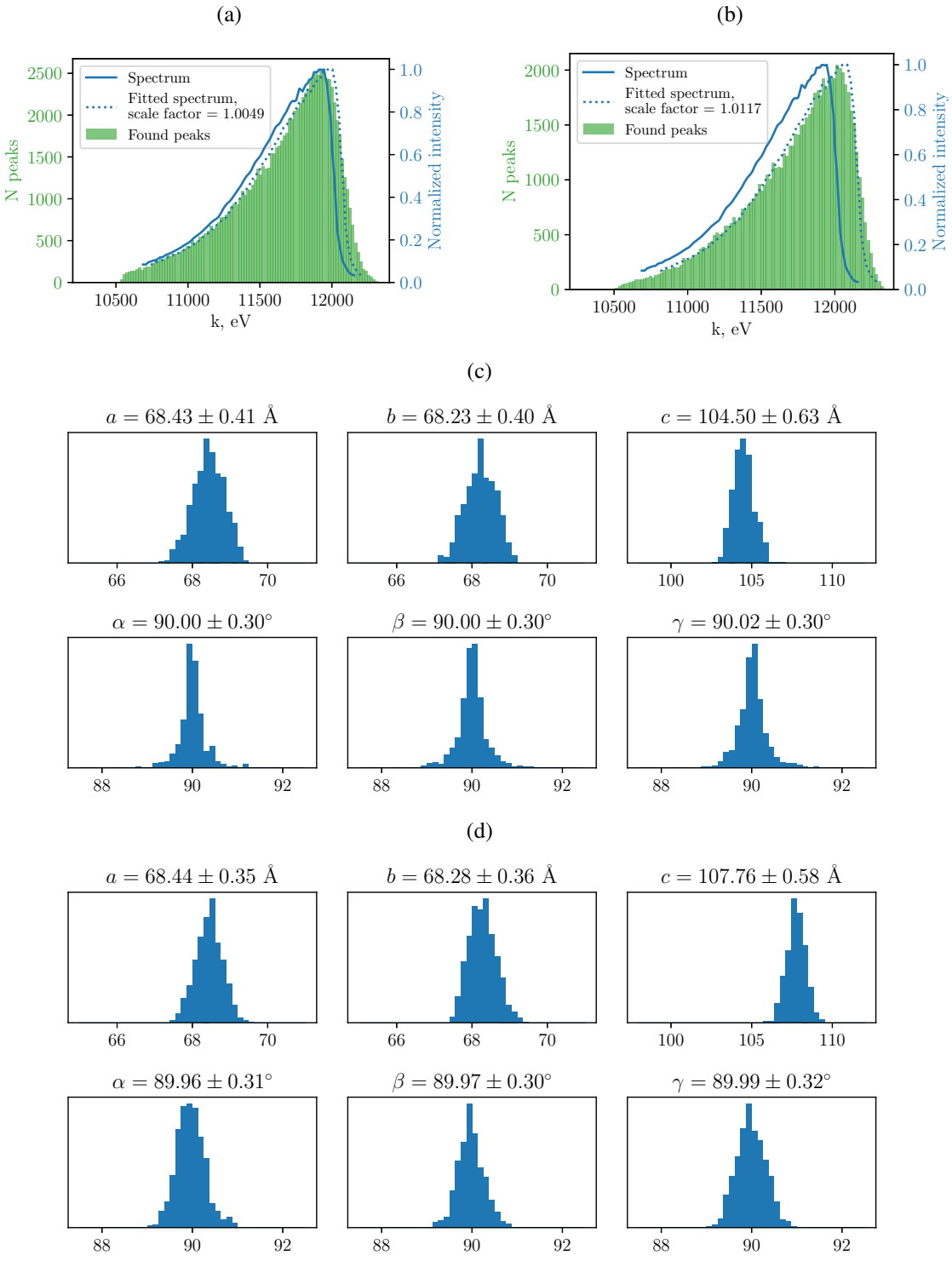


Figure 7.4: (a, b) Distributions of the X-ray energies producing each detected peak in datasets 1 and 2 respectively. The measured spectrum is shown in blue. By fitting the spectrum to the peak energies distribution, the unit cell scaling factor can be found. (c) and (d) show the resulting distributions of the unit cell parameters for two respective datasets after the scaling is applied.

reflection, which can only be found after the scaling of the unit cell described above. Indexing pink-beam diffraction data with *pinkIndexer* is by far the most computationally expensive step of pink-beam data processing therefore it is desirable to avoid re-indexing the data after the unit cell scaling. This can be

achieved by separating the integration and intensity scaling procedures.

As it follows from the necessity of the unit cell scaling procedure, in the first step after indexing the peaks have to be predicted in the broader range of energies than the range defined by the spectrum. For example, were the peaks in Fig. 7.4b only predicted up to 12.1 keV, which is a high energy cut-off in the measured spectrum, the large portion of the peaks would not have been predicted and the scaling factor would not have been determined correctly. In the end, that would not only lead to incorrect values of the unit cell parameters, but also errors in the reflection intensities scaling to the spectrum. According to Eqn. 7.1, scaling proportionally all three unit cell dimensions is equivalent to scaling the beam energy. If the peaks are predicted in the broader range of energies their predicted positions will not change with the unit cell scaling. Therefore, while the three-rings integration can be performed together with the spot prediction right after the indexing in *indexamajig*, the scaling of reflection intensities to the spectrum has to be applied separately, after the scaling of the unit cell.

7.1.3 Merging intensities

Fig. 7.4a and 7.4b again confirm that the peaks detected in the pink-beam diffraction patterns are indeed sampled by the whole X-ray energy range. However, the reflections coming from the X-ray energies in the tail of the spectrum with the lower spectral intensity $I_s(k)$ will accordingly have lower intensity and, therefore, lower signal-to-noise ratio. This problem will be even worse with the CCD detectors in particular, as the weak reflections will be smaller and, when integrated over a large radius necessary for integration of the stronger reflections, their intensities will be lost in the noise. Additionally, the higher scaling factors proportional to $1/I_s(k)$ will be applied to the intensities of these weaker reflections amplifying the noise, which may lead to larger errors in the merged intensities. Therefore, it can be beneficial to discard reflections sampled by the low-intensity tails of the spectrum with the normalized spectral intensity I_s below a certain cut-off I_{co} before merging. This is illustrated in Fig. 7.5, which shows the histogram of the peaks in dataset 1 highlighting the region which would be used for merging in the case of I_{co} = 0.4.

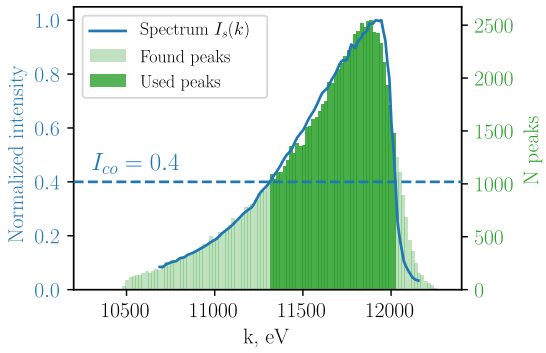


Figure 7.5: Distribution of the detected peaks in dataset 1. The highlighted region represents the peaks which would be used for merging in the case of spectral intensity cut-off I_{co} = 0.4.

To investigate this assumption, the intensities of the dataset 1 were scaled and merged after applying the spectral intensity cut-off, I_{co}, varying from 0.1 to 0.9. The overall $CC_{1/2}$ of the obtained merges as well as the average I/σ ratio are shown in Fig. 7.6a. While $CC_{1/2}$ shows generally downward trend with the increasing cut-off, the average I/σ demonstrates a clear maximum for I_{co} values in the range between

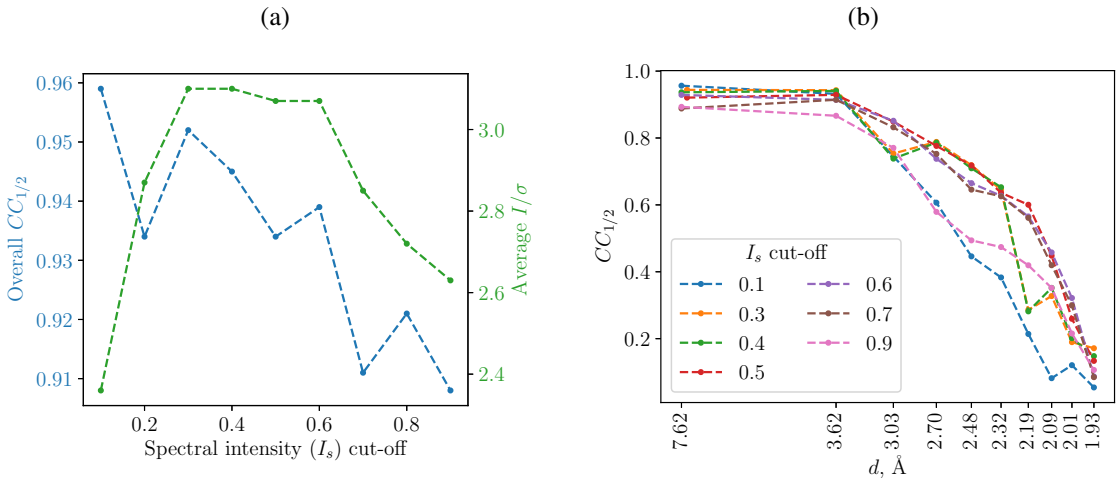


Figure 7.6: (a) Dependence of the overall $CC_{1/2}$ and the average I/σ on the spectral intensity cut-off used for merging of dataset 1. (b) $CC_{1/2}$ as a function of the intensity cut-off and resolution for dataset 1.

0.3 and 0.6. As can be seen on Fig. 7.6b, although $CC_{1/2}$ reaches its maximum at the low resolution with the $I_{co} = 0.1$, the best statistics at the high resolution is as well achieved with the cut-off values around 0.5, which corresponds to the spectral bandwidth of only 5%. If we define the dataset resolution as the resolution where $CC_{1/2}$ = 0.2, then increasing the cut-off from 0.1 to 0.5 improves the resolution from 2.2 to 2.0 Å.

An alternative to applying the constant cut-off to all reflections is to vary it with resolution. As the best $CC_{1/2}$ values at the low resolution are obtained with $I_{co} = 0.1$ and at high resolution with $I_{co} \simeq 0.5$, varying cut-off between these values with resolution might increase the overall $CC_{1/2}$. Below, I propose an approach to derive the resolution-dependent spectral intensity cut-off value from the data itself.

7.1.3.1 Resolution dependent energy cut-off

As the reflection intensities are in general decreasing with resolution, the peaks produced by the low-intensity tails of the spectrum are less likely to be detected at higher resolution. This effect is illustrated in Fig. 7.7a, which shows the distributions of the X-ray energies contributing to the found peaks in dataset 1 in three different resolution ranges. As the distribution becomes narrower with higher resolution, its width at each resolution can be used to determine the energies range for merging. The value $I_{co} = 0.5$ for high resolution reflections can be obtained if the energy is cut off at the 10th percentile of the found peaks. Thus the resolution-dependent intensity cut-off is determined from the X-ray energy at the 10th percentile of the found peaks in each resolution range. The dependence $I_{co}(d)$ for both datasets is shown on Fig. 7.7b.

A comparison of resulting $CC_{1/2}$ for both datasets merged with the constant $I_{co} = 0.5$ and resolution-dependent cut-off shows the improvement of both overall $CC_{1/2}$ and the average I/σ when the resolution-dependent cut-off is used (Fig. 7.8). Additionally, another important observation can be made here: if we compare $CC_{1/2}$ of two datasets, it can be seen that dataset 1, although containing 25% fewer crystals, has in fact approximately 0.2 Å higher resolution than dataset 2.

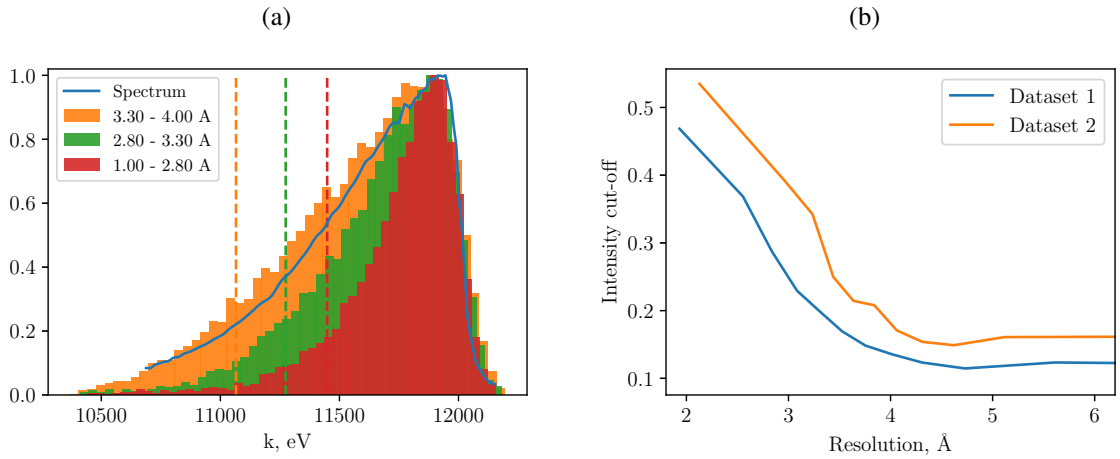


Figure 7.7: (a) Distributions of the X-ray energies contributing to the detected peaks in dataset 1 in three different resolution ranges. The dashed lines of the corresponding colors represent the energy cut-off at 10th percentile of the detected peaks. (b) Resolution-dependent intensity cut-off for two datasets based on 10th percentile of the detected peaks.

7.1.3.2 Discarding overlaps

As mentioned earlier, the low energy tail in the full undulator X-ray spectrum leads to a large number of overlapping reflections (Fig. 6.2). Therefore, the last necessary step before merging reflections with *partialator* is to discard these overlaps. This is achieved by rejecting all reflections which have at least one neighbor within a distance of two inner radii used for 'three-rings' integration (Fig. 4.2).

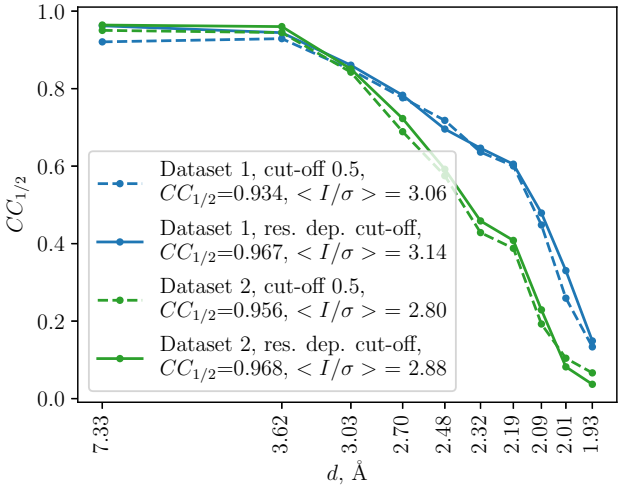


Figure 7.8: $CC_{1/2}$ as a function of resolution for two datasets merged with and without resolution-dependent energy cut-off

7.1.4 Lorentz factor correction

Apart from scaling of the integrated intensities with the spectral weight of the contributing X-ray energy, another scaling factor, called Lorentz factor, has to be applied. Consider again the illustration of the Ewald sphere in case of polychromatic beam shown in Fig. 6.1. Since the distance between the limiting Ewald

spheres increases with scattering angle, the fraction of the whole energy range giving rise to diffraction is much larger for low resolution reflections than for high resolution ones. As a result, reflection intensities will fall off much faster with resolution compared to monochromatic case, which will lead to largely overestimated Wilson B-factors. To account for this, a Lorentz factor, which is inversely proportional to the distance between the limiting Ewald spheres, is used.

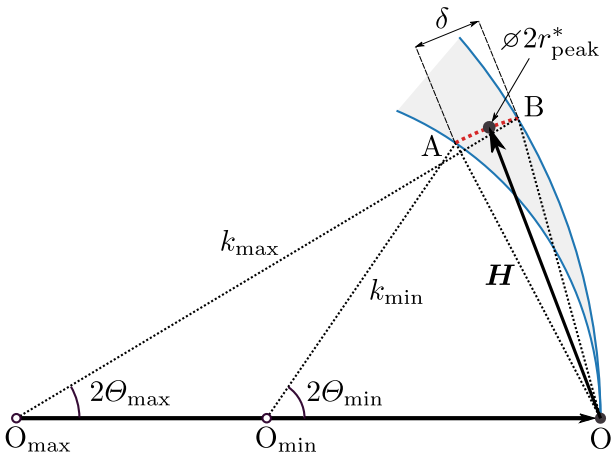


Figure 7.9: The fraction of the whole energy range $\Delta k = k_{\mathrm{max}} - k_{\mathrm{min}}$ contributing to the reflection depends on the reflection resolution H. It can be estimated for the reflection $\boldsymbol{H}$ as $2r^*_{\mathrm{peak}}/\delta$, where r^*_{peak} is the peak profile radius.

Fig. 7.9 illustrates how Lorentz factor is calculated for reflection $\boldsymbol{H}$. The distance δ between the limiting Ewald spheres at the resolution H can be calculated as

$$\begin{aligned} \delta(H) &= H \cdot \angle\mathrm{AOB} \simeq H(\Theta_{\mathrm{min}} - \Theta_{\mathrm{max}}), \\ \Theta_{\mathrm{min}} &= \arcsin\left(\frac{H}{2k_{\mathrm{min}}}\right), \\ \Theta_{\mathrm{max}} &= \arcsin\left(\frac{H}{2k_{\mathrm{max}}}\right). \end{aligned} \tag{7.2}$$

Lorentz factor is then estimated as a ratio of the distance δ between the limiting Ewald spheres to the diameter of the reciprocal lattice peak $2r^*_{\mathrm{peak}}$. For the low resolution reflections with the diameter greater than $\delta(H)$ Lorentz factor is set to 1:

$$L(\boldsymbol{H}) = \begin{cases} \delta/2r^*_{\mathrm{peak}} & \text{if } 2r^*_{\mathrm{peak}} < \delta, \\ 1 & \text{if } 2r^*_{\mathrm{peak}} > \delta. \end{cases} \tag{7.3}$$

As Lorentz factor doesn't depend on the lattice orientation but only on reflection resolution, it can be applied to already merged intensities. Fig. 7.10 shows Wilson plots for both datasets before and after Lorentz factor correction: with the correction applied B-factors decrease by a factor of about 2.5.

7.1.5 Structure refinements

For the final comparison of the proposed merging strategies the structures of datasets 1 and 2 were refined using PHENIX [120]. Because dataset 2 has similar unit cell parameters to the data processed with Precognition, the structure published in Meents *et al.* 2017 [2] was used as a start model for the

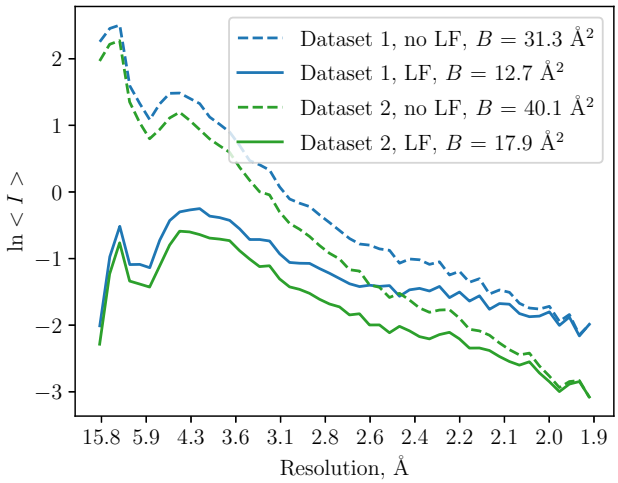


Figure 7.10: Wilson plot of two datasets with and without Lorentz factor scaling.

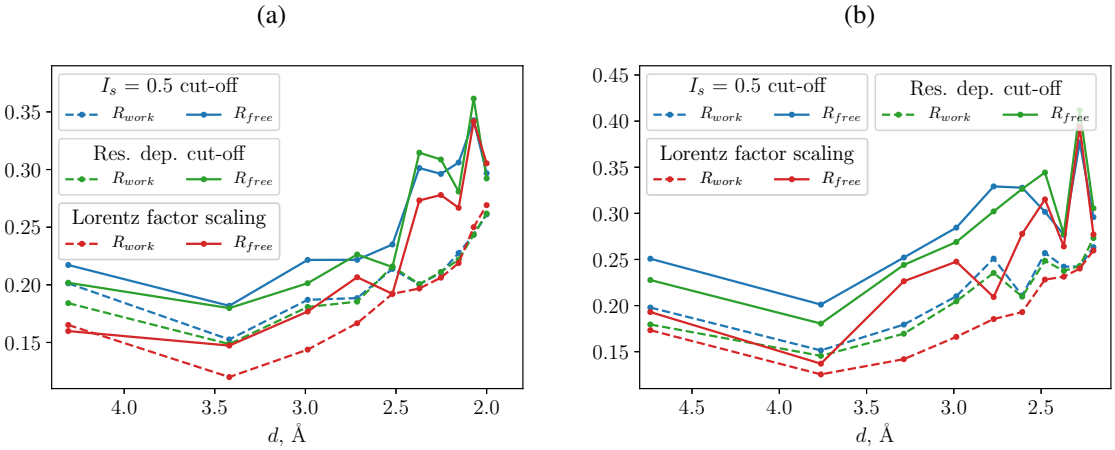


Figure 7.11: R_{work} and R_{free} as functions of resolution for (a) dataset 1 and (b) dataset 2, merged using constant spectral intensity cut-off $I_{co} = 0.5$ (blue), resolution-dependent cut-off (green) and resolution-dependent cut-off + Lorentz factor scaling (red).

refinement of dataset 2. For dataset 1, proteinase K structure obtained in the ESRF experiment described previously (Section 6.4) was used. Table 7.2 shows data processing and structure refinement parameters. The resulting R_{work} and R_{free} for both datasets are plotted in Fig. 7.11. Both R-factors are clearly improved by using the resolution-dependent cut-off and Lorentz factor scaling. It is further confirmed that the dataset 1 is of better quality compared to dataset 2 as, although refined to 0.2 Å higher resolution, it demonstrates lower R-factors.

Proteinase K structure refined from the data processed with Precognition, however, shows much better R-factors. This is unsurprising: in Precognition reflection intensities are determined by analytical peak profile fitting taking into account the point spread function of CCD detector, which is much more accurate compared to three-rings integration method in *CrystFEL*. Furthermore, the diffraction patterns processed with Precognition are inspected visually and hand-picked for merging. Besides, with Precognition only reflections with $I/\sigma(I) > 3$ were merged and used in the refinement. For a more accurate comparison, the structure of dataset 2 processed with *CrystFEL* was refined using only reflections with $I/\sigma(I) > 1.7$ (see

Dataset	1			2				Meents *et al.* 2017
Software	*CrystFEL*							Precognition
N hits	1424							1011
N indexed patterns / crystals	1015 / 1736							140 / 140
N merged	616			816				59
Space group	$P4_32_12$			$P4_32_12$				$P4_32_12$
Unit cell a, b, c, Å	68.33 68.33 104.5			68.35 68.35 107.8				68.3 68.3 108.3
Resolution range, Å	48.32 - 2.0 (2.072 - 2.0)			57.72 - 2.2 (2.279 - 2.2)				44.1 - 2.21 (2.35 - 2.21)
Merging strategy	$I_s = 0.5$ energy cut-off	Resolution dependent cut-off	+ Lorentz factor scaling	$I_s = 0.5$ energy cut-off	Resolution dependent cut-off	+ Lorentz factor scaling	+ $I/\sigma(I)$>1.7 cut-off	$I/\sigma(I)$>3
Completeness	99.7(99.5)			99.7(99.6)			60.7 (34.0)	61.9 (26.1)
Wilson B, Å^2	31.9	31.3	12.7	39.7	40.1	17.9	12.4	0.02
$CC_{1/2}$	0.934	0.967	0.967	0.956	0.968	0.968	0.968	n/a
Reflections N_{work}/N_{free}	17321 / 742	17331 / 777	17321 / 778	13512 / 555	13530 / 578	13529 / 578	8120 / 439	8342 / 793
R_{work}	0.195 (0.261)	0.189 (0.262)	0.179 (0.269)	0.201 (0.263)	0.192 (0.273)	0.180 (0.260)	0.154 (0.172)	0.154 (0.172)
R_{free}	0.234 (0.397)	0.227 (0.292)	0.215 (0.305)	0.265 (0.296)	0.254 (0.305)	0.23 (0.277)	0.193 (0.233)	0.196 (0.220)
RMS bonds, Å	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003
RMS angles, °	0.46	0.46	0.46	0.48	0.53	0.55	0.58	0.65
Average B, Å^2	37.7	34.6	17.1	42.4	42.3	20.5	12.1	6.16

Table 7.2: Data processing and structure refinement parameters for two proteinase K datasets processed with *CrystFEL* and merged using different strategies. The last two columns provide a comparison between the structures obtained with *CrystFEL* and Precognition [2].

the last two columns of Table 7.2). This value was chosen in a way to match the overall completeness of two datasets, which in the end was about 61% in both cases. With such $I/\sigma(I)$ cut-off the structure of dataset 2 demonstrated almost identical R-factors: R_{work}/R_{free} = 0.154/0.193 compared to 0.154/0.196 achieved with Precognition. This result proves that the advantage of semi-manual processing and more advanced integration in Precognition can be overcome by indexing and merging larger number of diffraction patterns achievable with *CrystFEL* thanks to the effectiveness of *pinkIndexer* and lack of requirement to visually evaluate the whole array of collected diffraction images.

7.1.6 Conclusion

Based on the analysis presented above, the following pink-beam serial crystallography data processing pipeline, shown schematically in Fig. 7.12, is developed:

1. If they are unknown, determine the unit cell parameters using monochromatic indexing algorithms in *indexamajig*.

2. Index diffraction patterns with *pinkIndexer* and predict reflections within a broader energy range than the range defined by the spectrum, using the pink-beam prediction procedure established in the previous section (6.4).

3. Calculate the X-ray energy producing each found peaks which was successfully predicted, fit the resulting energy distribution to the measured spectrum to determine the unit cell scaling factor, scale the unit cell accordingly.

4. Split the energies contributing to the found peaks into resolution bins, from the energy distribution in each bin determine the energy or spectral intensity cut-off value.

5. Apply the spectral intensity cut-off and divide the integrated reflection intensities by the spectral intensity of the central contributing X-ray energy.

6. Merge reflection intensities using *partialator*.

7. Apply Lorentz factor correction to the merged intensities.

* Use other *CrystFEL* programs, such as *geoptimiser* and *ambigator*, as intermediate steps when necessary.

Using this pipeline to process proteinase K dataset, collected using the full undulator bandwidth, it was possible to achieve results comparable to Precognition. Furthermore, due to the capability to determine the unit cell parameters without the need to perform monochromatic measurements, crystals with a different unit cell were discovered in the same dataset, which resulted in obtaining a second crystal structure of higher resolution and of better quality.

In addition to the advantage of automatic and effective data processing, which allowed to exploit 70% of the acquired data compared to 6% with Precognition, *CrystFEL*, being a free open-source software, is much more flexible and easily adjustable. It is not limited to CCD detectors, but, on the contrary, is initially developed to handle modern integrating detectors often assembled from multiple panels, which are much better suited for serial crystallography measurements due to their better performance and faster data acquisition rates. The possibility to refine geometry of a multi-panel detector and exclude bad detector regions from the processing gives it further advantage. Finally, thanks to the capabilities of *pinkIndexer* algorithm, it doesn't require strong diffraction with large numbers of visible peaks forming elliptical patterns for indexing. Even as few as 15 detected peaks can be sufficient to find the correct crystal orientation, making it perfect for serial crystallography as it usually involves measurement of very small crystals.

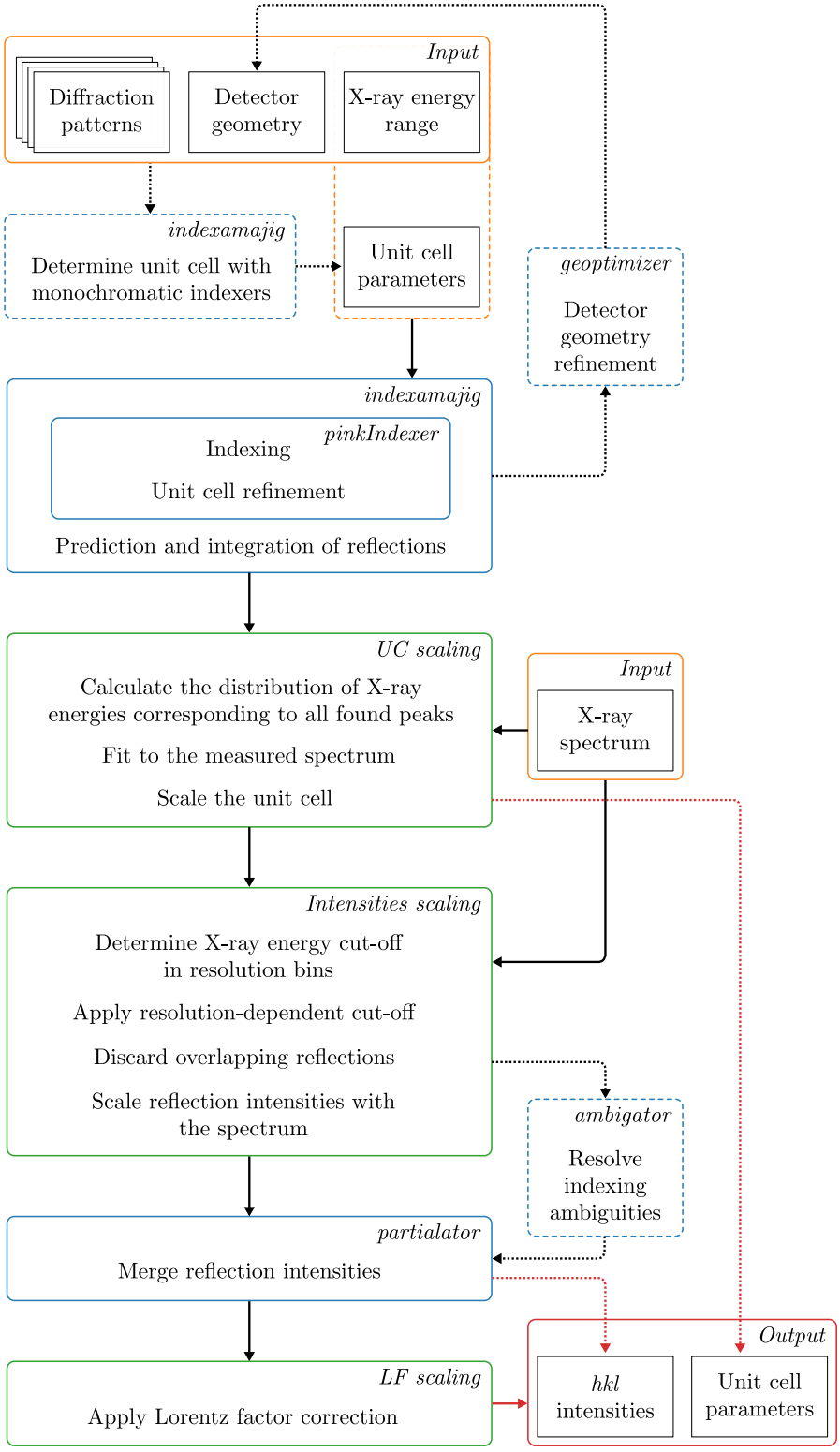


Figure 7.12: Pink-beam serial crystallography data processing pipeline with *CrystFEL*. *CrystFEL* programs are shown as blue boxes, green boxes are currently implemented as Python scripts.

7.2 Fixed-target experiment at ID09

During the 1 kHz fixed-target serial crystallography experiment at beamline ID09 at ESRF, described in detail in Section 6.4, we collected diffraction data from two lysozyme chips using the full undulator bandwidth without the multilayer monochromator. The X-ray spectrum produced by the undulator (Fig. 6.10) has a bandwidth of 5.1% (FWHM) and a maximum at 15.2 keV. The overall energy range spread between ∼11 and 16 keV is about 10 times broader than the whole range produced by the multilayer monochromator. The data was collected with 1 µs exposure and the beam size at the sample position was 100×100 µm^2. The number of photons was 3×10^9 per 1 µs exposure which was 17% less compared to 5 µs exposure with the multilayer monochromator.

As mentioned earlier, the background scattering in pink-beam experiments is typically much stronger than in monochromatic experiments. Because in our case experimental setup was not optimized for measurements with the full undulator beam, the background was significantly higher compared to the measurements with the multilayer monochromator. The example diffraction pattern in Fig. 7.13a shows overall higher background compared to 1 µs diffraction pattern in Fig. 6.15b. The difference is especially noticeable in the low resolution region around the beamstop and at the high resolution on the bottom detector panel. Additionally, the larger beam size resulted in a number of parasitic scattering streaks, which had to be masked together with the low resolution region during the data processing. The resulting mask is shown in red in Fig. 7.13b. As a consequence, intensities of the low-resolution reflections below 17 Å were not measured at all and the overall resolution of the data was lower compared to the data collected with the multilayer monochromator.

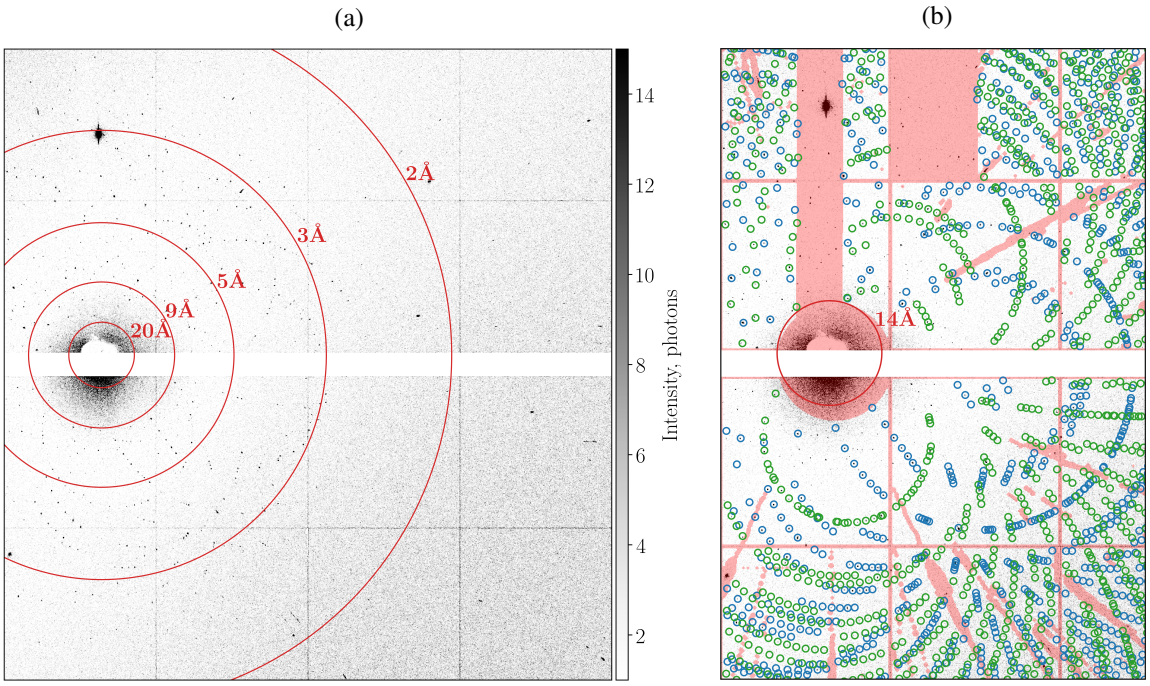


Figure 7.13: (a) Example double crystal diffraction pattern collected using full undulator beam at ID09. Resolution rings shown in red correspond to 15.2 keV - the peak energy in the spectrum. (b) Blue and green circles show predicted Bragg spot positions for two found crystal lattices.

Despite the far from optimal quality of the data, there are several reasons which make this dataset a good test case for the pink-beam processing pipeline with *CrystFEL*. First, the data was collected using the

novel JUNGFRAU detector and, therefore, can not be processed by other Laue crystallography software only compatible with CCD detectors. Secondly, since the experimental conditions were almost identical to the measurements with the multilayer monochromator, the two methods can be compared directly. Finally, as shown in the previous section, it is not possible to determine absolute unit cell parameters with *pinkIndexer*. Hence it is important to see if the humidity induced variations of the unit cell on the chip can be accounted for within the proposed pipeline.

This section describes my attempts to deal with the fluctuations of the unit cell and provides the analysis of the data quality dependence on the number of merged patterns compared to the data collected with the multilayer monochromator.

7.2.1 Indexing and integration. Dealing with unit cell variations on the chip

The larger beam size of 100×100 μm^2 compared to 60×60 μm^2 with the multilayer monochromator resulted in a larger fraction of multiple crystal hits. Luckily it was not a problem since *pinkIndexer* can index up to five lattices in a single pattern. Fig. 7.13b, for example, shows Bragg peak prediction for two crystals indexed in Fig. 7.13a. From 13632 patterns identified as hits 12393 were indexed: 11125 containing one, 1177 containing two and 91 containing three or more crystals. In total 13757 crystals were found.

The unit cell parameters determined from the multilayer data (Table 6.2) were used as an input for *pinkIndexer*. Following the procedure established in the previous section, the resulting distribution of the X-ray energies contributing to the found Bragg peaks is plotted in Fig. 7.14a. Compared to the distributions obtained above for proteinase K data (Fig. 7.7), it is much broader and does not match the shape of the spectrum. This is an expected effect of the unit cell size varying with the position on the chip as each crystal in this case has its own scaling factor between the real unit cell and the target unit cell used for indexing. Therefore, using the common unit cell scaling factor applied to each indexed crystal in this dataset is not sufficient. However, the individual scaling factors can be obtained for crystals with the large enough numbers of found peaks using the same procedure as for the whole dataset. Examples of the distributions of the peak X-ray energies fitted with the spectrum for individual patterns are shown in Fig. 7.15. After applying the individual scaling factors to the crystals with more than 100 correctly predicted peaks, which constitute approximately 33% of all indexed crystals, and the average scaling factor to the remaining 65% of crystals, the distribution becomes significantly narrower (Fig. 7.14b).

In order to assess the effects of the per-crystal unit cell scaling on the quality of the merged intensities, the data was merged using both common unit cell scaling factor for all crystals and individual scaling factors for crystals with more than 100 peaks, in both cases with the resolution-dependent intensity cut-off, shown in Fig. 7.16a. The resulting CC^* demonstrates slightly higher values at resolutions below 2.2 Å when the individual scaling factors are used (Fig. 7.16b). The effect is, however, very small, probably because only one third of the crystals are scaled individually. For further comparison structure refinements were carried out with both sets of merged intensities using the 5 μs structure from Section 6.4 as a model. The analysis metrics presented in Table 7.3 confirm the slight improvement in the data quality after the individual unit cell scaling with $R_{work}/R_{free} = 0.176/0.204$ compared to 0.176/0.211 achieved with the common scaling factor.

To demonstrate again the benefit of the resolution-dependent energy cut-off, the intensities were merged using the constant $I_{co} = 0.4$. The resulting CC^* and structure refinement statistics are also shown

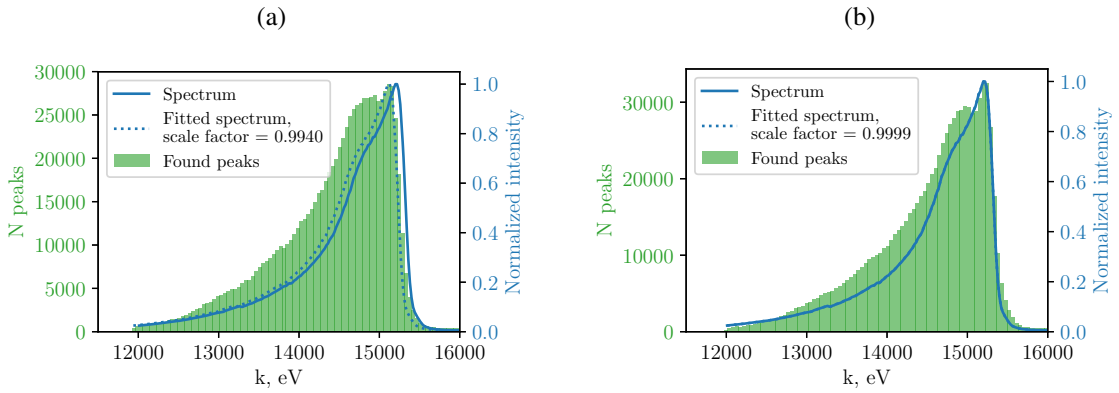


Figure 7.14: Distributions of the X-ray energies contributing to the detected peaks before (a) and after (b) the per-pattern unit cell scaling for crystals with > 100 peaks.

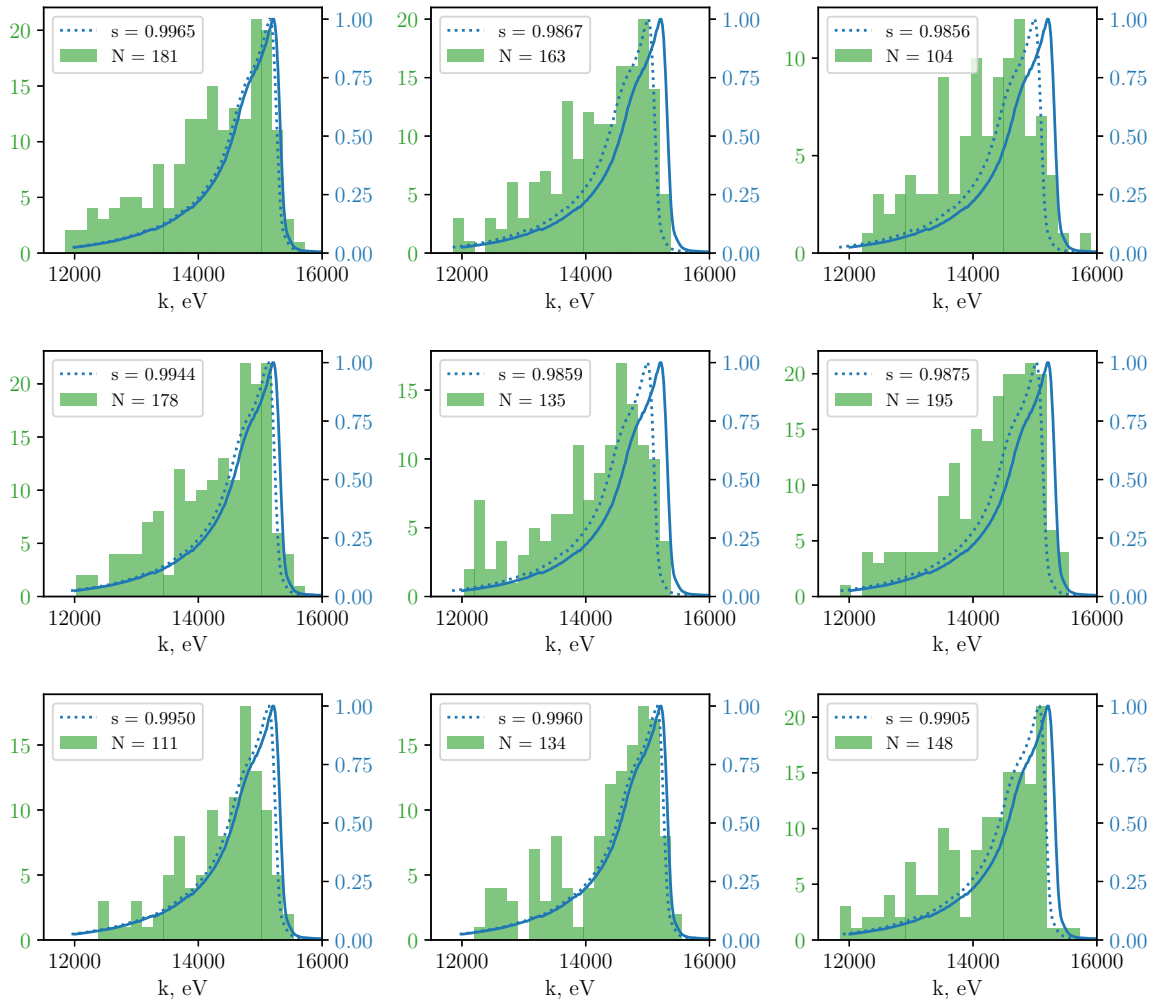


Figure 7.15: Examples of peak energy distributions for individual crystals. The numbers of correctly predicted peaks and the resulting unit cell scaling factors for each crystal are given in the legends.

in Fig. 7.16b and Table 7.3, both proving the advantage of resolution-dependent approach. The effects of the Lorentz factor scaling are evident from the obtained B-factors, which are approximately equal to the ones obtained with 2.5% bandwidth using the multilayer monochromator (Table 6.2).

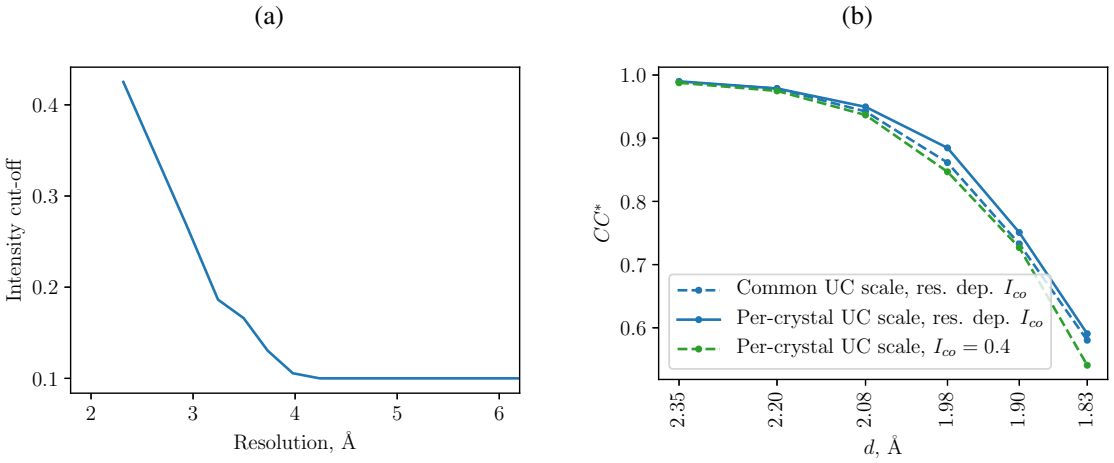


Figure 7.16: (a) Spectral intensity cut-off as a function of resolution. (b) CC^* as a function of resolution for three sets of merged intensities: common unit cell scaling factor + resolution-dependent I_{co}, per-crystal unit cell scaling factor + resolution-dependent I_{co} and per-crystal unit cell scaling factor + constant $I_{co} = 0.4$.

UC scaling Intensity cut-off	Common res. dep.	Per-crystal res. dep.	Per-crystal $I_{co} = 0.4$
Space group	$P4_32_12$		
Unit cell a, b, c, Å α, β, γ,°	79.75 79.75 37.93 90 90 90		
Resolution range, Å	17.83 - 1.85 (1.916 - 1.85)	17.83 - 1.85 (1.916 - 1.85)	8.916 - 1.85 (1.915 - 1.85)
Completeness, %	99.37 (98.79)	99.31 (98.98)	96.42 (98.57)
Mean $I/\sigma(I)$	9.13	9.9	9.85
Wilson B-factor, Å^2	19.72	19.42	20.95
$CC_{1/2}$	0.995	0.997	0.982
Reflections N_{work}/N_{free}	10879 / 738	10877/ 739	10565 / 717
R_{work}	0.176 (0.314)	0.176 (0.307)	0.181 (0.320)
R_{free}	0.211 (0.318)	0.204 (0.291)	0.213 (0.335)
RMS bonds, Å	0.010	0.005	0.004
RMS angles, °	1.34	1.07	1.04
Average B-factor, Å^2	20.49	21.36	22.72

Table 7.3: Data processing and structure refinement parameters for three sets of merged intensities.

7.2.2 Dependence of data quality on number of merged patterns

Similar to analysis performed in Section 6.4 on the data collected with 2.5% bandwidth, thirteen sub-datasets were randomly selected from the whole dataset of 12393 indexed patterns containing from 200 to 7000 indexed crystals. The datasets were merged separately and then subjected to structure refinement under exactly same conditions (same model, same refinement parameters, same R_{free}-flags) using reflections up to 2 Å resolution.

The resulting data completeness, percentage of reflections with $I/\sigma(I) > 2$ and CC^* as a function of resolution are shown in Fig. 7.17. Compared to the same plots for 2.5% bandwidth (Fig. 6.19), several differences can be pointed out. First, the changed behaviour of completeness, i.e. the drop off of completeness at low resolution is explained by the large masked region around the beamstop (Fig. 7.13). As a consequence, the low resolution reflections were only integrated when they were produced by the lower X-ray energies. Furthermore, among the low resolution reflections a larger fraction belongs to densely populated directions in the reciprocal space, e.g. (100) or (110), which often leads to two (or more) reflections from the same direction appearing at the same spot on the detector (Fig. 6.2a). Such reflections are then discarded as overlaps which further reduces low resolution completeness.

The other two metrics, percentage of reflections with $I/\sigma(I) > 2$ and CC^*, begin to fall down at approximately 0.3 Å lower resolution compared to the multilayer data. As explained above, this is a consequence of higher background and shorter exposure times used in the measurements with the full undulator bandwidth. It is also the reason why these structures were refined up to 2 Å resolution instead of 1.7 Å used for the multilayer data.

However, in this case the datasets consisting of a smaller number of indexed crystals show higher metrics compared to the multilayer case. The drop off at low resolution which can be seen for the merges with < 1500 patterns in Fig. 6.19b and 6.19c is not observed here. This effect, explained by the partiality problem being predominant at low resolution (Fig. 6.21), should be shifted to a lower resolution when the broader bandwidth is used. Indeed, the width of the wedge between the limiting Ewald spheres at the reflection resolution d can be calculated as $2d\sin(\Theta_{\max} - \Theta_{\min})/2$, where $2\Theta_{\min}$ and $2\Theta_{\max}$ are the scattering angles corresponding to $\lambda_{\min}$ and $\lambda_{\max}$ respectively: $2/\lambda \sin\Theta = d$. Using the small angle approximation, the width can be estimated as $w = d^2\Delta\lambda/2$. Therefore, considering that the FWHM of the undulator spectrum is 5.1% which is two times bigger than 2.5% produced by the multilayer and the full energy range is more than 7 times larger, the same decrease in the low resolution metrics for the undulator data for the crystals of the same kind should be observed at resolutions at least $\sqrt{2}$ times but, more likely, about $\sqrt{7}$ times lower. This means it would increase from 3 Å in case of multilayer to about 7 Å in case of the full undulator beam, which we can not observe in our case due to overall low number of reflections measured in the low resolution range.

Fig. 7.18a shows the dependence of R_{free} on the number of merged crystals. Although, due to the reasons stated previously, the best R_{free} values obtained for the datasets consisting of a larger number of patterns are higher than the values achieved with 2.5% bandwidth data (Fig. 6.20), they increase slower with reducing the number of patterns: datasets with fewer than 1000 patterns show lower R_{free} with the full bandwidth than with the multilayer. To compare the rate at which the data quality deteriorates, relative change in the R_{free} from the lowest value is plotted in Fig. 7.18b for both cases: with and without multilayer monochromator. As can be seen, without the multilayer R_{free} decreases about 2.5 times slower, which means that, provided the experimental setup is carefully optimized to reduce the

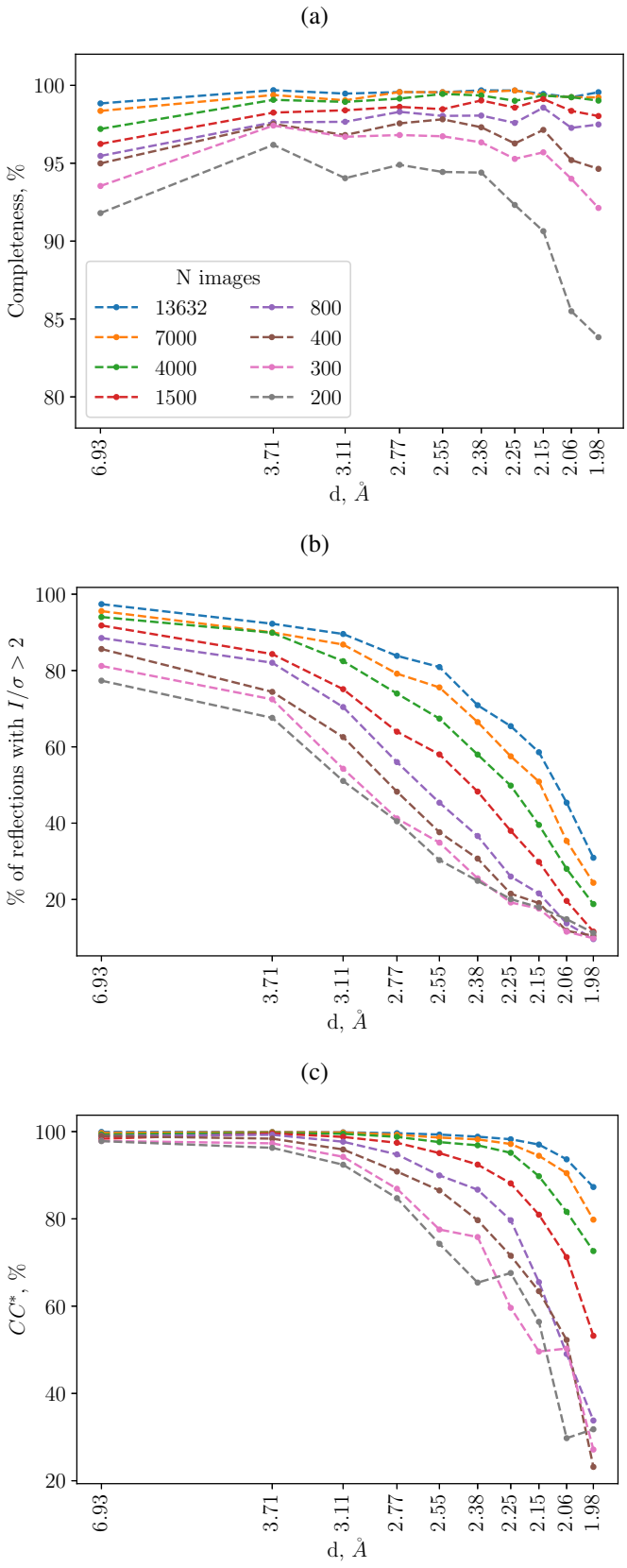


Figure 7.17: Completeness (a), percentage of reflections with $I/\sigma \geq 2$ (b) and CC^* (c) as a function of resolution for different numbers of merged crystals.

background scattering, the same data quality can be achieved with 2.5 times fewer diffraction patterns using the full undulator bandwidth compared to when the multilayer monochromator is used. Since these measurements were performed in the same experimental geometry with the detector covering only half of the scattering area (Fig. 6.11), following from the conclusion made in Section 6.4 that 1500 diffraction patterns collected with 2.5% X-ray bandwidth are enough for a high-quality dataset, we can suggest that using the full undulator beam only 600 diffraction patterns should be sufficient to obtain the structure of a similar high quality.

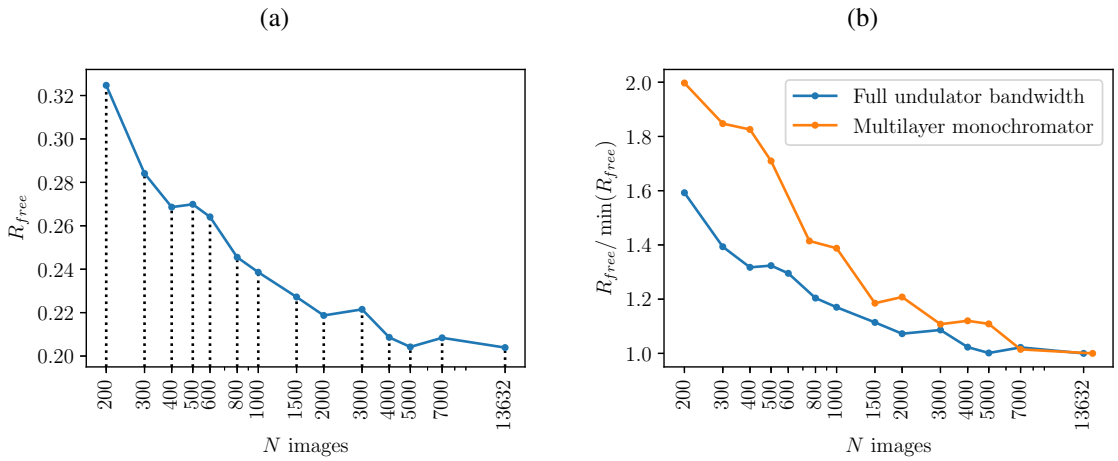


Figure 7.18: (a) R_{free} as a function of the number of merged crystals for the lysozyme data measured with the full undulator bandwidth. (b) Comparison of the relative increase in R_{free} with reducing the number of merged crystals in case of the full undulator bandwidth of 5.1% (FWHM) and multilayer monochromator bandwidth of 2.5% (FWHM).

7.2.3 Conclusion

This section demonstrated that the pink-beam data processing pipeline in *CrystFEL* can indeed handle polychromatic diffraction data collected with the modern integrating detector, which makes it the only currently available software capable of such processing. Furthermore, it has been shown that given the data is sufficiently strong, i.e. there is large enough number of detected diffraction spots per each crystal, the variations of the unit cell parameters due to fluctuating humidity, which is a known drawback of the fixed-target approach, can be at least partially accounted for. A direct comparison between the data collected with the full undulator bandwidth and with 2.5% bandwidth of the multilayer monochromator suggests that in the former case about 2.5 times less data is required to obtain a dataset of a similar quality.

7.3 Liquid jet experiment at BioCARS

Both examples presented in the two previous sections involved analysis of the strong polychromatic diffraction patterns collected in a serial fashion from relatively big crystals using the fixed-target approach. In this section I show the processing of weak and sparse pink-beam diffraction data collected from lysozyme microcrystals below 10 µm in size with 100 ps exposure times. The experiment was performed at the BioCARS beamline at the APS using the double flow-focusing liquid injector for sample delivery [53] and JUNGFRAU detector.

7.3.1 Analysis of sparse pink-beam diffraction data

From 200000 images acquired during 200 seconds of data collection, 8248 images with more than 15 peaks were classified as hits. Similar to $A_{2A}AR$ data described by Martin-Garcia *et al.* 2019 [115], the found diffraction patterns are relatively sparse containing on average 40 peaks per hit. For comparison, proteinase K and lysozyme datasets described previously contain 110 and 305 peaks per hit respectively. As a consequence, this data can not be indexed using the traditional Laue indexing approaches which require much larger number of diffraction peaks per pattern [115]. However, *pinkIndexer* is perfectly capable to index even such sparse data: from 8248 hits 8240 were indexed, 702 of them as multiple crystals. Fig. 7.19a shows the distributions of all found hits as well as the hits indexed as single and multiple crystals as functions of the number of detected peaks. It can be seen that *pinkIndexer* yields almost 100% indexing fraction even for extremely weak diffraction patterns with only 15 found peaks. Furthermore, it can successfully find multiple lattices in the sparse patterns. Example of such double crystal diffraction pattern containing in total 63 detected peaks (blue squares) is demonstrated in Fig. 7.20: each of the two found crystal lattices (yellow and green circles) predicts 31 of the detected peaks. The resulting distribution of the numbers of correctly predicted peaks for each crystal is shown in Fig. 7.19b. It gives an average of 37 predicted peaks per crystal.

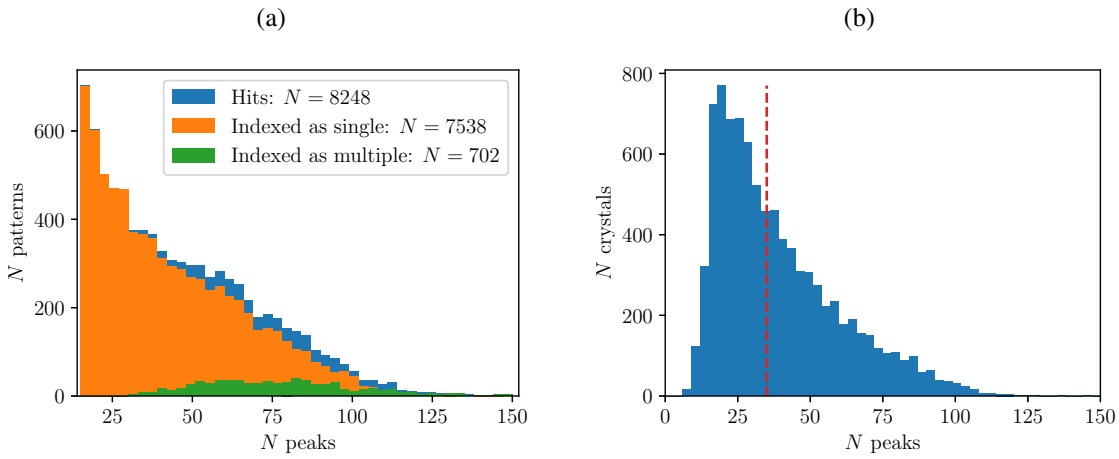


Figure 7.19: (a) Distribution of all found hits (blue), hits indexed as single crystal (orange) and hits indexed as two or more crystals (green) as function of number of found peaks. (b) Distribution of indexed crystals as a function of number of corresponding Bragg peaks.

Following the polychromatic data processing pipeline (Fig. 7.12), the next steps after indexing are unit cell scaling, resolution-dependent spectral intensity cut-off and scaling of integrated reflection intensities.

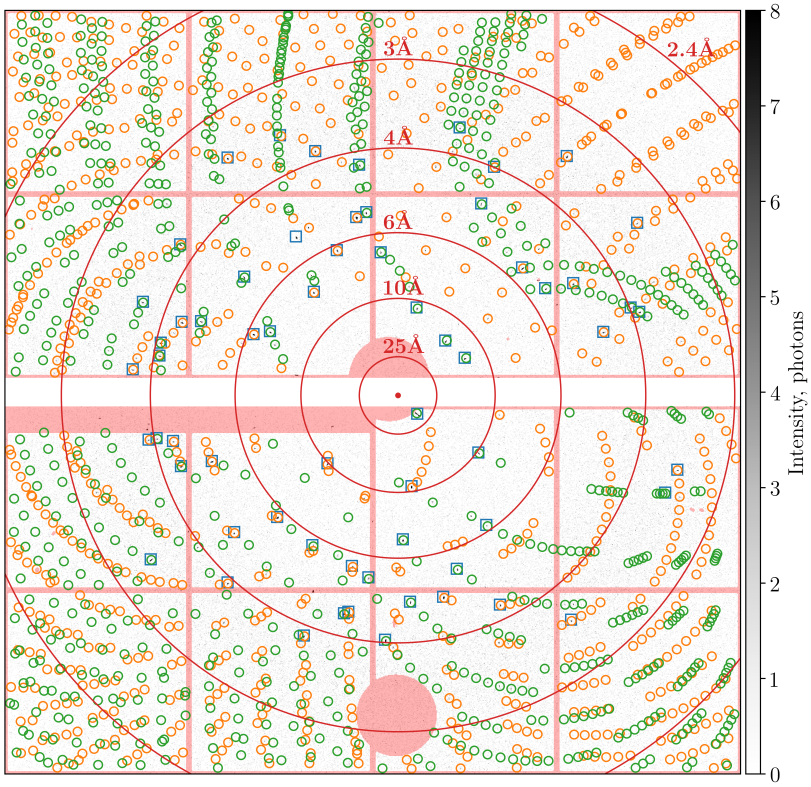


Figure 7.20: Example double crystal lysozyme diffraction pattern. Blue squares show positions of the found peaks, yellow and green circles show predicted Bragg spot positions for two found crystal lattices. Red circles show resolution rings corresponding to the X-ray energy of 12 keV. Masked regions of the detector are shown in pink.

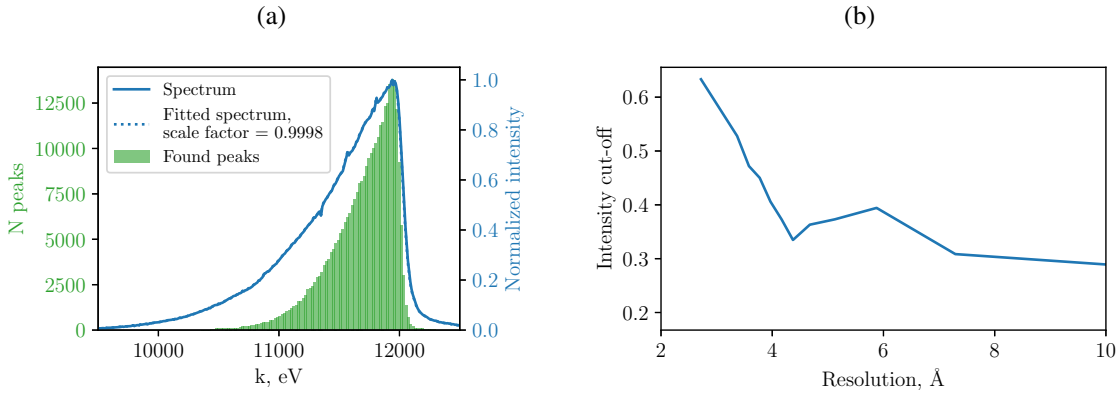


Figure 7.21: (a) Distributions of the X-ray energies contributing to the detected peaks. (b) Spectral intensity cut-off as a function of resolution.

Thus, the distribution of the X-ray energies producing all found peaks is plotted in Fig. 7.21a. As expected for the weak data it is significantly narrower than the measured X-ray spectrum. However, the detected peaks still originate from a broad range of energies of more than 10% bandwidth, confirming the conclusion made from the analysis of simulated patterns in Section 6.3 that even sparse pink-beam diffraction data can not be considered monochromatic. As the peak energies distribution is narrower compared to the two previous examples, the resulting spectral intensity cut-off determined from the width

of the distribution in different resolution bins is in this case higher than in previous cases (Fig. 7.21b). The reflection intensities were then merged with *partialator* and the crystal structure was refined using PHENIX. The resulting merging and refinement statistics are given in the first column in Table 7.4.

7.3.2 Dependence of data quality on sparsity of diffraction data

Although the average number of detected peaks per crystal is only 37, there are stronger diffraction patterns in the dataset with up to 120 predicted peaks per crystal (Fig. 7.19b). In order to examine if it is possible to improve the quality of the merged intensities by discarding weakly diffracting crystals, 11 sub-datasets were created by selecting crystals with the minimum number of detected peaks varying from 35 to 85. Additionally, crystals with fewer than 35 predicted peaks (red dashed line in Fig. 7.19b) were sorted out and merged separately to investigate the quality achievable with only weak sparse patterns with 20 predicted peaks per crystal on average.

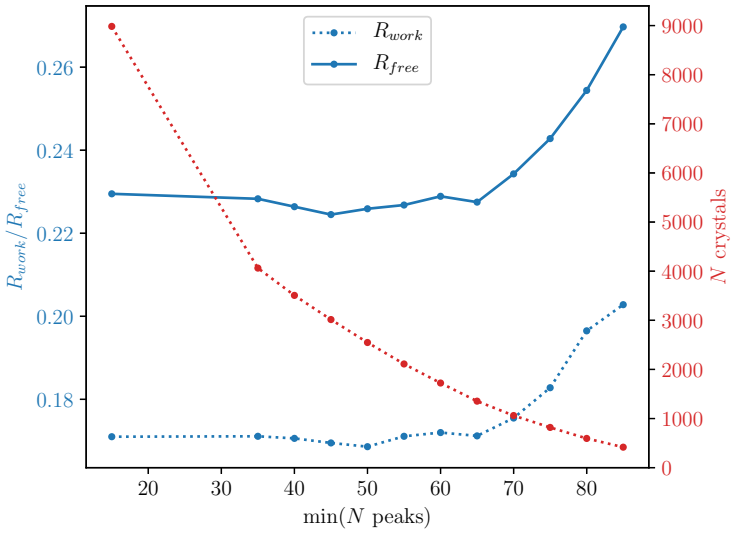


Figure 7.22: R_{work}, R_{free} and the number of crystals as functions of the minimum number of peaks per crystal.

The dependence of R_{work} and R_{free} on the minimum number of peaks as well as the number of crystals in each sub-dataset are shown in Fig. 7.22. As it can be seen, despite decreasing number of merged crystals from almost 9000 to about 1500 when merging only crystals with more than 65 peaks, both R-factors do not significantly change. When the minimum number of peaks is further increased and the number of merged crystals falls down below 1500, R-factors only become worse. It shows that only very little improvement can be gained by discarding the weak diffraction data from merging. The best achieved $R_{work}/R_{free} = 0.170/0.224$ obtained merging 3015 crystals with N peaks > 45 are only slightly lower than 0.171/0.230 obtained from the whole dataset. The CC^*, shown for both datasets in Fig. 7.23, is also similar.

The results obtained for the dataset consisting of 3345 weakest crystals, with N peaks < 35, are significantly worse. The refinement statistics shown in Table 7.4 is in fact comparable to the structure refined from 595 strongest diffracting crystals with N peaks > 80, although CC^* of the dataset consisting of crystals with N peaks < 35 falls off much faster at high resolution (Fig. 7.23).

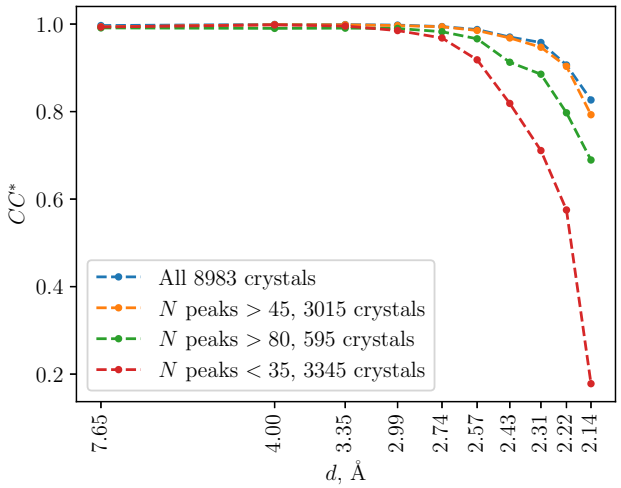


Figure 7.23: CC^* as a function of resolution for the four compared datasets.

Dataset	All crystals	N peaks > 45	N peaks > 80	N peaks < 35
Space group	$P4_32_12$			
Unit cell a, b, c, Å α, β, γ,°	79.58 79.58 37.91 90 90 90			
N crystals	8983	3015	595	3345
Resolution range, Å	22.07 - 2.1 (2.175 - 2.1)		22.07 - 2.2 (2.279 - 2.2)	
Completeness (%)	99.73 (99.86)	99.52 (99.31)	99.13 (97.15)	99.53 (99.21)
Mean $I/\sigma(I)$	10.6	8.5	5.5	5.8
Wilson B-factor, Å^2	17.38	18.06	18.11	21.37
$CC_{1/2}$	98.9	98.0	96.5	97.9
Reflections N_{work}/N_{free}	7497 / 749	7480 / 747	6495 / 647	6537 / 655
R_{work}	0.171 (0.249)	0.170 (0.255)	0.197 (0.278)	0.192 (0.312)
R_{free}	0.230 (0.341)	0.224 (0.316)	0.254 (0.345)	0.257 (0.319)
RMS bonds, Å	0.003	0.003	0.004	0.004
RMS angles, °	0.92	0.92	0.93	0.93
Average B-factor, Å^2	18.98	19.47	18.78	20.95

Table 7.4: Data analysis and refinement parameters of the four compared datasets.

7.3.3 Conclusion

The first major advantage of pink-beam serial crystallography is that it substantially reduces sample consumption compared to serial crystallography with the monochromatic beam while allowing to collect low dose data by exposing each crystal only once as opposed to conventional data collection. The second advantage is the possibility to use much shorter exposure times which would not only reduce the data collection time but also allow to perform time-resolved measurements of essential biological reactions at time scales starting from hundreds of picoseconds. Both advantages would be mostly beneficial when very small crystals are used. Firstly, serial approach is often required when only small crystals can be grown and the total amount of material is limited. Secondly, for the time-resolved measurements, such as

laser pump-probe or mix-and-inject, the crystals should be sufficiently small for the reaction to take place within a larger fraction of the sample volume. For example, laser penetration depth in the crystals of the photoactive yellow protein is only 30 μm [30], while the typical size of the crystals used in traditional time-resolved Laue experiments is above 100 μm (Section 3.4). Therefore, the capability of the software to process weak diffraction data from small crystals is a major requirement for the success of the method.

Here I showed that the pink-beam data processing pipeline in *CrystFEL* is suitable for the weak diffraction data. Discarding the weakest diffraction patterns in this case resulted in only slight improvement of the data quality. Although the structure obtained from approximately 3000 sparse diffraction patterns with the average of 20 detected Bragg peaks per crystal is, unsurprisingly, of a significantly worse quality compared to the whole dataset, it is nevertheless comparable to the structure obtained from 600 strongest patterns. This result suggests that, given enough data, a high quality set of intensities can be obtained with *CrystFEL* even from very weak polychromatic diffraction patterns. Additionally, it further highlights the advantage of the automatic processing in *CrystFEL* as it becomes essential with the larger number of diffraction patterns required in this case.

7.4 Discussion

Compared to other available pink-beam crystallography software, the processing pipeline presented here offers several advantages which are especially important for serial crystallography. First, it is integrated in *CrystFEL* software suite, which is specifically designed for serial crystallography and allows to process large amounts of data automatically with only minimal manual intervention. It can handle diffraction data collected with modern detectors assembled from multiple panels, and provides the possibility to accurately refine geometry of the said multi-panel detectors. Secondly, it does not require prior knowledge of the unit cell parameters as their estimations can be obtained using the monochromatic indexing algorithms integrated in *CrystFEL*. Lastly, it is currently the only software which can process sparse pink-beam data from small, weakly diffracting crystals often used in serial crystallography.

Possible further improvements of the pipeline, which should reduce the number of required diffraction patterns and improve quality of the merged data, include the following:

1. Automatic determination of the reflection profile radius. As mentioned in Section 6.4, the profile radius determination in *indexamajig*, based on calculation of the distance between the predicted reciprocal lattice points and an infinitely thin Ewald sphere in case of monochromatic X-rays, can not be applied to polychromatic data. Hence, currently reflection profile radius has to be set manually. For the X-ray beam with relatively narrow bandwidth, such as 2.5% bandwidth produced by multilayer monochromator, it is possible to determine profile radius using the same approach as for monochromatic data by taking into account only low-resolution reflections where the wedge between the limiting Ewald spheres is still thin. For the broader bandwidth, such as full undulator harmonic, a more general approach has to be implemented. Rather than calculating the distances between the reciprocal lattice points and the Ewald sphere, it can possibly be determined by minimizing the distances between the detected spots on the detector and corresponding predicted reflection positions.

2. Using integrated intensities to refine X-ray spectrum and resolve overlaps. While the distribution of the X-ray energies contributing to the detected peaks can usually give the general shape of the incident spectrum and allows to determine the absolute values of the unit cell parameters, it strongly depends on the strength of the diffraction data (Fig. 7.21a) and the variation of the unit cell parameters (Fig. 7.21a, 7.15). Therefore, we have to rely on the measured spectrum to scale the integrated reflection intensities accordingly. By comparing integrated intensities of equivalent reflections sampled by different X-ray energies in different crystal orientations, it should be possible to obtain a normalization curve for intensity scaling, which would include all energy dependent factors, e.g. varying detector response, in addition to the incident spectrum. Furthermore, using the normalization curve and integrated intensities of overlapping reflections measured in different crystal orientations, it should be possible to deconvolute the overlaps by finding a least-squares solution of a system of linear equations, similar to the procedure described in [127].

3. Refinement of experimental parameters and reflection profile fitting. Currently, prediction refinement procedure in *pinkIndexer* optimizes lattice parameters, orientation of the crystal and the detector center. In the case of monochromatic X-rays, the only other parameter which is refined in *CrystFEL* for each pattern individually is reflection profile radius. Degree of crystal mosaicity, another parameter which is important for accurate peak prediction, is not currently used in the diffraction model implemented in *CrystFEL*. The choice of using only profile radius and omitting crystal mosaicity is caused by the small number of the detected peaks per crystal, which is usually insufficient to determine both profile size and mosaicity. Despite this lack of modeling of mosaicity, *CrystFEL* is very successful in accurate prediction and integration of diffraction spots in the monochromatic diffraction patterns. Compared to still diffraction images measured with monochromatic beam, where higher crystal mosaicity is rarely noticeable, effects of high crystal mosaicity on polychromatic diffraction patterns are more severe. They result in largely elongated peaks in radial direction, as can be seen in the example diffraction patterns in Fig. 6.2c and Fig. 6.16. In order to accurately predict and integrate reflections in such patterns, three-rings integration method used in *CrystFEL* is clearly not applicable. Given that polychromatic diffraction patterns often contain large number of detected peaks, especially compared to the monochromatic case, it should be possible to determine both reflection profile radius and crystal mosaicity for each individual crystal. These can then be used for accurate prediction of reflection positions and peak profile shapes. Integration of reflection intensities based on analytical profile fitting can then be implemented instead of three-rings approach, which should allow to achieve quality of the merged intensities similar to the traditional Laue processing software without using larger amounts of data.

CHAPTER 8

Summary and outlook

Data analysis is an integral part of any research. Almost every emerging experimental technique leads to the development of the specific data analysis methods. Serial crystallography is a relatively new method and although it already has established analysis routine there are still aspects open for improvement and new developments. In this thesis I presented several serial crystallography experiments conducted at three different facilities: LCLS, ESRF and APS, showing how careful and thorough approach to the data analysis helps to extract maximum information from the available data and advance the field of serial crystallography. In particular, evaluation of the humidity variations and their influence on the quality and consistency of the data in the fixed-target experiments presented in Sections 5.2 and 6.4 provided valuable feedback and led to a significant improvement of the experimental setup. Humidity chamber of the improved design provides a more stable and homogeneous humidity environment while enabling direct access to the crystals on the fixed-target chip for external manipulation. As a result, with the advantages Roadrunner fixed-target setup [65] has over the liquid jets, i.e. lower background, lower sample consumption and higher hit rate, it becomes an attractive sample delivery option for time-resolved serial crystallography experiments.

The major topic of this thesis was establishing the method of serial synchrotron crystallography with a polychromatic (or pink) X-ray beam. More specifically, the goal was to extend existing data analysis routines in *CrystFEL* [74] used for serial crystallography with monochromatic X-rays to the polychromatic case. Serial crystallography with the pink beam offers two main advantages. First, due to the higher number of Bragg peaks and a large fraction of fully integrated reflections in a still diffraction pattern it requires fewer crystals to obtain a complete high-quality set of reflection intensities. Second, since the achievable flux is two orders of magnitude higher compared to the monochromatic beam, it allows to collect data with much lower exposure times.

In general, there are two types of polychromatic X-ray spectra available at the undulator beamlines at the third generation synchrotrons. Spectrum of an undulator harmonic has a typical bandwidth of about 5% and a long low energy tail of about 15-20%. This low energy tail can be cut off using a multilayer monochromator which narrows the spectrum to about 2-3% and makes it more symmetric. As shown in Section 6.3, the standard monochromatic indexing algorithms can not accurately index diffraction patterns produced by the full undulator bandwidth. They are, however, sufficiently accurate in indexing diffraction data produced by the multilayer beam. The proof-of-principle serial crystallography experiment using the multilayer monochromator at beamline ID09 at the ESRF was presented in Section 6.4. There, diffraction patterns were indexed using available monochromatic indexing algorithms in *CrystFEL* and

only peak prediction and integration algorithms had to be adapted to the polychromatic beam. Using this extended data processing pipeline we demonstrated that only 1500 diffraction patterns recorded with 2.5% bandwidth of the multilayer and 1 μs exposure were sufficient to reach the optimal quality of the data. Further increasing the number of patterns only slightly improved structure refinement statistics but did not significantly affect the quality of the resulting electron density maps.

In Chapter 7 I presented the full data analysis pipeline for pink-beam serial crystallography, using the new indexing algorithm for polychromatic diffraction data recently developed by Gevorkov *et al.* [5] and further modification to the standard monochromatic pipeline in *CrystFEL*. With this new pipeline it was possible to process pink-beam serial diffraction data recorded with the modern charge integrating detector from both strongly and very poorly diffracting crystals, which makes it the only currently available software for such processing. Furthermore, it was shown that using the full undulator bandwidth instead of the multilayer monochromator the required number of diffraction patterns reduced by the factor of 2.5. With the future developments of the integration and merging procedures proposed in Section 7.4, the total required number of diffraction patterns should be further decreased.

With the significantly lower sample consumption, fast data collection with modern detectors and very short exposure times, pink-beam serial crystallography opens many new opportunities. Using 2.5% bandwidth of the multilayer monochromator with Roadrunner fixed-target setup and JUNGFRAU detector it was possible to obtain the high-quality structure from less than 30 seconds of data collection. With even shorter data collection times using the full undulator bandwidth it makes the method perfectly suitable for fast ligand screening for pharmaceutical research.

Extending the method to time-resolved serial crystallography is of course the next logical step: short exposures of around 100 ps achievable with single bunches offer the possibility to perform such experiments with the time resolution at the sub-nanosecond range, which should cover a vast range of biological processes. Performing time-resolved experiments in a serial fashion presents a huge advantage over the standard time-resolved Laue technique described in Section 3.4. First, it allows to extend the technique to study non-cyclic and irreversible reactions. It can be applied to more radiation-sensitive proteins as the total dose inflicted on each crystal is significantly lower. Finally, thanks to the capabilities of the new software, it can use much smaller crystals, below 10 μm in size, which is a massive improvement over a typical crystal size of 100-2000 μm used in the standard time-resolved Laue crystallography experiments [27–30]. Smaller crystal size should allow more uniform and efficient laser activation in pump-probe experiments. Additionally, it should make the method suitable for mix-and-inject experiments, where the small crystal size is essential to reduce the diffusion times [47]. Such experiments, already performed at the FELs, open the opportunity for time-resolved studies of a very large class of enzymatic reactions [8, 26, 49]. With the advantage of requiring fewer diffraction patterns compared to monochromatic experiments at FELs and overcoming limitations of the traditional Laue crystallography at synchrotrons, time-resolved pink-beam crystallography may soon become an alternative technique to both FEL and standard Laue experiments.

With more polychromatic diffraction beamlines for macromolecular crystallography becoming available at many synchrotron radiation facilities such as ESRF, APS, Max IV, and PETRA III in the near future, the method of pink-beam serial crystallography will undoubtedly become more popular. Furthermore, some FEL endstations, for example ARAMIS beamline at SwissFEL, also offer polychromatic beams for serial crystallography experiments [128]. Hence, data analysis methods for pink-beam diffraction

developed as part of this dissertation can also become applicable in the field of serial femtosecond crystallography.

Bibliography

[1] Berman, H., Henrick, K., and Nakamura, H. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980, 2003.

[2] Bragg, W. H. and Bragg, W. L. The reflection of x-rays by crystals. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 88(605):428–438, 1913.

[3] Chapman, H. N., Fromme, P., Barty, A., et al. Femtosecond X-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.

[4] Kendrew, J. C., Bodo, G., Dintzis, H. M., et al. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 1958.

[5] Perutz, M. F., Rossmann, M. G., Cullis, A. F., et al. Structure of Hæmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by X-ray analysis, 1960.

[6] Robertson, J. M. Vector maps and heavy atoms in crystal analysis and the insulin structure. *Nature*, 143(3611):75, 1939.

[7] Rossmann, M. G. *The Molecular Replacement Method*. New York: Gordon & Breach, 1972.

[8] Kruse, A. C., Manglik, A., Kobilka, B. K., and Weis, W. I. Applications of molecular replacement to g protein-coupled receptors. *Acta Crystallographica Section D: Biological Crystallography*, 69(11):2287–2292, 2013.

[9] Scapin, G. Molecular replacement then and now. *Acta Crystallographica Section D Biological Crystallography*, 2013.

[10] Zeldin, O. B., Gerstel, M., and Garman, E. F. Raddose-3d: time-and space-resolved modelling of dose in macromolecular crystallography. *Journal of applied crystallography*, 46(4):1225–1230, 2013.

[11] Blake, C. and Phillips, D. Effects of x-irradiation on single crystals of myoglobin. 1962.

[12] Garman, E. F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D: Biological Crystallography*, 66(4):339–351, 2010.

[13] Weik, M., Ravelli, R. B., Kryger, G., et al. Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences*, 97(2):623–628, 2000.

[14] Henderson, R. 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.

[15] Owen, R. L., Rudiño-Piñera, E., and Garman, E. F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proceedings of the National Academy of Sciences*, 103(13):4912–4917, 2006.

[16] Howells, M. R., Beetz, T., Chapman, H. N., 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.

[17] Roedig, P., Duman, R., Sanchez-Weatherby, J., et al. Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering. *Journal of Applied Crystallography*, 49(3):968–975, 2016.

[18] Authier, A. *Dynamical Theory of X-ray Diffraction*. IUCr monographs on crystallography. Oxford University Press, 2004.

[19] Berman, L. E., Yin, Z., Dierker, S. B., et al. Performance of the double multilayer monochromator on the nsls wiggler beam line x25. *AIP Conference Proceedings*, 417(1):71–79, 1997.

[20] Rossmann, M. G. Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. *Journal of Applied Crystallography*, 12(2):225–238, 1979.

[21] Winkler, F., Schutt, C. t., and Harrison, S. The oscillation method for crystals with very large unit cells. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 35(6):901–911, 1979.

[22] Cianci, M., Bourenkov, G., Pompidor, G., et al. P13, the EMBL macromolecular crystallography beamline at the low-emittance PETRA III ring for high- and low-energy phasing with variable beam focusing. In *Journal of Synchrotron Radiation*, volume 24, pages 323–332, 2017.

[23] Margiolaki, I. and Wright, J. P. Powder crystallography on macromolecules. *Acta Crystallographica Section A*, 64(1):169–180, Jan 2008.

[24] Bourgeois, D. and Weik, M. Kinetic protein crystallography: a tool to watch proteins in action. *Crystallography reviews*, 15(2):87–118, 2009.

[25] Bourgeois, D., Schotte, F., Brunori, M., and Vallone, B. 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.

[26] Šrajer, V. and Schmidt, M. Watching proteins function with time-resolved x-ray crystallography. *Journal of physics D: Applied physics*, 50(37):373001, 2017.

[27] Šrajer, V., Teng, T. Y., Ursby, T., et al. Photolysis of the Carbon Monoxide Complex of Myoglobin: Nanosecond Time-Resolved Crystallography. *Science*, 274(5293):1726–1729, 1996.

[28] Schotte, F., Lim, M., Jackson, T. A., et al. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300(5627):1944–1947, 2003.

[29] Schmidt, M., Henning, R., Ihee, H., et al. Protein energy landscapes determined by five-dimensional crystallography. *Acta Crystallographica Section D Biological Crystallography*, 69(12):2534–2542, 2013.

[30] Schotte, F., Cho, H. S., Kaila, V. R., et al. Watching a signaling protein function in real time via 100-ps time-resolved laue crystallography. *Proceedings of the National Academy of Sciences*, 109(47):19256–19261, 2012.

[31] Ihee, H., Rajagopal, S., Srajer, V., et al. Visualizing reaction pathways in photoactive yellow protein from nanoseconds to seconds. *Proceedings of the National Academy of Sciences*, 102(20):7145–7150, 2005.

[32] Schmidt, M., Pahl, R., Srajer, V., et al. Protein kinetics: Structures of intermediates and reaction mechanism from time-resolved x-ray data. *Proceedings of the National Academy of Sciences*, 101(14):4799–4804, 2004.

[33] Bourgeois, D., Vallone, B., Schotte, F., et al. Complex landscape of protein structural dynamics unveiled by nanosecond laue crystallography. *Proceedings of the National Academy of Sciences*, 100(15):8704–8709, 2003.

[34] Bourgeois, D., Vallone, B., Arcovito, A., et al. Extended subnanosecond structural dynamics of myoglobin revealed by laue crystallography. *Proceedings of the National Academy of Sciences*, 103(13):4924–4929, 2006.

[35] Schmidt, M., Nienhaus, K., Pahl, R., et al. Ligand migration pathway and protein dynamics in myoglobin: A time-resolved crystallographic study on L29W MbCO. *Proceedings of the National Academy of Sciences*, 102(33):11704–11709, 2005.

[36] Šrajer, V., Ren, Z., Teng, T.-Y., et al. Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved laue x-ray diffraction. *Biochemistry*, 40(46):13802–13815, 2001.

[37] Knapp, J. E., Pahl, R., Srajer, V., and Royer, W. E. Allosteric action in real time: Time-resolved crystallographic studies of a cooperative dimeric hemoglobin. *Proceedings of the National Academy of Sciences*, 103(20):7649–7654, 2006.

[38] Neutze, R., Wouts, R., van der Spoel, D., Weckert, E., and Hajdu, J. Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406(6797):752, 2000.

[39] Chapman, H. N., Barty, A., Bogan, M. J., et al. Femtosecond diffractive imaging with a soft-x-ray free-electron laser. *Nature Physics*, 2(12):839, 2006.

[40] Stellato, F., Oberthür, D., Liang, M., et al. Room-temperature macromolecular serial crystallography using synchrotron radiation. *IUCrJ*, 1(4):204–212, 2014.

[41] Broennimann, C., Eikenberry, E. F., Henrich, B., et al. The PILATUS 1M detector. *Journal of Synchrotron Radiation*, 13(2):120–130, 2006.

[42] Hart, P., Boutet, S., Carini, G., et al. The CSPAD megapixel x-ray camera at LCLS. In *X-Ray Free-Electron Lasers: Beam Diagnostics, Beamline Instrumentation, and Applications*, volume 8504, page 85040C. International Society for Optics and Photonics, 2012.

[43] Allahgholi, A., Becker, J., Bianco, L., et al. Front end ASIC for AGIPD, a high dynamic range fast detector for the European XFEL. *Journal of Instrumentation*, 11(1):C01057, 2016.

[44] Mozzanica, A., Bergamaschi, A., Brueckner, M., et al. Characterization results of the JUNGFRAU full scale readout ASIC. *Journal of Instrumentation*, 11(2):C02047, 2016.

[45] Barends, T. R., Foucar, L., Ardevol, A., et al. Direct observation of ultrafast collective motions in co myoglobin upon ligand dissociation. *Science*, 350(6259):445–450, 2015.

[46] Pande, K., Hutchison, C. D., Groenhof, G., et al. Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science*, 352(6286):725–729, 2016.

[47] Schmidt, M. Mix and Inject: Reaction Initiation by Diffusion for Time-Resolved Macromolecular Crystallography. *Advances in Condensed Matter Physics*, 2013:1–10, 2013.

[48] Kupitz, C., Olmos Jr, J. L., Holl, M., et al. Structural enzymology using x-ray free electron lasers. *Structural Dynamics*, 4(4):044003, 2017.

[49] Stagno, J., Liu, Y., Bhandari, Y., et al. Structures of riboswitch rna reaction states by mix-and-inject xfel serial crystallography. *Nature*, 541(7636):242, 2017.

[50] Grünbein, M. L. and Nass Kovacs, G. Sample delivery for serial crystallography at free-electron lasers and synchrotrons. *Acta Crystallographica Section D Structural Biology*, 75(2):178–191, feb 2019.

[51] Martiel, I., Müller-Werkmeister, H. M., and Cohen, A. E. Strategies for sample delivery for femtosecond crystallography. *Acta Crystallographica Section D Structural Biology*, 75(2):160–177, feb 2019.

[52] DePonte, D., Weierstall, U., Schmidt, K., et al. Gas dynamic virtual nozzle for generation of microscopic droplet streams. *Journal of Physics D: Applied Physics*, 41(19):195505, 2008.

[53] Oberthuer, D., Knoška, J., Wiedorn, M. O., et al. Double-flow focused liquid injector for efficient serial femtosecond crystallography. *Scientific Reports*, 7(1), 2017.

[54] Wiedorn, M. O., Oberthür, D., Bean, R., et al. Megahertz serial crystallography. *Nature Communications*, 9(1):4025, 2018.

[55] Grünbein, M. L., Bielecki, J., Gorel, A., et al. Megahertz data collection from protein microcrystals at an X-ray free-electron laser. *Nature Communications*, 9(1):3487, 2018.

[56] Landau, E. M. and Rosenbusch, J. P. Lipidic cubic phases: A novel concept for the crystallization of membrane proteins. *Proceedings of the National Academy of Sciences*, 93(25):14532–14535, 1996.

[57] Weierstall, U., James, D., Wang, C., et al. Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nature communications*, 5:3309, 2014.

[58] Zhang, H., Unal, H., Gati, C., et al. Structure of the angiotensin receptor revealed by serial femtosecond crystallography. *Cell*, 161(4):833–844, 2015.

[59] Kang, Y., Zhou, X. E., Gao, X., et al. Crystal structure of rhodopsin bound to arrestin by femtosecond x-ray laser. *Nature*, 523(7562):561, 2015.

[60] Hunter, M. S., Segelke, B., Messerschmidt, M., et al. Fixed-target protein serial microcrystallography with an X-ray free electron laser. *Scientific Reports*, 4:6026, 2014.

[61] Mueller, C., Marx, A., Epp, S. W., et al. Fixed target matrix for femtosecond time-resolved and in situ serial micro-crystallography. *Structural Dynamics*, 2(5):054302, 2015.

[62] Roedig, P., Vartiainen, I., Duman, R., et al. A micro-patterned silicon chip as sample holder for macromolecular crystallography experiments with minimal background scattering. *Scientific Reports*, 5:10451, 2015.

[63] Oghbaey, S., Sarracini, A., Ginn, H. M., 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.

[64] Owen, R. L., Axford, D., Sherrell, D. A., et al. Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallographica Section D: Structural Biology*, 73(4):373–378, 2017.

[65] Roedig, P., Ginn, H. M., Pakendorf, T., et al. High-speed fixed-target serial virus crystallography. *Nature Methods*, 14(8):805–810, 2017.

[66] Meents, A., Wiedorn, M. O., Srajer, V., et al. Pink-beam serial crystallography. *Nature Communications*, 8(1):1281, 2017.

[67] Kirian, R., Wang, X., Weierstall, U., et al. Femtosecond protein nanocrystallography—data analysis methods. *Optics Express*, 18(6):5713, 2010.

[68] Kirian, R. A., White, T. A., Holton, J. M., et al. Structure-factor analysis of femtosecond microdiffraction patterns from protein nanocrystals. *Acta Crystallographica Section A: Foundations of Crystallography*, 67(2):131–140, 2011.

[69] White, T. A., Barty, A., Stellato, F., et al. Crystallographic data processing for free-electron laser sources. *Acta Crystallographica Section D Biological Crystallography*, 69(7):1231–1240, 2013.

[70] White, T. A. Post-refinement method for snapshot serial crystallography. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647), 2014.

[71] Ginn, H. M., Brewster, A. S., Hattne, J., et al. 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.

[72] Kroon-Batenburg, L. M., Schreurs, A. M., Ravelli, R. B., and Gros, P. Accounting for partiality in serial crystallography using ray-tracing principles. *Acta Crystallographica Section D: Biological Crystallography*, 71(9):1799–1811, 2015.

[73] Barty, A., Kirian, R. A., Maia, F. R., et al. Cheetah: software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data. *Journal of applied crystallography*, 47(3):1118–1131, 2014.

[74] White, T. A., Kirian, R. A., Martin, A. V., et al. CrystFEL: A software suite for snapshot serial crystallography. *Journal of Applied Crystallography*, 45(2):335–341, 2012.

[75] White, T. A., Mariani, V., Brehm, W., et al. Recent developments in CrystFEL. *Journal of Applied Crystallography*, 49(2):680–689, 2016.

[76] Sauter, N. K. Xfel diffraction: developing processing methods to optimize data quality. *Journal of synchrotron radiation*, 22(2):239–248, 2015.

[77] Kabsch, W. Processing of x-ray snapshots from crystals in random orientations. *Acta Crystallographica Section D: Biological Crystallography*, 70(8):2204–2216, 2014.

[78] Waterman, D. G., Winter, G., Parkhurst, J. M., et al. The dials framework for integration software. *CCP4 Newslett. Protein Crystallogr*, 49:13–15, 2013.

[79] Powell, H. R. The rossmann fourier autoindexing algorithm in mosflm. *Acta Crystallographica Section D: Biological Crystallography*, 55(10):1690–1695, 1999.

[80] Duisenberg, A. J. Indexing in single-crystal diffractometry with an obstinate list of reflections. *Journal of applied crystallography*, 25(2):92–96, 1992.

[81] Kabsch, W. Xds. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):125–132, 2010.

[82] Ginn, H. M., Roedig, P., Kuo, A., et al. Taketwo: an indexing algorithm suited to still images with known crystal parameters. *Acta Crystallographica Section D: Structural Biology*, 72(8):956–965, 2016.

[83] Beyerlein, K. R., White, T. A., Yefanov, O., et al. 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.

[84] Gevorkov, Y., Yefanov, O., Barty, A., et al. Xgandalf–extended gradient descent algorithm for lattice finding. *Acta Crystallographica Section A: Foundations and Advances*, 75(5):694–704, 2019.

[85] Gevorkov, Y., Barty, A., Brehm, W., et al. pinkindexer–a universal indexer for pink-beam x-ray and electron diffraction snapshots. *Acta Crystallographica Section A: Foundations and Advances*, 76(2), 2020.

[86] Yefanov, O., Mariani, V., Barty, A., et al. Accurate determination of segmented X-ray detector geometry. *Optics Express*, 23(22):28459, 2015.

[87] Brehm, W. and Diederichs, K. Breaking the indexing ambiguity in serial crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 70(1):101–109, 2014.

[88] White, T. A. Processing serial crystallography data with *CrystFEL*: a step-by-step guide. *Acta Crystallographica Section D*, 75(2):219–233, Feb 2019.

[89] Karplus, P. A. and Diederichs, K. Linking crystallographic model and data quality. *Science*, 336(6084):1030–1033, 2012.

[90] Martin-Garcia, J. M., Conrad, C. E., Nelson, G., et al. Serial millisecond crystallography of membrane and soluble protein microcrystals using synchrotron radiation. *IUCrJ*, 4(4):439–454, 2017.

[91] Zhang, H., Qiao, A., Yang, D., et al. Structure of the full-length glucagon class b g-protein-coupled receptor. *Nature*, 546(7657):259, 2017.

[92] Zhang, H., Han, G. W., Batyuk, A., et al. Structural basis for selectivity and diversity in angiotensin II receptors. *Nature*, 544(7650):327–332, 2017.

[93] Karnik, S. S., Unal, H., Kemp, J. R., et al. International union of basic and clinical pharmacology. xcix. angiotensin receptors: interpreters of pathophysiological angiotensinergic stimuli. *Pharmacological reviews*, 67(4):754–819, 2015.

[94] Porrello, E. R., Delbridge, L., and Thomas, W. G. The angiotensin ii type 2 (at2) receptor: an enigmatic seven transmembrane receptor. *Frontiers in bioscience (Landmark edition)*, 14:958–972, 2009.

[95] Guimond, M.-O. and Gallo-Payet, N. How does angiotensin at2 receptor activation help neuronal differentiation and improve neuronal pathological situations? *Frontiers in endocrinology*, 3:164, 2012.

[96] Aquila, A., Hunter, M. S., Doak, R. B., et al. Time-resolved protein nanocrystallography using an x-ray free-electron laser. *Opt. Express*, 20(3):2706–2716, Jan 2012.

[97] Kern, J., Alonso-Mori, R., Hellmich, J., et al. Room temperature femtosecond x-ray diffraction of photosystem ii microcrystals. *Proceedings of the National Academy of Sciences*, 109(25):9721–9726, 2012.

[98] Kupitz, C., Basu, S., Grotjohann, I., et al. Serial time-resolved crystallography of photosystem ii using a femtosecond x-ray laser. *Nature*, 513(7517):261, 2014.

[99] Suga, M., Akita, F., Hirata, K., et al. Native structure of photosystem ii at 1.95 å resolution viewed by femtosecond x-ray pulses. *Nature*, 517(7532):99, 2015.

[100] Suga, M., Akita, F., Yamashita, K., et al. An oxyl/oxo mechanism for oxygen-oxygen coupling in psii revealed by an x-ray free-electron laser. *Science*, 366(6463):334–338, 2019.

[101] Ayyer, K., Yefanov, O. M., Oberthür, D., et al. Macromolecular diffractive imaging using imperfect crystals. *Nature*, 530(7589):202, 2016.

[102] Morgan, A. J., Ayyer, K., Barty, A., et al. Ab initio phasing of the diffraction of crystals with translational disorder. *Acta Crystallographica Section A: Foundations and Advances*, 75(1):25–40, 2019.

[103] Kiefersauer, R., Than, M. E., Dobbek, H., et al. A novel free-mounting system for protein crystals: transformation and improvement of diffraction power by accurately controlled humidity changes. *Journal of applied crystallography*, 33(5):1223–1230, 2000.

[104] Sayre, D. The squaring method: a new method for phase determination. *Acta Crystallographica*, 5(1):60–65, 1952.

[105] Bragg, W. L. and Perutz, M. F. The structure of haemoglobin. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 213(1115):425–435, 1952.

[106] Dinapoli, R., Bergamaschi, A., Henrich, B., 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.

[107] Weinert, T., Olieric, N., Cheng, R., et al. Serial millisecond crystallography for routine room-temperature structure determination at synchrotrons. *Nature Communications*, 8(1), 2017.

[108] Schlichting, I., Almo, S. C., Rapp, G., et al. Time-resolved X-ray crystallographic study of the conformational change in Ha-Ras p21 protein on GTP hydrolysis. *Nature*, 345(6273):309–315, 1990.

[109] Stoddard, B. L., Cohen, B. E., Brubaker, M., Mesecar, A. D., and Koshland, D. E. Millisecond Laue structures of an enzyme–product complex using photocaged substrate analogs. *Nature Structural Biology*, 5(10):891–897, 2002.

[110] Ren, Z. How to process laue data using precognition. 01 2008.

[111] Ren, Z., Bourgeois, D., Helliwell, J. R., et al. Laue crystallography: Coming of age. *Journal of Synchrotron Radiation*, 6(4):891–917, 1999.

[112] Ren, Z. and Moffat, K. Laue Crystallography for Studying Rapid Reactions. *Journal of Synchrotron Radiation*, 1(1):78–82, 2002.

[113] Shrive, A. K., Clifton, I. J., Hajdu, J., and Greenhough, T. J. Laue film integration and deconvolution of spatially overlapping reflections. *Journal of Applied Crystallography*, 23(3):169–174, Jun 1990.

[114] Ren, Z. and Moffat, K. Deconvolution of energy overlaps in laue diffraction. *Journal of applied crystallography*, 28(5):482–494, 1995.

[115] Martin-Garcia, J. M., Zhu, L., Mendez, D., et al. High-viscosity injector-based pink-beam serial crystallography of microcrystals at a synchrotron radiation source. *IUCrJ*, 6(3), 2019.

[116] Tolstikova, A., Levantino, M., Yefanov, O., et al. 1 kHz fixed-target serial crystallography using a multilayer monochromator and an integrating pixel detector. *IUCrJ*, 6(5):927–937, Sep 2019.

[117] Lieske, J., Cerv, M., Kreida, S., et al. On-chip crystallization for serial crystallography experiments and on-chip ligand-binding studies. *IUCrJ*, 6(4), Jul 2019.

[118] Cammarata, M., Eybert, L., Ewald, F., et al. Chopper system for time resolved experiments with synchrotron radiation. *Review of Scientific Instruments*, 80(1):015101, 2009.

[119] Redford, S., Andrä, M., Barten, R., et al. First full dynamic range calibration of the JUNGFRAU photon detector. In *Journal of Instrumentation*, volume 13, page C01027, 2018.

[120] Adams, P. D., Afonine, P. V., Bunkóczi, G., et al. PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):213–221, 2010.

[121] Masuda, T., Suzuki, M., Inoue, S., et al. Atomic resolution structure of serine protease proteinase K at ambient temperature. *Scientific Reports*, 7:45604, 2017.

[122] Sauter, C., Otálora, F., Gavira, J. A., et al. Structure of tetragonal hen egg-white lysozyme at 0.94 Å from crystals grown by the counter-diffusion method. *Acta Crystallographica Section D: Biological Crystallography*, 57(8):1119–1126, 2001.

[123] Meents, A., Gutmann, S., Wagner, A., and Schulze-Briese, C. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proceedings of the National Academy of Sciences*, 107(3):1094–1099, 2009.

[124] Seitlich, T., Kühnel, K., Schulze-Briese, C., Shoeman, R. L., and Schlichting, I. Cryoradiolytic reduction of crystalline heme proteins: analysis by UV-Vis spectroscopy and X-ray crystallography. In *Journal of Synchrotron Radiation*, volume 14, pages 11–23, 2007.

[125] Yano, J., Sauer, K., Pushkar, Y., et al. X-ray damage to the Mn4Ca complex in single crystals of photosystem II: A case study for metalloprotein crystallography. *Proceedings of the National Academy of Sciences*, 102(34):12047–12052, 2005.

[126] Corbett, M. C., Latimer, M. J., Poulos, T. L., et al. Photoreduction of the active site of the metalloprotein putidaredoxin by synchrotron radiation. *Acta Crystallographica Section D: Biological Crystallography*, 63(9):951–960, 2007.

[127] Campbell, J. and Hao, Q. Evaluation of reflection intensities for the components of multiple laue diffraction spots. ii. using the wavelength-normalization curve. *Acta Crystallographica Section A: Foundations of Crystallography*, 49(6):889–893, 1993.

[128] Follath, R., Flechsig, U., Milne, C., et al. Optical design of the aramis-beamlines at swissfel. In *AIP Conference Proceedings*, volume 1741, page 020009. AIP Publishing, 2016.

List of publications

[1] **A. Tolstikova**, M. Levantino, O. Yefanov, V. Hennicke, P. Fischer, J. Meyer, A. Mozzanica, S. Redford, E. Crosas, N. L. Opara, M. Barthelmess, J. Lieske, D. Oberthuer, E. Wator, I. Mohacsi, M. Wulff, B. Schmitt, H. N. Chapman, and A. Meents. 1 kHz fixed-target serial crystallography using a multilayer monochromator and an integrating pixel detector. *IUCrJ*, 6(5):927–937, Sep 2019.

[2] A. Meents, M. O. Wiedorn, V. Srajer, R. Henning, I. Sarrou, J. Bergtholdt, M. Barthelmess, P. Y.A. Reinke, D. Dierksmeyer, **A. Tolstikova**, S. Schaible, M. Messerschmidt, C. M. Ogata, D. J. Kissick, M. H. Taft, D. J. Manstein, J. Lieske, D. Oberthuer, R. F. Fischetti, and H. N. Chapman. Pink-beam serial crystallography. *Nature Communications*, 8(1):1281, 2017.

[3] H. Zhang, G. W. Han, A. Batyuk, A. Ishchenko, K. L. White, N. Patel, A. Sadybekov, B. Zamlynny, M. T. Rudd, K. Hollenstein, **A. Tolstikova**, T. A. White, M. S. Hunter, U. Weierstall, W. Liu, K. Babaoglu, E. L. Moore, R. D. Katz, J. M. Shipman, M. Garcia-Calvo, S. Sharma, P. Sheth, S. M. Soisson, R. C. Stevens, V. Katritch, and V. Cherezov. Structural basis for selectivity and diversity in angiotensin II receptors. *Nature*, 544(7650):327–332, 2017.

List of additional publications

[1] Y. Gevorkov, A. Barty, W. Brehm, T.A. White, **A. Tolstikova**, M.O. Wiedorn, A. Meents, R.-R. Grigat, H.N. Chapman, and O. 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.

[2] S. Pandey, R. Bean, T. Sato, I. Poudyal, J. Bielecki, J. Cruz Villarreal, O. Yefanov, V. Mariani, T.A. White, C. Kupitz, M. Hunter, M.H. Abdellatif, S. Bajt, V. Bondar, A. Echelmeier, D. Doppler, M. Emons, M. Frank, R. Fromme, Y. Gevorkov, G. Giovanetti, M. Jiang, D. Kim, Y. Kim, H. Kirkwood, A. Klimovskaia, J. Knoska, F.H.M. Koua, R. Letrun, S. Lisova, L. Maia, V. Mazalova, D. Meza, T. Michelat, A. Ourmazd, G. Palmer, M. Ramilli, R. Schubert, P. Schwander, A. Silenzi, J. Sztuk-Dambietz, **A. Tolstikova**, H.N. Chapman, A. Ros, A. Barty, P. Fromme, A.P. Mancuso, and M. Schmidt. Time-resolved serial femtosecond crystallography at the european xfel. *Nature Methods*, 17(1):73–78, 2020.

[3] O. Yefanov, D. Oberthür, R. Bean, M.O. Wiedorn, J. Knoska, G. Pena, S. Awel, L. Gumprecht, M. Domaracky, I. Sarrou, P. Lourdu Xavier, M. Metz, S. Bajt, V. Mariani, Y. Gevorkov, T. White, **A. Tolstikova**, P. Villanueva-Perez, C. Seuring, S. Aplin, A. Estillore, J. Küpper, A. Klujev, M. Kuhn, T. Laurus, H. Graafsma, D. Monteiro, M. Trebbin, F. Maia, F. Cruz-Mazo, A. Ganan-Calvo, M. Heymann, C. Darmanin, B. Abbey, M. Schmidt, P. Fromme, K. Giewekemeyer, M. Sikorski, R. Graceffa, P. Vagovic, T. Kluyver, M. Bergemann, H. Fangohr, J. Sztuk-Dambietz, S. Hauf, N. Raab, V. Bondar, A.P. Mancuso, H.N. Chapman, and A. Barty. Evaluation of serial crystallographic structure determination within megahertz pulse trains. *Structural dynamics*, 6(6):064702, 2019.

[4] P. Lindenberg, L.R. Arana, L.K. Mahnke, P. Rönfeldt, N. Heidenreich, G. Doungmo, N. Guignot, R. Bean, H.N. Chapman, D. Dierksmeyer, J. Knoska, M. Kuhn, J. Garrevoet, V. Mariani, D. Oberthuer, K. Pande, S. Stern, **A. Tolstikova**, T.A. White, K.R. Beyerlein, and H. Terraschke. New insights into the crystallization of polymorphic materials: From real-time serial crystallography to luminescence analysis. *Reaction Chemistry and Engineering*, 4(10):1757–1767, 2019.

[5] Y. Gevorkov, O. Yefanov, A. Barty, T.A. White, V. Mariani, W. Brehm, **A. Tolstikova**, R.-R. Grigat, and H.N. Chapman. Xgandalf - extended gradient descent algorithm for lattice finding. *Acta Crystallographica Section A: Foundations and Advances*, 75:694–704, 2019.

[6] M.O. Wiedorn, D. Oberthür, R. Bean, R. Schubert, N. Werner, B. Abbey, M. Aepfelbacher, L. Adriano, A. Allahgholi, N. Al-Qudami, J. Andreasson, S. Aplin, S. Awel, K. Ayyer, S. Bajt, I. Barák, S. Bari, J. Bielecki, S. Botha, D. Boukhelef, W. Brehm, S. Brockhauser, I. Cheviakov, M.A. Coleman, F. Cruz-Mazo, C. Danilevski, C. Darmanin, R.B. Doak, M. Domaracky, K. Dörner, Y. Du, H. Fangohr, H. Fleckenstein, M. Frank, P. Fromme, A.M. Gañán-Calvo, Y. Gevorkov, K. Giewekemeyer, H.M. Ginn, H. Graafsma, R. Graceffa, D. Greiffenberg, L. Gumprecht, P. Göttlicher, J. Hajdu, S. Hauf, M. Heymann, S. Holmes, D.A. Horke, M.S. Hunter, S. Imlau, A. Kaukher, Y. Kim, A. Klyuev, J. Knoška, B. Kobe, M. Kuhn, C. Kupitz, J. Küpper, J.M. Lahey-Rudolph, T. Laurus, K. Le Cong, R. Letrun, P.L. Xavier, L. Maia, F.R.N.C. Maia, V. Mariani, M. Messerschmidt, M. Metz, D. Mezza, T. Michelat, G. Mills, D.C.F. Monteiro, A. Morgan, K. Mühlig, A. Munke, A. Münnich, J. Nette, K.A. Nugent, T. Nuguid, A.M. Orville, S. Pandey, G. Pena, P. Villanueva-Perez, J. Poehlsen, G. Previtali, L. Redecke, W.M. Riekehr, H. Rohde, A. Round, T. Safenreiter, I. Sarrou, T. Sato, M. Schmidt, B. Schmitt, R. Schönherr, J. Schulz, J.A. Sellberg, M.M. Seibert, C. Seuring, M.L. Shelby, R.L. Shoeman, M. Sikorski, A. Silenzi, C.A. Stan, X. Shi, S. Stern, J. Sztuk-Dambietz, J. Szuba, **A. Tolstikova**, M. Trebbin, U. Trunk, P. Vagovic, T. Ve, B. Weinhausen, T.A. White, K. Wrona, C. Xu, O. Yefanov, N. Zatsepin, J. Zhang, M. Perbandt, A.P. Mancuso, C. Betzel, H. Chapman, and A. Barty. Megahertz serial crystallography. *Nature Communications*, 9(1):4025, 2018.

[7] S. Awel, R.A. Kirian, M.O. Wiedorn, K.R. Beyerlein, N. Roth, D.A. Horke, D. Oberthür, J. Knoska, V. Mariani, A. Morgan, L. Adriano, **A. Tolstikova**, P.L. Xavier, O. Yefanov, A. Aquila, A. Barty, S. Roy-Chowdhury, M.S. Hunter, D. James, J.S. Robinson, U. Weierstall, A.V. Rode, S. Bajt, J. Küpper, and H.N. Chapman. Femtosecond x-ray diffraction from an aerosolized beam of protein nanocrystals:. *Journal of Applied Crystallography*, 51(1):133–139, 2018.

[8] C. Kupitz, J.L. Olmos, M. Holl, L. Tremblay, L. Pande, S. Pandey, D. Oberthür, M. Hunter, M. Liang, A. Aquila, J. Tenboer, G. Calvey, A. Katz, Y. Chen, M.O. Wiedorn, J. Knoska, A. Meents, V. Majriani, T. Norwood, I. Poudyal, T. Grant, M.D. Miller, W. Xu, **A. Tolstikova**, A. Morgan, M. Metz, J.M. Martín-García, J.D. Zook, S. Roy-Chowdhury, J. Coe, N. Nagaratnam, D. Meza, R. Fromme, S. Basu, M. Frank, T. White, A. Barty, S. Bajt, O. Yefanov, H.N. Chapman, N. Zatsepin, G. Nelson, U. Weierstall, J. Spence, P. Schwander, L. Pollack, P. Fromme, A. Ourmazd, G.N. Phillips, and M. Schmidt. Structural enzymology using x-ray free electron lasers. *Structural Dynamics*, 4(4):044003, 2017.

[9] D. Oberthuer, J. Knoška, M.O. Wiedorn, K.R. Beyerlein, D.A. Bushnell, E.G. Kovaleva, M. Heymann, L. Gumprecht, R.A. Kirian, A. Barty, V. Mariani, **A. Tolstikova**, L. Adriano, S. Awel, M. Barthelmess, K. Dörner, P.L. Xavier, O. Yefanov, D.R. James, G. Nelson, D. Wang, G. Calvey, Y. Chen, A. Schmidt, M. Szczepek, S. Frielingsdorf, O. Lenz, E. Snell, P.J. Robinson, B. Šarler, G. Belšak, M. Maček, F. Wilde, A. Aquila, S. Boutet, M. Liang, M.S. Hunter, P. Scheerer, J.D. Lipscomb, U. Weierstall, R.D. Kornberg, J.C.H. Spence, L. Pollack, H.N. Chapman, and S. Bajt. Double-flow focused liquid injector for efficient serial femtosecond crystallography. *Scientific Reports*, 7:44628, 2017.

[10] K.R. Beyerlein, D. Dierksmeyer, V. Mariani, M. Kuhn, I. Sarrou, A. Ottaviano, S. Awel, J. Knoska, S. Fuglerud, O. Jönsson, S. Stern, M.O. Wiedorn, O. Yefanov, L. Adriano, R. Bean, A. Burkhardt, P. Fischer, M. Heymann, D.A. Horke, K.E.J. Jungnickel, E. Kovaleva, O. Lorbeer, M. Metz, J. Meyer,

A. Morgan, K. Pande, S. Panneerselvam, C. Seuring, **A. Tolstikova**, J. Lieske, S. Aplin, M. Roessle, T.A. White, H.N. Chapman, A. Meents, and D. Oberthuer. Mix-and-diffuse serial synchrotron crystallography. *IUCrJ*, 4:769–777, 2017.

[11] T.A. White, V. Mariani, W. Brehm, O. Yefanov, A. Barty, K.R. Beyerlein, F. Chervinskii, L. Galli, C. Gati, T. Nakane, **A. Tolstikova**, K. Yamashita, C.H. Yoon, K. Diederichs, and H.N. Chapman. Recent developments in crystfel. *Journal of Applied Crystallography*, 49:680–689, 2016.

Acknowledgments

I would like to start by thanking my Master's supervisor, Salavat Khasanov, who introduced me to the field of crystallography and taught me the fundamentals, which was incredibly beneficial especially at the beginning of my PhD.

I would like to express my gratitude to Henry Chapman for giving me the opportunity to join his group and for the freedom and support that I received during my time there to pursue my ideas. My special thanks go to my second supervisor, Thomas White, for his guidance throughout my research, numerous helpful discussions and his invaluable comments to this thesis.

It has been a great pleasure working with all the former and current members of the Coherent Imaging Division at CFEL. In particular, I would like to thank Yaroslav Gevorkov and Wolfgang Brehm for our joint work on the development of the pink-beam data analysis and many good discussions on the subject. I would like to further thank Dominik Oberthür for his instructions and help with structure refinements. I thank Oleksandr Yefanov for his generous help in all areas, for an endless supply of cookies and for never failing to find extra work for me to do. I also thank Valerio Mariani and Kanupriya Pande for their support and unfailing sense of humour which helped me a lot during my PhD.

I would like to specially thank Alke Meents for making all the pink-beam experiments possible and inviting me to join them. I am grateful to all the people who have participated in the experiments presented in the thesis. In particular, I would like to thank beamline scientists, Matteo Levantino and Michael Wulff from ID09 (ESRF) and Vukica Šrajer and Robert Henning from BioCARS (APS), for their great assistance with the experiments and discussions about the data analysis.

Finally, I would like to thank all my friends and my dog Bender for the emotional support and encouragement during the writing of this thesis and, most importantly, I thank my family, my mom and grandma, for their constant love and support.

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 13.03.2020

Aleksandra Tolstikova