Universidade Estadual de Campinas

Instituto de Física "Gleb Wataghin"

Ana Carolina Rodrigues

Development of an automated data processing pipeline for Serial Synchrotron Crystallography on MANACÁ (Sirius)

Desenvolvimento do fluxo de processamento automático de dados para Cristalografia Serial em Síncrotron na MANACÁ (Sirius)

campinas
2021

Ana Carolina Rodrigues

Development of an automated data processing pipeline for Serial Synchrotron Crystallography on MANACÁ (Sirius)

Desenvolvimento do fluxo de processamento automático de dados para Cristalografia Serial em Síncrotron na MANACÁ (Sirius)

Dissertation presented to the Institute of Physics "Gleb Wataghin"at University of Campinas in partial fulfillment of the requirements for the degree of Master in Physics, in the area of Physics.

Dissertação apresentada ao Instituto de Física "Gleb Wataghin"da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Física, na Àrea de Física.

Supervisor/Orientador: Ana Carolina de Mattos Zeri

Co-supervisor/Coorientador: Eduardo Granado Monteiro da Silva

Este trabalho corresponde à versão final da dissertação defendida pela aluna Ana Carolina Rodrigues, e orientada pela Profa. Dra. Ana Carolina de Mattos Zeri.

Campinas
2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Física Gleb Wataghin
Lucimeire de Oliveira Silva da Rocha - CRB 8/9174

R618d

Rodrigues, Ana Carolina, 1996-
Development of an automated data processing pipeline for Serial Synchrotron Crystallography on MANACÁ (Sirius) / Ana Carolina Rodrigues. – Campinas, SP : [s.n.], 2021.

Orientador: Ana Carolina de Mattos Zeri.
Coorientador: Eduardo Granado Monteiro da Silva.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Física Gleb Wataghin.

1. Projeto Sirius. 2. Cristalografia serial em síncrotron. 3. Python (Linguagem de programação de computador). 4. Processamento eletrônico de dados. I. Zeri, Ana Carolina de Mattos. II. Silva, Eduardo Granado Monteiro da, 1974-. III. Universidade Estadual de Campinas. Instituto de Física Gleb Wataghin. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Desenvolvimento do fluxo de processamento automático de dados para Cristalografia Serial em Síncrotron na MANACÁ (Sirius)
Palavras-chave em inglês:
Sirius Project
Serial synchrotron crystallography
Python (Computer program language)
Electronic data processing
Área de concentração: Física
Titulação: Mestra em Física
Banca examinadora:
Ana Carolina de Mattos Zeri [Orientador]
Maria Cristina Nonato Costa
Júlio Criginski Cezar
Data de defesa: 25-08-2021
Programa de Pós-Graduação: Física

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-6180-9522
- Currículo Lattes do autor: http://lattes.cnpq.br/2201008890458407

MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DA ANA CAROLINA RODRIGUES - RA 154599 APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA “GLEB WATAGHIN”, DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 25/ 08 / 2021.

COMISSÃO JULGADORA:

- Profa. Dra. Ana Carolina de Mattos Zeri - Orientadora (CNPEM/LNLS)

- Profa. Dra. Maria Cristina Nonato Costa (USP)

- Prof. Dr. Júlio Criginski Cézar (CNPEM/LNLS)

OBS.: Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

CAMPINAS

2021

To my father, my family and friends.
To the hardest plants to grow,
and every tear they have dropped in this land.

Acknowledgements

I would like to thank my family, without them I couldn't be here. It was on the hardest times that we've learned most with each other. My father Joaquim, thank you seems not enough words to express my feelings, but we are better at feelings than words and that is totally fine. You and my mother, Maria José, have noticed the light in me and showed me how to better use my potential, even when I couldn't understand well your teachings. In my graduation speech I said that you took the most difficult responsibility in raising a child, that is showing how life is hard sometimes. Now, I would like to add that you also taught me that humility is more important than everything that a human being might have knowledge. Humility to use your acknowledgment, without hurting anyone. I know I failed sometimes on that, but I love you two with all my heart and I will carry these teachings wherever I am. You will always be with me and I will be with you too.

My sister Ana Clara, you can always count on me. We have a long path together and you have always listened to me, even on my most boring times. Let's walk together and we will be giants! To all my aunts, uncles and cousins: I love you so much, I'm sorry for being so distant sometimes, but the memories I have from you on my side, whenever I needed, are always on my mind.

I would like to thank my advisor Prof. Dr. Ana Carolina de Mattos Zeri, for trusting me since the beginning of my career of crystallography. You must have been pretty brave on my first day of internship when I said "I know absolutely nothing about crystallography". You were very patient and took off all the pressure on me in the biggest decisions of my career. Thank you for the opportunity to work on the beautiful project of Manacá beamline. We are dreamers who put words in practice and warriors ready for all battles.

Thank you so much, Prof. Dr. Eduardo Granado, for always being open-minded for us. You received the project with open arms and aggregated a lot to it. I would be much more lost during the pandemic year, if it wasn't for your advises and our weekly meetings.

I would also like to thank CNPEM, mainly LNLS, for being a home for me since the beginning of my scientific career. Thanks for being a unique place for doing research in Brazil. The freedom and support all staffs gives to the scientists are extremely inspiring to me. Special thanks to Prof. Dr. Júlio Criginski Cezar for not giving up on me, even when I was too lost to understand that I owned a undergraduate research project at LNLS. You were the hand that pushed me up and I am very grateful to you for the time I spent at PGM beamline.

I am very grateful to CNPq for the research funding (143239/2019-8). The financial support from Brazilian's funding research agencies are essential for young scientists to start their career. They have to be valued and their continuity must be guaranteed for future scientists.

Finally, but not less important, thanks to all my friends that have always been a big support on my life. Special thanks to Gabriel Minoru, Letícia and Paloma for all revisions and emotional force during the writing of this dissertation and all great moments we had together. I love you! To my biggest friends, Alencar e Marlene, that has accompanied my growth, were kind with me and showed me the light I have inside me. We are above, and this has its beauty too.

To my therapists, Rosângela and Maria Lúcia, and all my psychiatrists, you were completely important for me to be alive here and to possess the self-knowledge I have today.

Nala, you are sou unique. In this moment I'm writing and I see you taking your daily nap on my bed. You are so important to me as a dog and as my friend for all moments. I promise to love you and carry you with me wherever I go. I will always be here for you, my heart.

Thank you for reading this excessively long acknowledgment until the end.

This work was financed by the "Ministry of Science, Technology and Innovation" and the "National Council for Scientific and Technological Development – CNPq".

Resumo

Manacá (MAcromolecular micro and NAno CrystAllography), primeira linha de luz em fase de comissionamento no Sirius (LNLS/CNPEM), é dedicada a diversos experimentos de cristalografia aplicadas em pequenas e macromoléculas, incluindo a recente técnica de cristalografia serial (SX). SX utiliza o feixe de microfoco e alto fluxo de fontes de luz avançadas, como o Sirius, para adquirir padrões únicos de difração de milhares de cristais por experimento, possivelmente, à temperatura ambiente.

Um dos maiores desafios para SX é a união de conjuntos de dados parciais, aleatoriamente orientados no espaço recíproco, como se tivessem sido obtidos a partir da rotação de cristal único. Quanto ao dado final, o objetivo principal é extrair os melhores resultados dentro de uma variedade de combinações possíveis entre os subconjuntos de dado. Para solucionar isso, tem sido aplicados diferentes algoritmos, cada um com seus respectivos parâmetros a serem ajustados para toda coleta de dados. O objetivo dessa dissertação é fornecer uma rotina de processamento automático de dados para experimentos de SX realizados na Manacá. Construímos um *script*, em *Python*, que agirá como uma interface entre os usuários e os programas que estão sendo desenvolvidos para SX (*CrystFEL* [1], *nXDS*, *cctbx.xfel*, *ccCluster*[2], *BLEND*[3], *xscale_isocluster*, entre outros). Os usuários poderão chamar, de forma prática, diferentes programas de processamento, otimizar os parâmetros que melhor ajustem os seus dados, explorar visualmente as figuras de mérito e escolher as melhores opções de processamento.

A fim de testar nossa rotina de processamento automático de dados, realizamos três experimentos durante o comissionamento científico da Manacá. Primeiramente, foram medidos 21 cristais de *Ame*GH128, no modo de oscilação (9.15 keV, tamanho de feixe variando de 20 a 60 μm), em 77K. Desse experimento, obtivemos um total de cerca de 64800 padrões de difração. Em seguida, coletamos 64 cristais de lisozima criocongelados, em modo *grid-scan* (12.69 keV, tamanho do feixe de 25 μm, fluxo de 10^{12} ph/s). Nesse experimento, foram coletadas 2910 imagens no total. O mesmo experimento foi realizado em 20 cristais de lisozima em temperatura ambiente, utilizando um porta-amostra desenvolvido internamente pelo nosso grupo, selado com Kapton. Nessas condiçõs, nós obtivemos um total de 601 imagens.

Os padrões de difrações obtidos foram usados como entrada para duas principais vias de processamento de SX: *Hierarchical Cluster Analysis* HCA (*ccCluster*) e imagens instantâneas (*CrystFEL*). Nosso *script* pôde indexar os padrões de difração com sucesso, de acordo com os parâmetros de célula unitária já conhecidos. Nosso fluxo de processamento de dados é versátil para verificação da qualidade dos dados e ajuste de parâmetros. Na rotina de imagens instantâneas, melhor qualidade final dos dados pode ser atingida, refinando a taxa de indexação e aumentando o número de cristais coletados. Todavia, a ferramenta de processamento de dados de SX é funcional e customizável para os usuários da Manacá.

SX é uma promessa para o estudo de dinâmica de proteínas resolvida no tempo, e na descoberta de proteínas que são difíceis de cristalizar, como estruturas de proteínas de membrana. Possibilitar essa técnica na Manacá terá grande impacto em diversas áreas, desde descobrimento de drogas para tratamentos efetivos de doenças até otimização da produção de biocombustíveis renováveis.

Palavras-chave: Cristalografia serial em síncrotron, projeto Sirius, Python, processamento automático de dados.

Abstract

Manacá (MAcromolecular micro and NAno CrystAllography), first beamline in commissioning at Sirius (LNLS/CNPEM), is dedicated to a range of crystallography experiments applied to small and macro molecules, including the most recent technique of serial crystallography (SX). SX takes advantage of microfocus beam and higher flux from advanced light sources, such as Sirius, to acquire unique diffraction patterns from several tens of thousands of microcrystals per experiment, possibly, at room temperature.

One of the biggest challenges for SX is to merge partial datasets, randomly oriented in the reciprocal space, as if they had been obtained from a single crystal rotation. Regarding final data, the main goal is to extract the best results from a range of all possible subdataset combination. To solve this, it has been applied different algorithms, each with their respective parameters that should be adjusted to every data collection. The aim of this dissertation is to provide an automated data processing pipeline for SX experiments on Manacá. We have built a script, in Python, that will act as an interface between users and new software being developed for SX (*CrystFEL* [1], *nXDS*, *cctbx.xfel*, *ccCluster*[2], *BLEND*[3], *xscale_isocluster*, among others). Users will be able to, with minimal effort, call different data processing programs, adjust parameters that best fit their data, visually explore the figures of merit and choose the best data processing options.

In order to test our data processing pipelines, we have performed three experiments during the Manacá's scienctific commisioning. Firstly, it was measured 21 crystals of *Ame*GH128, in oscillation mode (9.15 keV, beam size varying from 20 to 60 μm), at 77K. From that experiment, we obtained a total of around 64800 diffraction patterns. Afterwards, we collected 64 cryocooled lysozyme crystals, in grid-scan mode (12.69 keV, beam size 25 μm, flux at sample 10^{12} ph/s). In this experiment, it was collected 2910 images in total. The same experiment was done on 20 lysozyme crystals at room-temperature, using an in-house build sample device, sealed with Kapton. In this condition, we obtained a total of 601 images.

The diffraction patterns were used as an input for the two main branches of SX data processing: Hierarchical Cluster Analysis HCA (*ccCluster*) and snapshot images routine (*CrystFEL*). Our script could successfully index the diffraction patterns, according to the already known unit cell parameters. Our SX automatic data processing pipeline is versatile for immediate data quality verification and software parameters adjustments. In the snapshot routine, better final data quality might be achieved by tuning the indexing rate and increasing the number of crystals measured. Nevertheless, the SX data processing tool is functional and customizable for Manacá future users.

Serial Crystallography is a huge promise for the study of time-resolved protein dynamics and in the revealing of proteins structures that are extremely difficult to crystallize, as membrane protein structures. Enabling this technique on Manacá will have a great impact in a variety of fields, from drug discovery for effective treatment of diseases to the optimization of renewable biofuels production.

Keywords: Serial Synchrotron Crystallography, project Sirius, Python, automatic data processing

List of Figures

List of Tables

List of Abbreviations

CC Pearson's correlation coefficient

CNPEM Centro Nacional de Pesquisa em Energia e Materiais

DCM double crystal monochromator

GA Genetic Algorithms

GDVN gas dynamic virtual nozzles

HCA Hierarchical Clustering Analysis

KB Kirkpatrick–Baez

LCP lipidic cubic phase

LNBr Laboratório Nacional de Biorrenováveis

LNBio Laboratório Nacional de Biotecnologia

LNLS Laboratório Nacional de Luz Síncrotron

LNNano Laboratório Nacional de Nanotecnologia

MAD Multiple Anomalous Diffraction

MIR Multiple Isomorphous Replacement

PDB Protein Data Bank

SAD Single Anomalous Diffraction

SASE Self-Amplified Spontaneous Emission

SIR Single Isomorphous Replacement

SNR Signal to Noise Ratio

SFX Serial Femtosecond Crystallography

SSX Serial Synchrotron Crystallography.

SX Serial Crystallography

XFELs X-ray Free Electron Lasers

Contents

Chapter 1

Motivation

X-ray crystallography is a well-established technique, initially developed by M. von Laue (1914) and W. H. Bragg and W. L. Bragg (1915). Since that, it has incorporated a number of technological advances that have been essential for the expansion and establishment of the technique for over a hundred years. From the discovery of enzymes and virus crystallization, in 1946, by Stanley W. M, Northrop J. H. e Sumner J. B – which were awarded the Chemistry Nobel Prize – biological samples became of great interest for X-ray crystallography.

The first protein structures solved were myoglobin and hemoglobin, by Max Perutz and John Kendrew, Nobel Prize of Chemistry in 1962. Their work opened doors for the development of the field of macromolecular crystallography and they are one of the pioneering structures that set the foundation for the Protein Data Bank (PDB) archive, in 1971. In this year of 2021, PDB is completing 50 years and approaches the number of a hundred-eighty thousand structures deposited. The increase in number of entries is directly impacted by technological advances that have emerged since the 70s, as can be seen in figure 1.1.

Greater availability of X-ray sources, such as synchrotron sources, use of cryogenics to maintain samples, and computational developments have a huge importance on the technique development. However, in the last twenty years, we have faced a stability on the number of structures deposited per year, implying a stability on the technique capability. In order to continue expanding, crystallography has to be updated with the most recent developments originated by new X-ray sources, as X-ray Free Electron Lasers (XFELs) and fourth generation synchrotrons like Sirius (LNLS-CNPEM, Brazil).

Nowadays, many macromolecular beamlines around the world aim fore complete experiment automation, using robotic arms for sample delivery and remote data collection. Which improves experiments efficiency and reduces its costs from displacement.

The new sources, which are extremely brilliant, impose a certain limitation regarding radiation damage, especially for biological samples. In this context, it was revived the concept of measuring multiple crystals to solve a structure as it is allowed to achieve only a few patterns, or in the case of XFELs, a unique pattern from each crystal (diffraction-before-destruction). Besides that, a higher flux and smaller spot size can provide satisfactory intensity reflections, even for small crystals or big unit cell ones. In both cases, they would have been discarded from experiments due to bad diffraction performance. All of that set a base on the development of a recent crystallography

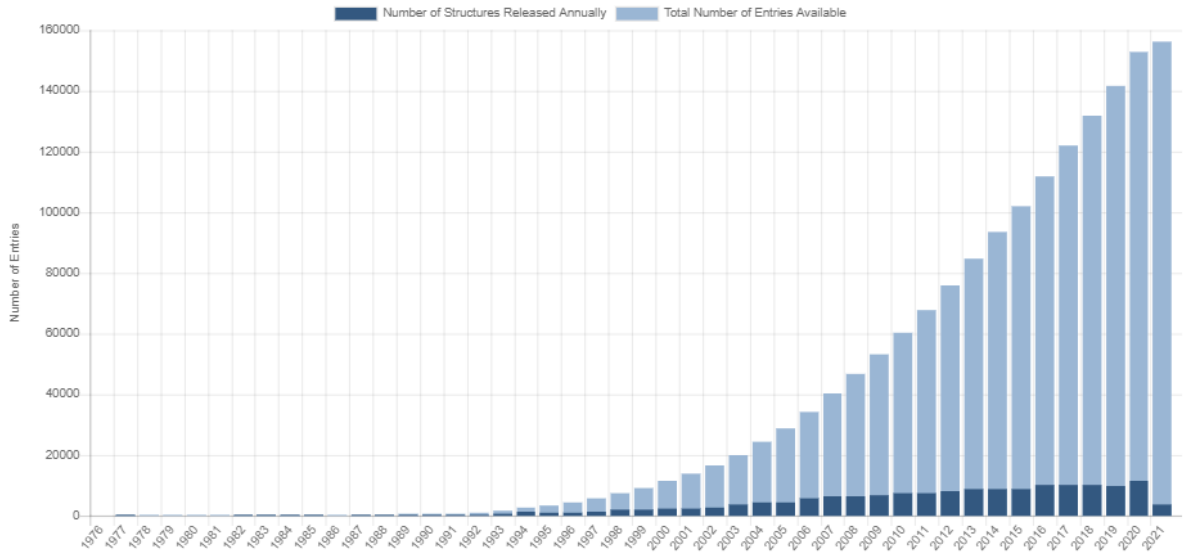


Figure 1.1: PDB Statistics: Growth of Structures from X-ray Crystallography Experiments Released per Year

technique named serial crystallography (SX).

Serial crystallography is a promising technique to allow experiments in large scale, as massive X-ray screening of drugs, time-resolved crystallography and room-temperature measurements. Added to the recent advances on protein folding prediction with high accuracy by Alphafold2[4] and high resolution electron cryo-microscopy, upgrades in X-ray crystallography, such as SX, certainly have a lot to contribute to the expansion of our knowledge on biological process.

The aim of this dissertation is to allow Manacá users to automatically process their data, especially when dealing with SX experiments. We have developed an automatic data processing pipeline , written in Python, that establishes communication with SX packages, mainly CrystFEL and ccCluster. The routine is flexible to software parameters optimization, and shows final data quality statistics, assisting users to take decisions during the experiment.

Chapters 2 and 3 give a general background on X-ray crystallography theory and experimental methods. Chapter 4 exposes the recent developments on serial crystallography data processing. Chapter 5 is dedicated to explore the main features of Manacá beamline at Sirius. In chapter 6, it is described the development of an automatic SX data processing pipeline. The pipeline was evaluated in the analysis of data sets collected during Manacá commissioning, as it is showed in chapter 6. Chapter 7 summarizes the main results of this dissertation and gives an outlook on future developments.

Codes are available at:
https://github.com/anananacr/HCAmanaca.git (*ccCluster*)
https://github.com/anananacr/SSXmanaca.git (*CrystFEL*).

Chapter 2

Introduction to X-ray crystallography

This chapter gives a general background on X-ray crystallography theory, in sequential order since data acquisition until structure solution.

2.1 Diffraction from periodic structures

2.1.1 The crystal lattice

X-ray crystallography is based on scattering of coherent electromagnetic waves by the electronic density of atoms in a crystal. The set o vectors $\mathbf{r}_n$ (eq. 2.1.1), that corresponds to the atom positions in a crystal, is called direct or real lattice:

$$\mathbf{r}_n = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3, \tag{2.1.1}$$

where n_i (i=1, 2, 3) are integers and $\mathbf{a}_i$ (i=1, 2, 3) are the elementary translations of the direct lattice. This vector defines the unit cell of the lattice, also written in terms of its length a, b, c, and the respective angles between them (α, β, γ) as figure 2.1 shows.

Molecular crystals are composed from units units of small molecules (tens of atoms per molecule) or macro molecules (thousands of atoms per molecule) packed together by weak dispersion or bipo-

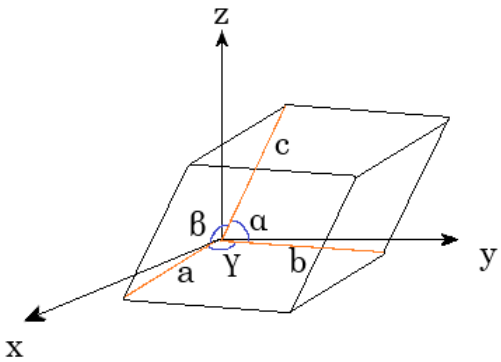


Figure 2.1: Unit cell parameters representation of a direct lattice.

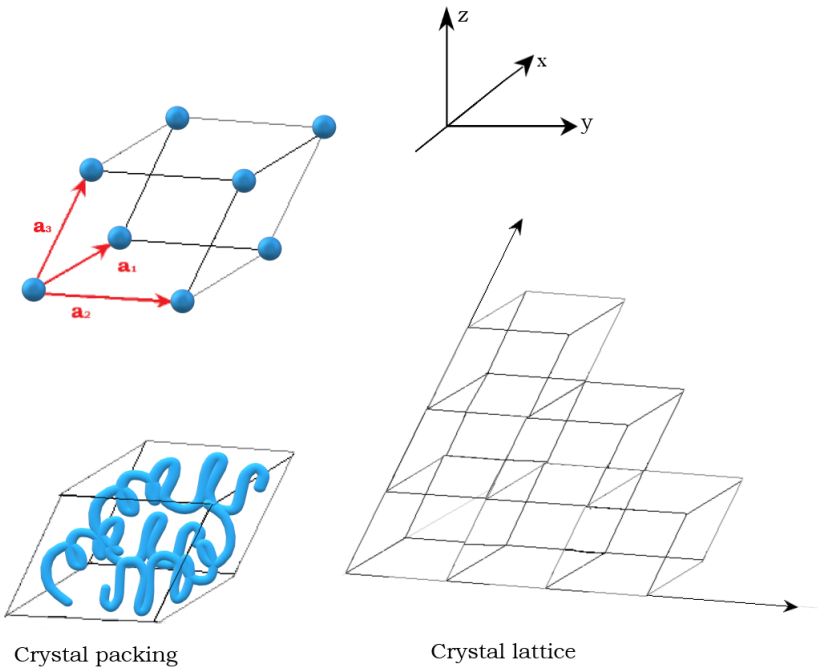


Figure 2.2: Crystal packing and crystal lattice representation.

lar forces. Even for molecules, crystal packing definition still can be applied to described a direct lattice basis, given the pattern that repeats periodically in all directions (figure 2.2). However, differently from most ionic and covalent crystals, the interest, here, is more on the molecular structure that composes the unit cell, rather than the crystal structure itself[5].

In order to obtain the molecular structure, we begin defining the scattering density function $\rho(\mathbf{r})$, that will give rise to the X-ray interaction. Given its periodicity, it can be expanded in a Fourier Series (eq. 2.1.2). Therefore, we can expect the exponential part to have its maximum value when the space vector $\mathbf{r}$ coincides with a direct lattice vector in $\mathbf{r}_n$. This fact imposes a restriction on the set of wave vector $\mathbf{G}$ from $\rho(\mathbf{r})$ (eq. 2.1.3).

$$\rho(\mathbf{r}) = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{2.1.2}$$

$$\mathbf{G}\cdot\mathbf{r}_n = 2\pi m, \tag{2.1.3}$$

where m is an integer.

If we write $\mathbf{G}$ on a vector basis $\mathbf{g}_i$ (eq. 2.1.4), there will be a unidirectional and unequivocally correspondence between the direct lattice vectors, $\mathbf{r}_n$, and the reciprocal lattice, $\mathbf{G}$ (eq. 2.1.5). The vector $\mathbf{G}$ can be defined unambiguously by its coordinates h,k,l and it is commonly used to label diffracted beams, which means reflections recorded by the detector.

$$\mathbf{G} = h\mathbf{g}_1 + k\mathbf{g}_2 + l\mathbf{g}_3, \tag{2.1.4}$$

where h, k, l are integers.

$$\mathbf{g}_i \cdot \mathbf{a}_j = 2\pi\delta_{ij} \tag{2.1.5}$$

with i, j=1, 2, 3.

Equation 2.1.5 can also be understood in terms of

$$\mathbf{g}_1 = 2\pi \frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)} \tag{2.1.6}$$

and its cyclic permutations.

Using the definitions above, the X-ray scattering intensity will be given by equation 2.1.7, where $\mathbf{K}$=$\mathbf{k}$-$\mathbf{k}_0$ [6] is the scattering vector, $\mathbf{k}$ and $\mathbf{k}_0$, are the scattered and incident wave vectors, respectively.

$$I(\mathbf{K}) \propto \left|\frac{A_0}{R'}\right|^2 \left|\sum_{\mathbf{G}} \rho_{\mathbf{G}} \int e^{i(\mathbf{G}-\mathbf{K})\cdot\mathbf{r}} d\mathbf{r}\right|^2 \tag{2.1.7}$$

For a crystal composed of many identical unit cells, the main contribution for the integral in equation 2.1.7 is when $\mathbf{G}$=$\mathbf{K}$. The integral can be seen as an approximate representation of the δ-functions, so that

$$\int e^{i(\mathbf{G}-\mathbf{K})\cdot\mathbf{r}} d\mathbf{r} = \begin{cases} \text{V, for } \mathbf{K}=\mathbf{G} \\ \sim 0 \text{ otherwise} \end{cases} \tag{2.1.8}$$

The Laue condition $\mathbf{K}$=$\mathbf{G}$ can be represented by the Ewald construction. We select an arbitrary point of the reciprocal lattice as the origin, and the incident wave vector $\mathbf{k}_0$ is drawn pointing towards the origin. Considering an elastic scattering regime, we have $k = k_0 = \frac{2\pi}{\lambda}$, λ is the radiation wavelength. All possible scattered wave vectors centered around the starting of $\mathbf{k}_0$, set up an sphere on the reciprocal space, which is called the Ewald sphere (fig. 2.3).

A vector that connects $\mathbf{k}$ and $\mathbf{k}_0$ ends, the subtraction between them, is exactly equal to the scattering vector $\mathbf{K}$. The Laue condition says a reflection will happens when $\mathbf{K}$=$\mathbf{G}$, that means as reciprocal lattice points cross the surface of the Ewald sphere.

As figure 2.3 shows, not all reciprocal points cross the sphere surface at the same time. In order to obtain information from the whole reciprocal space, you have to either use a continuum of wavelengths or vary crystal orientation. In theory, reciprocal lattice points and Ewald sphere thickness are infinitely small, but due to some aspects from the beam and the real crystals, the reflections shape might be distorted. This gives rise to reflection partiality, that shall be better discussed further.

Here, we will present the interpretation of the Laue condition, but it's worth to mention that the Bragg condition interpretation is equally valid.

2.1.2 The structure factor

The scattering condition only gives the position where a reflection will appear. In order to obtain their integrated intensity, we need first to calculate the Fourier coefficients ρ_{hkl} of the

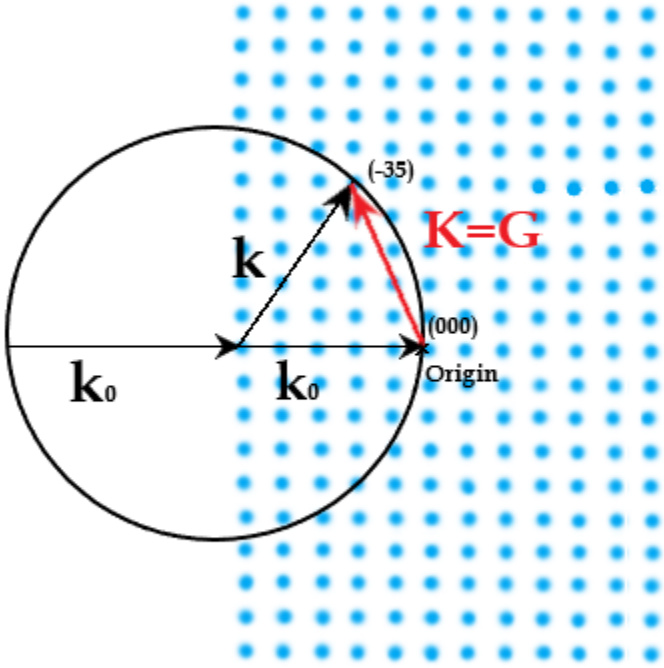


Figure 2.3: Ewald sphere representation and the reciprocal lattice representation, highlighting a diffraction beam that satisfies the Laue condition.

electronic density. Those are calculated by eq. 2.1.9, integrated over the unit cell.

$$\rho_{hkl} = \frac{1}{V_{cell}} \int_{cell} \rho(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} d\mathbf{r} \tag{2.1.9}$$

To solve this equation, it is convenient to separte the $\mathbf{r}$ vector in $\mathbf{r} = \mathbf{r}_n + \mathbf{r}_\alpha + \mathbf{r'}$, as it is established in figure 2.4.

Therefore, the eq. 2.1.9 leads to eq. 2.1.10. Given the scattering density symmetry the integral we find the atomic scattering factor in function of $sin\theta/\lambda$, equation 2.1.11, where 2θ is the angle between $\mathbf{k}$ and $\mathbf{k}_0$.

$$\rho_{hkl} = \frac{1}{V_{cell}} \sum_{\alpha} e^{-i\mathbf{G}\cdot\mathbf{r}_\alpha} \int \rho_\alpha(\mathbf{r'}) e^{-i\mathbf{G}\cdot\mathbf{r'}} d\mathbf{r'} \tag{2.1.10}$$

$$f_\alpha = 4\pi \int \rho_\alpha(r') r'^2 \frac{sin[4\pi r'(sin\theta/\lambda)]}{4\pi r'(sin\theta/\lambda} dr' \tag{2.1.11}$$

Finally, the summation over α in eq. 2.1.10 defines the so called structure factor F_{hkl} (eq.

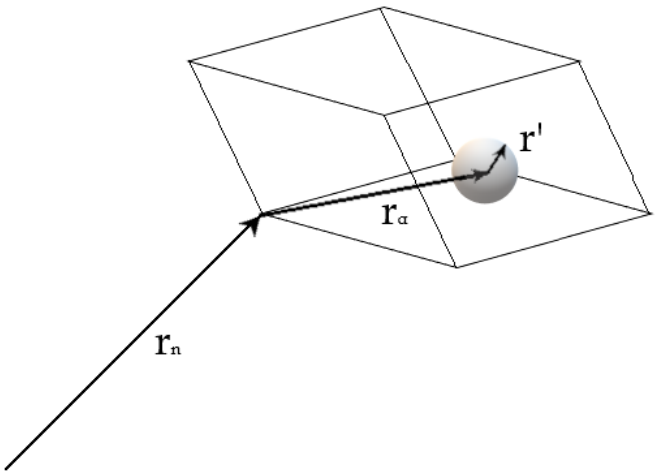


Figure 2.4: Definition of position vectors **r**, where $\mathbf{r}_n$ indicates the unit cell origin, $\mathbf{r}_\alpha$ points to the center of the atoms inside the unit cell, and **r'** to a point within the atom.

2.1.12).

$$F_{hkl} = \sum_{\alpha} f_\alpha e^{-i\mathbf{G}_{hkl}\cdot\mathbf{r}_\alpha} \tag{2.1.12}$$

Obtaining F_{hkl}, amplitude and phase, gives directly the electronic density in the unit cell, as it is the inverse Fourier transform of the structure factors:

$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h,k,l=-\infty}^{+\infty} F_{hkl} e^{2\pi i \mathbf{G}_{hkl}\cdot\mathbf{r}} \tag{2.1.13}$$

From an X-ray crystallography experiment, integrated intensities I_{hkl} of reflections, measured by detectors, are given by the Darwin's formula (eq. 2.1.14) [7].

$$I_{hkl} = I_0 r_e^2 \frac{V_{crystal}}{{V_{cell}}^2} \frac{\lambda^3}{\omega} LPA|F_{hkl}|^2, \tag{2.1.14}$$

where I_0 is the incident intensity, λ the X ray wavelength, ω is the crystal angular velocity during exposure, L, P and A are the Lorentz, polarization and transmission factors, respectively, r_e is the classic electron radius, and F_{hkl} is the structure factor.

Equation 2.1.14 states the quadratic relation between I_{hkl} and F_{hkl}. Hence, from the experiment we might obtain F_{hkl} amplitudes, but the phase information is lost. It gives a rise to the phase problem.

2.1.3 The phase problem

Since X-ray crystallography experiments do not use lens, the only quantity measured to determine the structure factor F_{hkl}, an imaginary number, are the amplitudes. In order to overcome

the phase problem, there are some strategies that enables the electronic density model calculation of a molecular structure.

In small molecules, less than 1000 atoms per molecule, it is used some chemical constraints to obtain the phases from the difference between distinct Fourier components[8]. Those are called direct methods and it is also exploited on macromolecules.

Another constriction that gives useful information phases is the Patterson function, that is analyzed by Patterson maps. The Patterson function is a Fourier transform of the measured intensities (eq. 2.1.15), with phases set to zero[9].

$$P(\mathbf{u}) = \mathcal{F}[|F(\mathbf{h})|^2] = \mathcal{F}[F(\mathbf{h})F(\mathbf{\text{-}h})] = \rho(\mathbf{r}) \times \rho(-\mathbf{r}) = \int_V \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})d\mathbf{r} \tag{2.1.15}$$

where $\mathbf{h}$= $\mathbf{G}_{hkl}$ is the reciprocal lattice vector[9].

The Patterson map P(u) will give a notion about the atomic distribution, as eq. 2.1.15 has larger values when u equals the interatomic distance. If there are N atoms in the unit cell, the Patterson map will have N(N-1) peaks, and their intensities are proportional to the number of electrons of that atom.

In addition, the Friedel pairs law is well know on the calculation of the electron density model. It assumes the density ρ(r) as approximately real valued, so the structure factor of centrosymmetric reflections $\mathbf{h}$ and $\mathbf{\text{-}h}$ are complex conjugates eq. 2.1.16. Therefore, their squared amplitude will be equal (eq. 2.1.17). In other words, the so called Friedel pairs hkl and $\overline{hkl}$, will have the same intensity.

$$F(h) = \mathcal{F}[\rho(r)] = \int_V \rho(\mathbf{r})e^{i\mathbf{h}\cdot\mathbf{r}}d\mathbf{r} \Rightarrow F(\mathbf{h}) = F^*(-\mathbf{h}) \tag{2.1.16}$$

$$|F(\mathbf{h})|^2 = |F(-\mathbf{h})|^2 \tag{2.1.17}$$

In a particular regime, incident light on absorption edges of a certain molecule element, it will occur a resonance and the scattering will be anomalous. Therefore, the scattering factor amplitude gains a frequency-dependent factor and turns into a complex number (eq. 2.1.18)

$$f_\alpha = f_0 + \Delta f' + if'' \tag{2.1.18}$$

where f_0 is the non-anomalous scattering factor, $\Delta f'$ and f'' are the real and imaginary dispersion corrections, respectively. Given the imaginary part f'', in this regime, the Friedel's law is, generally, not valid.

2.2 Molecular structure determination

After having the intensity reflection lists there are a few methods in crystallography know to overcome the phase problem described in the last section. Here, we are going to shortly describe the most used ones for macro molecules.

2.2.1 Molecular replacement

Molecular replacement bases on the substitution of an already solved molecular structure that has a similarity (>30%) with the molecule you want to solve. It has an input of the amino acid sequence and the electronic density. The program will rotate and translate the known structure in the unit cell, until the best fitted structure minimizes the difference between the calculated diffraction data from the replaced model and the reflections observed in the experiment.

2.2.2 Direct methods

Direct methods are also called, *de-novo* or experimental phasing. They are more complex and usually requires time and laboratory specifics. The most popular ones are the isomorphous replacement and anomalous diffraction.

The isomorphous replamecement comprehends the single isomorphous replacement (SIR) and multiple isomorphous replacement (MIR). It relies on the obtainment of a derivative, where the investigated structure should have one or more heavy atoms into its structure. Derivative and native crystals have to be isomorphous, i.e. the molecular structure and lattice should be equal, except for the heavy atom density. From both diffraction intensities, you can locate the heavy atom, defining the heavy atom substructure phase. Therefore, you can calculate structure factor phases of the native and derivative crystals[9].

The anomalous might be single anomalous diffraction (SAD) or multiple anomalous diffraction (MAD) , as well. It is based on the effect of anomalous scattering, described in last section. The phases are obtained from the difference in the measured intensities of Friedel pairs of reflection (SAD) or the same difference at two different X-ray wavelengths (MAD). Usually, heavier atoms, naturally present on proteins or by soaking, are used to obtain a significant anomalous scattering.

2.2.3 Structure refinement and validation

After having your model, there are a few programs that automatically refines your model, posing chemical restraints and some prior knowledge, to improve your data quality. The data quality are measured by the figures of merit. Here we describe the most popular ones for crystallographic data validation.

- Completeness: Percentage of the theoretical number of reciprocal space reflections;
- Multiplicity or redundancy: The average number of observations for a given reflection hkl;
- Signal to noise ratio (SNR): The ratio between the intensity of Bragg peaks to their average standard error, respectively.

$$SNR = \frac{< I >}{< \sigma(I) >} \tag{2.2.1}$$

- Pearson correlation coefficient (CC): obtained when the same set of observations are made twice (x_j,y_j) [10].

$$CC_{=} \frac{\sum_j (x_j - \bar{x})(y_i - \bar{y})}{\left[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2\right]^{1/2}} \tag{2.2.2}$$

- $CC_{1/2}$: Intra-data-set correlation coefficient, is calculated based on CC, with unique reflections randomly assigned to two half data sets.

$$CC_{1/2} = \frac{\sum(a_i - \bar{a})(b_i - \bar{b})}{\left[\sum(a_i - \bar{a})^2 \sum(b_i - \bar{b})^2\right]^{1/2}} \tag{2.2.3}$$

  where a_i and b_i are the intensities of unique reflections merged across the observations randomly assigned to subdatasets A and B, respectively, and $\bar{a}$ and $\bar{b}$ are their averages [11].

- CC*: the closest estimation for CC_{true}, which is the correlation between the arithmetic average of the half data set intensities I_1 and I_2 and the true intensities J. CC* provides a statistic that not only assesses data quality, but also allows direct comparison of crystallographic model quality and data quality on the same scale [11].

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}} \tag{2.2.4}$$

- R_{meas}: It is a data quality indicator that compares intensities or amplitudes of unique reflection with the average intensity of Friedel pairs of a unique reflection. The $\sqrt{\frac{n}{n-1}}$ factor is a redundancy independence correction of the know R_{sym} or R_{merge}, that should not be used to evaluate data quality anymore [12].

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{j=1}^{n} |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}} \tag{2.2.5}$$

  where the sums over hkl extend only over unique reflections with more than one observation.

- R_{split}: Also called R_{mrgdI}, it has been most used by SFX community, and it is directly related to $R_{p.i.m}$ by approximately an $\sqrt{2}$ factor[12]. $R_{p.i.m}$ gives a measuring of the averaged intensities precision, since it includes precision improvement according to the increase in multiplicity with the $1/\sqrt{(n)}$ factor[11].

$$R_{split} = R_{mrgdI} \approx \sqrt{2} R_{p.i.m.} = \sqrt{2} \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{j=1}^{n} |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}} \tag{2.2.6}$$

- R_{factor}: It is a model quality indicator used in the refinement process to compare the observed structure factor with the model obtained from data. The objective is to minimize the rate R_{work}/R_{free} during refinement cycles, where R_{free} is a data portion that will not be included in the refinement process as a reference.

$$R = \frac{\sum_{hkl} ||F_{obs}| - |F_{calc}||}{\sum_{hkl} |F_{obs}|} \tag{2.2.7}$$

Chapter 3

Experimental methods in X-ray crystallography

In this chapter, we explain the principles of crystallography experimental methods. The beginning of crystallography is dated after the accidental discovery of X-rays by Wilhelm Roentgen, in 1895. Since then, X-rays beams have improved in many of their features (brilliance, coherence, emittance and others). In the first section, we explain the X-ray sources available nowadays and their main concepts. After that, we describe crystallography experimental methods that have been developed for sample delivery, focusing on the most recent technique of Serial Crystallography (SX), which is the principal study object of this dissertation.

3.1 X-ray sources

Historically, X-ray first tubes consisted of two metal electrodes, where it was applied a high voltage in the gas bulb. The ionized gas creates free electrons, which starts a chain reaction. The positive charged atoms are attracted to the cathode, that knock the surface electrons. Applying a high voltage, they hit the anode and tube walls in a high velocity. Therefore, these electrons excite atoms to higher energy levels generating a broad energy spectrum in their decay, with X-ray straight lines due to characteristic elements of matter.

Nowadays, X-ray tubes use thermionic emission to produce electrons from a heated cathode. The anode can be rotated for better cooling (rotating anodes) and can achieve up to 10^{10}ph/s·mm^2 [9].

3.1.1 Synchrotron light sources

Synchrotron radiation sources began as an spurious effect of accelerator facilities for particle physics. The light emitted by charged particles moved in relativistic regime, when submitted to a radial acceleration, emit a focused beam at the tangent. The energy spectrum is broad, from infrared to hard X-rays, which is suitable for the study of a wide range of materials. In 1970s, the second generation facilities were specifically build for synchrotron experiments. They had bending magnets as main source of radiation. The skip for the third generation came with the need for

higher flux at sample and higher energies. For that, insertion devices (wiggler and undulators) became the source of radiation, and as they do not disturb the electron orbit they can use much higher magnetic fields.

The most recent developments at accelerators is the advance of fourth generation synchrotrons. The crucial innovation of the fourth-generation machines is to employ a narrower vacuum pipe to circulate the electrons in, allowing stronger magnetic fields to be used and more compact bending and focusing magnets[13]. The next generation of synchrotrons are a great promise for submicrometer beam at sample, high flux (around 10^{12}) and low emittance (<0.25 nm·rad).

3.1.2 X-ray Free Electron Lasers (XFELs)

X-ray Free Electrons Lasers (XFELs) have grown to supply the need for brighter and coherent sources. It uses the process of microbunching and Self-Amplified Spontaneous Emission (SASE), to compress the electron bunchs and self-amplify the emitted radiation. Electrons are emitted from an electron gun, accelerated to relativistic regime and submitted to long undulators, where the X-rays are produced. FELs have achieved another level of peak brilliance in light sources, and gave rise to many novel techniques, one of them is the serial crystallography, main subject of this dissertation.

3.2 Data collection techniques

3.2.1 Laue crystallography

Historically, Laue crystallography was the first crystallography technique, due to broad peak X-ray sources available at that time. It is based on the acquisition of diffraction patterns from the same crystal with more than one X-ray wavelength. There one will have two or more Ewald sphere radius proving crystal's reciprocal lattice. Recently, the technique is coming back with the advance of pink beams at advanced light sources. Specially for serial crystallography it has been demonstrated to be a faster way to obtain a complete dataset[14].

3.2.2 Single-crystal oscillation

Single-crystal oscillation is the current most popular method of crystallography. It became the standard data-collection mode after the development of methods that avoided crystals dehydration, enabling them to receive a higher radiation dose. It is based on monochromatic incident radiation, i.e one Ewald sphere. The crystal is rotated in many angles to have maximum proof of the reciprocal lattice. With the auxiliary of cryogenic techniques, and starting from multiple orientations, a single crystal can be enough to solve a target structure.

3.2.3 Serial Femtosecond Crystallography (SFX)

The most recent technique of serial crystallography emerged with the rising of new extremely brilliant sources, initially with XFELs that have femtosecond pulses of high intensity. It imposed

a great challenge, especially for biological samples, due to radiation damage. XFELs pulses were capable to evaporate macromolecular crystals, even at cryogenic temperatures, as soon as the first pulse hits them. The breakthrough idea was to give up on single crystal measurements, as they couldn't hold on a complete data collection for multiple crystals. They noticed that even if you destroy your crystal, at least one or a few diffraction patterns could be achieved with the high brilliance of the source, that was named *diffraction-before-destruction* [15]. The issue is that you have to measure many multiple crystals (tens of thousands patterns) in order to achieve a complete dataset. Furthermore, they should be randomly oriented to prove the entire reciprocal space and be able to solve a structure. Several advances in sample-delivery, sample-holders, data processing have been made since the beginning of Serial Femtosecond Crystallography (SFX) . The actual challenge transposed to data processing is that you have thousands of diffraction patterns, not ordered, to be merged as it has come from a single crystal. For that, many serial crystallography software have been developed, mostly to handle XFELs data.

3.2.4 Serial Synchrotron Crystallography (SSX)

With the rise of the first microfocus beamlines and high brilliant source, with fourth generation synchrotrons, a great interest of bringing the serial technique developments to synchrotron facilities have emerged [8] [16] [17] [18]. The picoseconds pulse of synchrotons ideally covers a large range of macromolecular systems where the biological interest is predominantly in the slower dynamics (μs–s), that produce well diffracting microcrystals [19]. SSX can also enable to solve difficult to crystallize samples, as membrane proteins, which are quite unknown until now. Giving up on freezing crystals enables room temperature experiments, as time-resolved and protein folding dynamics intermediates revealing.

The two main sample delivery methods applied to synchrotron techniques, that is planned to be covered on Manacá beamline, are: fixed-target and flow-focusing method (figure 3.1). The fixed target method is based on the rastering of crystals disposed on chips, mesh-grids or nylon loops. Flow-focus comprehends by a jet sample delivery in front of the beam. Their main two setups are the gas dynamic virtual nozzles (GVDN) [20] [21] and lipidic cubic phase (LCP) jet [22] [23] [24].

The GVDN creates a micrometer-sized liquid jet containing protein crystals. It has a high sample consumption, but a low background scattering. The LCP consists on a viscous medium, which contains the protein crystals, can reduce sample consumption. It is of great interest to hydrophobic molecules target, such as membrane proteins, that are better crystallized on greasy solutions.

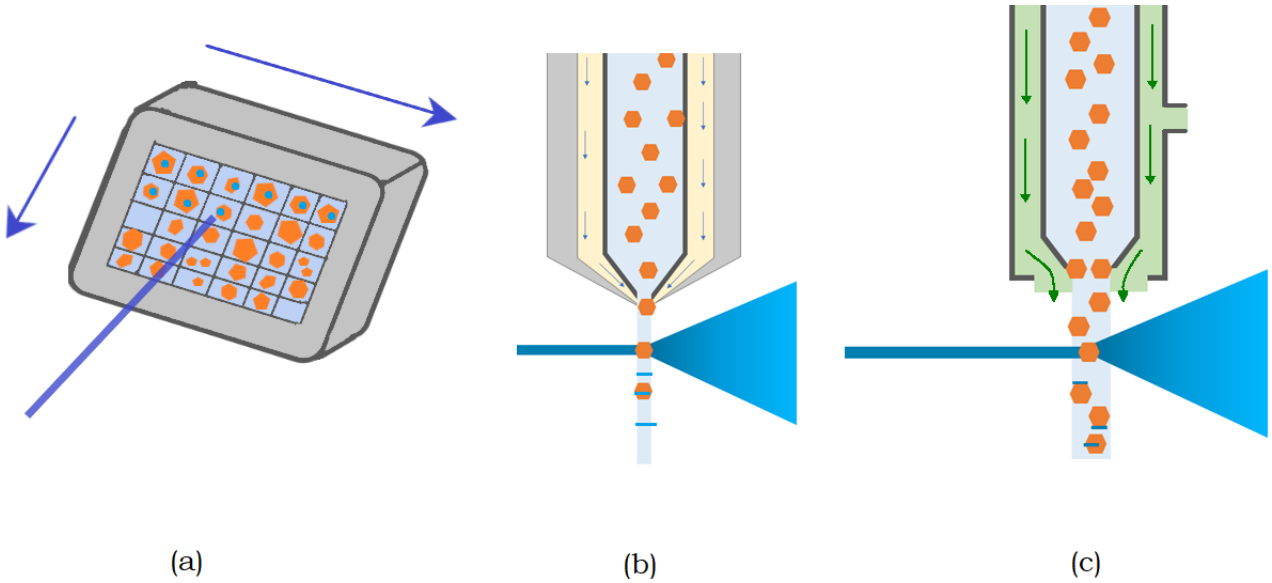


Figure 3.1: Sample delivery methods for serial crystallography experiments: (a) fixed-target (b) gas dynamic virtual nozzles (c) lipidic cubic phase jet

Chapter 4

Manacá beamline (Sirius - LNLS)

Manacá is the first beamline in operation for users at Sirius. It is one of the longest beamlines and is in the first phase of Sirius project. In this chapter, we describe the recent developments at Sirius and its current status. In the last section, we deepen the details on the Manacá beamline, with its main features already available and future ideas.

4.1 Sirius project

Sirius is one of the most complex scientific infra-structures constructed in Brazil. It is a fourth generation synchrotron, and is projected to be one of the most brilliant of its energy range (3 GeV, 0.25 nm·rad of emittance). The storage ring, 518 m of circumference, is designed to support 38 experimental stations, housing 20 magnetic cells with Five Bend Acromat (5BA) lattice. Sirius is planned to cover a wide energy range, including Hard X-rays (until 120 keV).

Sirius is one of the Brazilian Center for Research in Energy and Materials (CNPEM) installations, in Campinas, São Paulo. CNPEM holds four national laboratories, with open facilities for scientists of many fields. They are the Brazilian Biotechnology (LNBio), Nanotechnology (LNNano), Biorenewables (LNBR), Synchrotron Light Source (LNLS) National Laboratories. LNLS operates the only synchrotron light sources in South America, until now. The first machine, UVX (fig. 4.1), a second generation light source started to operate, in 1997, and it attended the scientific community until its shutdown, in 2019, with the end of Sirius' civil constructions.

(a) (b)

Figure 4.1: Brazilian Synchrotron Light Source installations (a) Sirius[25], a fourth generation source (3 GeV, 518 m of circumference) (b) UVX[26], second generation machine (1.37 GeV, 93.2 m of circumference.

Sirius current phase plans to provide 14 beamlines, covering a wide range of materials and techniques. Their names are acronyms related to Brazilian fauna and flora species. The first beamlines and their main aspects are summarized in figure 4.2 and table 4.1

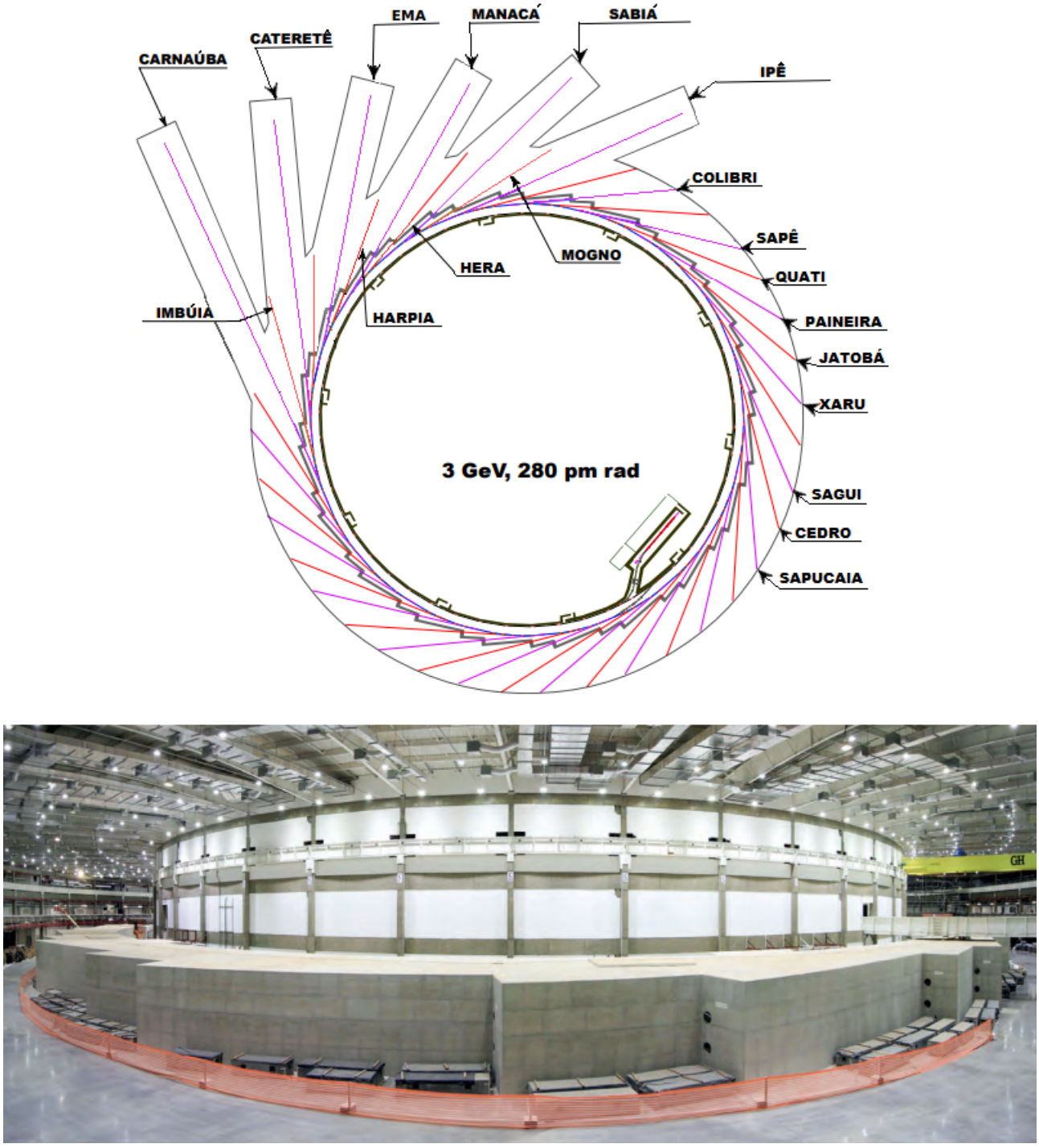


Figure 4.2: Sirius designed beamlines and panoramic vision of experimental hall before beginning of optical and experimental hutches construction [26].

Beamlines	Main technique	Energy range	Source
Carnaúba	X-Ray Nanoscopy	2.05 - 15 keV	ID
Imbúia	Infrared Micro and Nanospectroscopy	550 - 3500 cm^{-1}	BM
Cateretê	Coherent and Time-resolsed X-ray Scattering	5 - 20 keV	ID
Ema	X-ray Spectroscopy e Diffraction in Extreme Conditions	2.7 - 30 keV	ID
Manacá	Macromolecular Micro and Nanocrystallography	5 - 20 keV	ID
Sabiá	Soft X-Ray Absorption Spectroscopy and Imaging	100 - 2000 eV	ID
Mogno	X-ray Micro- and Nanotomography	22 | 39 | 67.5 keV	BM
Ipê	Resonant Inelastic X-ray scattering and Photoelectron spectroscopy	100 - 2000 eV	ID
Sapê	Angle-Resolved PhotoEmission Spectroscopy	8 - 70 eV	BM
Quati	X-ray Spectroscopy with Temporal Resolution	4.5 - 35 keV	BM
Paineira	Powder X-ray Diffraction	5 - 30 keV	ID
Jatobá	Full X-ray Scattering and PDF Analysis	40 - 70 keV	BM
Cedro	Circular Dichroism	3 - 9 eV	BM
Sapucaia	Small Angle X-ray Scattering	6 - 17 keV	ID

Table 4.1: Sirius beamlines in assembly or commisioning, currently, and their mains aspects (https://www.lnls.cnpem.br/beamlines/).

4.2 Manacá beamline

Manacá-de-cheiro is a tree of Solanaceae family, typically founded in the Brazilian Atlantic Forest. As a tribute to this native tree, Manacá (MAcromolecular micro and NAno CrystAllography) is an acronym for techniques that will be available for its users. The Manacá beamline will cover a range of crystallography techniques applied to macromolecular and small molecule samples. Additionally, taking advantage of the micro and nanofocus of Sirius, it will enable us to implement the recent technique of serial crystallography (SX). The initial design (figure 4.3) aims two experimental stations in-line, that may enable two experiments being run or prepared, in parallel.

	MicroManacá	**NanoManacá**
Beam at sample (μm^2)	10x6 to 100x80	0.7x0.5
Energy range (keV)	6 - 20	6-20
Energy resolution	10^{-4}	10^{-4}
Divergence at sample (mrad)	0.5	0.5

Table 4.2: MicroManacá, first experimental hutch of Manacá, and NanoManacá, second hutch, key specifications.

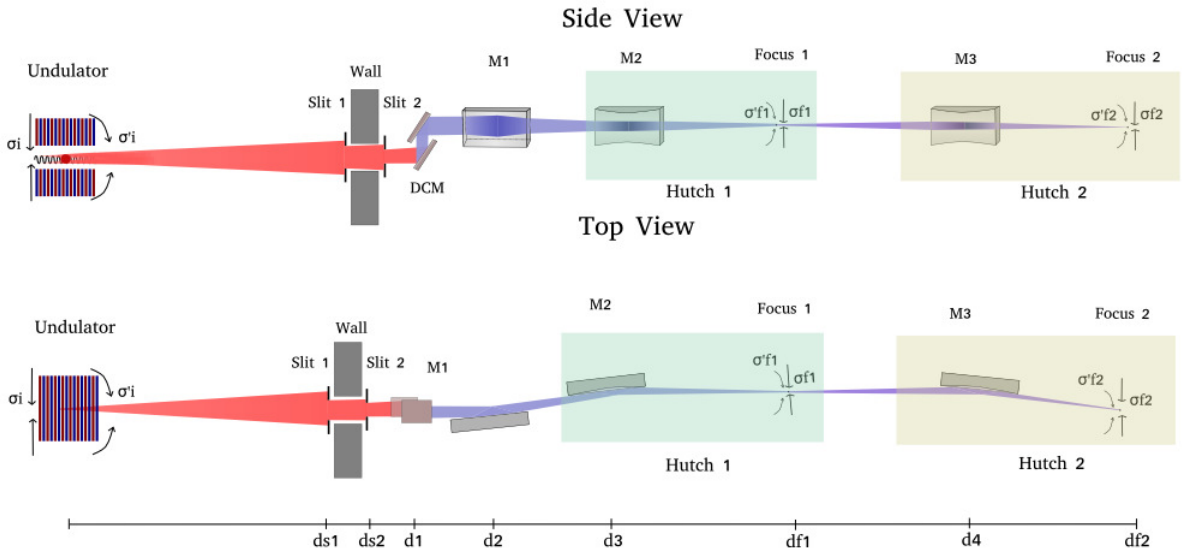


Figure 4.3: Schematic of Manacá beamline main components.

It is an undulator beamline, with double crystal monochromator (DCM), internally developed by Sirius support groups [27]. The optics consists on a pair of Kirkpatrick–Baez (KB) mirrors, the first (vertical focusing) stays in the Optical hutch and the second (horizontal focusing) in the MicroManacá, first experimental station. There we have microfocus at the samples, and it is already open for users. In the future we will have another mirror, that will refocus the beam vertically, achieving the nanofocus at NanoManacá station. Their main characteristics are summarized in table 4.2.

The first data collection at Manacá was performed in July 2020, with the beginning of scientific commissioning. It first opened to COVID-19 related subjects, and nowadays it is receiving proposals for all users. In MicroManacá hutch, we have installed the automatic sample changer (Stäubli robotic arm), already in performance for cryogenic samples (figure 4.4). Currently, we are performing the robot first tests for remote data collection with MXCuBE.

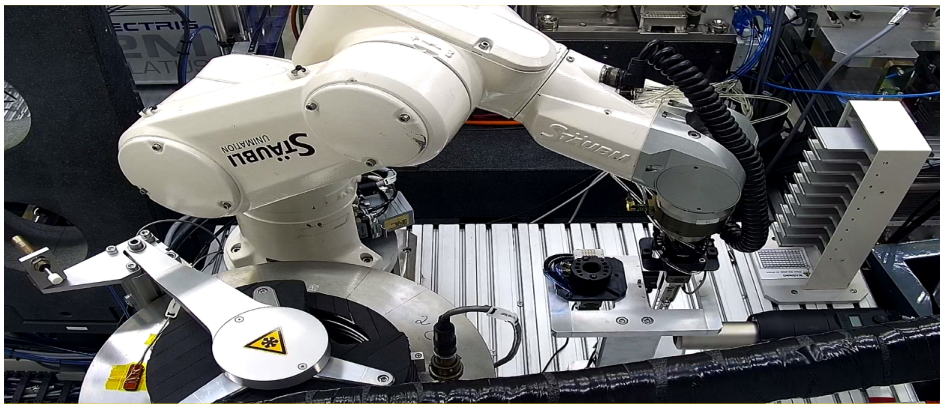


Figure 4.4: Automatic sample changer in operation at MicroManacá for cryogenic samples.

The beamline is equipped with a Pilatus 2M (Dectris) detector, positioned in a granite table, that can smoothly move back and forth by pneumatic system. It is also available an on-axis video microscope for sample alignment, airbearing goniometer, cryojet and fluorescence detector (figure 4.5).

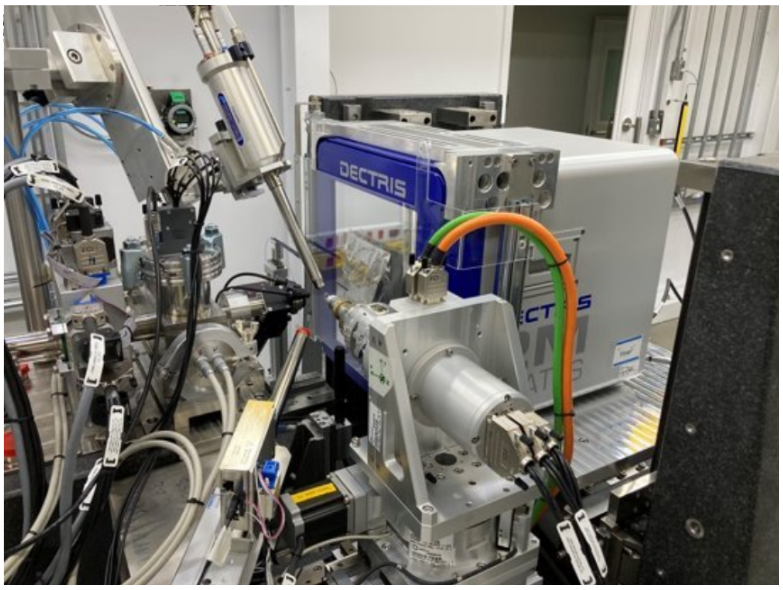


Figure 4.5: Current status of MicroManacá facility: Pilatus 2M (Dectris) detector, cryojet, on-axis video microscope, airbearing goniometer, beam stopper, slits and beam positioning monitors (BPM).

Chapter 5

Serial crystallography data processing

Serial crystallography data processing takes several steps from the oscillation method data processing, hence all conventional crystallography packages (*XDS*, *CCP4*, *cctbx* and others) are used at some point of the data processing. The main idea comes from the multiple crystals experiments. Firstly, occurs the individual analysis of all data coming from the same crystal, if this information is known, as in needle crystals multiple datasets or small oscillation experiments. After that, proceed data selection aiming the most isomorphic subdataset, in order to achieve the best data quality. Finally, user can continue with the traditional merging and, optionally, scaling and post refinement. With that, one should have all files needed to solve and refine the target structure.

In this chapter, we describe some data selection strategies that have been developed by the SX community, and the most used packages currently available. Afterwards, we show our pipeline, written in Python, that communicates with these SX packages, mostly *CrystFEL*[1] and *ccCluster*[2].

5.1 Data selection methods

The three main ideas for SX data selection are: Hierarchical Clustering Analysis (HCA), Down-weighting of outliers (snapshot images routine), and Genetic Algorithms (GA). Each one has its pros and cons, depending on the dataset collected.

The HCA needs some partiality of the subdatasets collected, as they will solve beforehand small oscillations with conventional crystallography packages. After that, they use machine learning algorithms to agglomerate isomorphic subdataset in clusters. The clusters are represented by a dendrogram where the x axis is the subdataset identification number and the y axis is the distance between them. The metric distance set in the algorithm can be unit-cell variation, intensity correlation coefficient (cc) or many others. In the same field, there is a study [28] that used another machine learning algorithm, the K-means clustering, that has also been proven powerful for data segmentation. In this field, these are the main packages that have been developed for SX: *ccCluster*[2] (ESRF - Grenoble, France), *BLEND*[3] (*CCP4*), *xscale_isocluster* (*XDS*), *xds_nonisomorphism* (*XDS*).

Downweighting or rejection of outliers are the most used strategy for snapshot images experiments, as subdatasets have extremely low partiality. Here, the images are treated individually, since they can't be solved as small oscillation. The main idea is to find all Bragg peaks in every pattern, index each image individually and fit the distribution of unit-cell parameters to find the average for these crystals to be used as a reference. After that, it should be set a tolerance in the unit-cell for the indexing methods and try to index the whole dataset. Then, crystals' reflections list is integrated and merged in a final dataset. The current packages in development for SX are: *CrystFEL* (DESY - Hamburg, Germany)[1], *cctbx.xfel* (LCLS, USA), *nXDS* (*XDS*).

Genetic algorithms (GAs) are a well known algorithm in biology and it can be used for automatic pipelines automation. It uses evolution and natural selection concepts to maximize or minimize a target function [29]. Applied to SX, the subdatasets are the genes, and the final dataset is the individual (chromosomes). The subdatasets are randomly separated into groups and rated by weights set in the GA. The algorithm applies selection, crossover and random mutations in the chromosomes to maximize the group rate, until the user gets an acceptable final dataset. GAs are straightforward way to automatize data selection, but it is much expensive computationally, compared to the other methods. In the last years it has not been widely applied in SX, as it relies on huge datasets.

5.2 *ccCluster* routine

ccCluster[2] has been the most popular software used for HCA. It was developed in the context of fixed-target sample delivery with small oscillations. It is comparatively the less expensive data selection method, but it depends on conventional crystallography packages to solve small wedges of diffraction patterns. *ccCluster* flow is briefly explained in figure 5.1.

5.2.1 *ccCalc.py*

ccCluster is mainly written in Python. *ccCalc.py* is the first script in the process. It uses a list of reflections list (`*.HKL` or `*.mtz`) as an input and calculates the correlation between each subdataset, constructing the cc distance matrix. There are two metrics currently available to calculate the distance between datasets: intensity-based correlation coefficient ('cc'), eq. 5.2.1 and unit-cell variation ('cell'), eq. 5.2.2. The first part of our automatic pipeline users indicate a list of paths, the output directory, the distance chosen ('cc' or 'cell'). Interval of heights in the dendrogram, commonly called threshold, should also be passed to the pipeline, and the number of points to be analyzed. The clusters merged will be equally spaced in that interval.

$$d(a,b) = (1 - cc^2_{(a,b)})^{1/2} \tag{5.2.1}$$

where $cc^2_{(a,b)}$ is directly obtained using a cctbx method (*miller_array.correlation.coefficient*), which calculates the correlation from common reflections in each pair of unmerged datasets (a,b).

$$d(a,b) = max\left[\left|\frac{A_a - A_b}{min(A_a, A_b)}\right|, \left|\frac{B_a - B_b}{min(B_a, B_b)}\right|, \left|\frac{C_a - C_b}{min(C_a, C_b)}\right|\right] \tag{5.2.2}$$

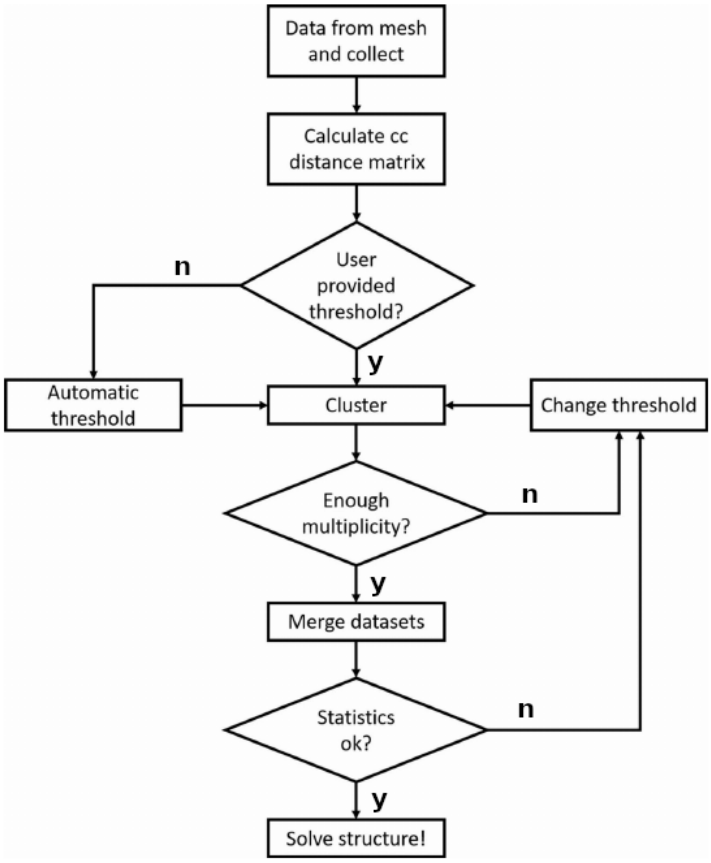


Figure 5.1: *ccCluster* flowchart diagram [2]

where A_i, B_i and C_i are the unit-cell lengths. The unit-cell method is highly sensitive to detector distance refinement, and it is less precise with wedges smaller than 10° rotation [2].

The distance between two clusters X and Y, is defined by the average linkage method (eq. 5.2.3) that will build a dendrogram, a tree representation of the clusters as in figure 5.2.

$$D(X,Y) = \frac{1}{N_X + N_Y} \sum d(a,b); a \in X, b \in Y \tag{5.2.3}$$

N_X and N_Y are the number of datasets in clusters X and Y.

5.2.2 *ccCluster.py*

The automatic pipeline calls the *ccCluster.py* script for each threshold passed. It will run `XSCALE`, merging the biggest cluster in that height. Our script summarizes the main control cards, given by `XSCALE` output and plot them comparatively.

ccCluster.py has a currently available option for automatic threshold estimation, that might give an idea of an acceptable clustering strategy. It computes the maximum variation of the number

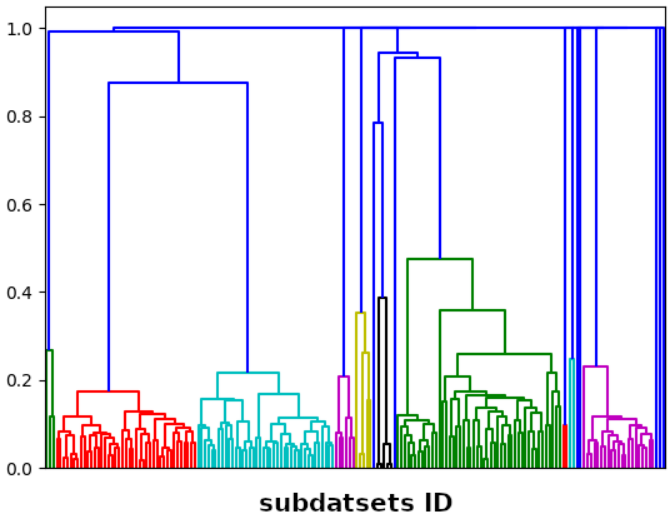


Figure 5.2: Dendrogram example for lysozyme and trypsin subdatasets, with 2° wedges, from crystals measured at LNLS on MX2 beamline.

of datasets in the biggest cluster in steps of threshold. The threshold with maximum variation will be given as the estimated threshold.

After deciding the merged dataset, the user can continue either with the `scaled.hkl` file created inside the threshold directory or to run a *ccCluster* internal script that runs POINTLESS and give the `clustered.mtz` file.

5.3 *CrystFEL* routine

CrystFEL [1] is the main software used for snapshot serial crystallography. It has been developed since 2015 in the context of X- ray free electron lasers (XFELs). It deals with low partiality dataset, where reflections are not entirely record, as it is on oscillation crystallography in sequential diffraction patterns. As the orientations are mostly random, the program treats each images individually. The package has six main steps:
1- Peak search, 2- Geometry file corrections, 3- Indexing, 4- Integration, 5- Merging, scaling and post refinement, 6- Figures of merit calculation.

The principal steps of *CrystFEL* are summarized in the software flowchart (figures 5.3 and 5.4).

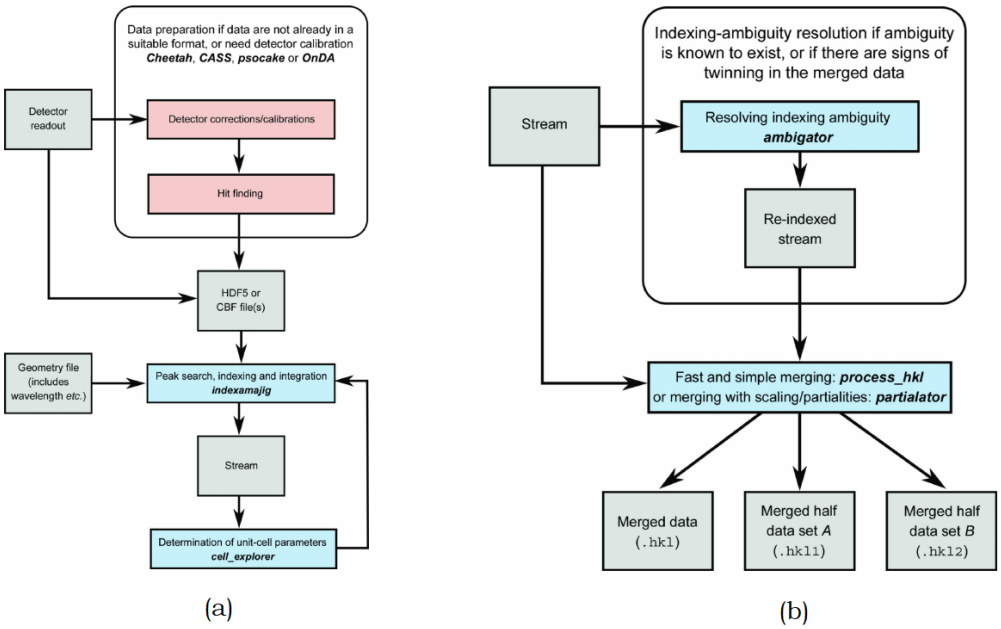


Figure 5.3: *CrystFEL* flowchart [30] (a) *indexamjig* and *cell_explorer* (b) merging, scaling and post refinement scripts.

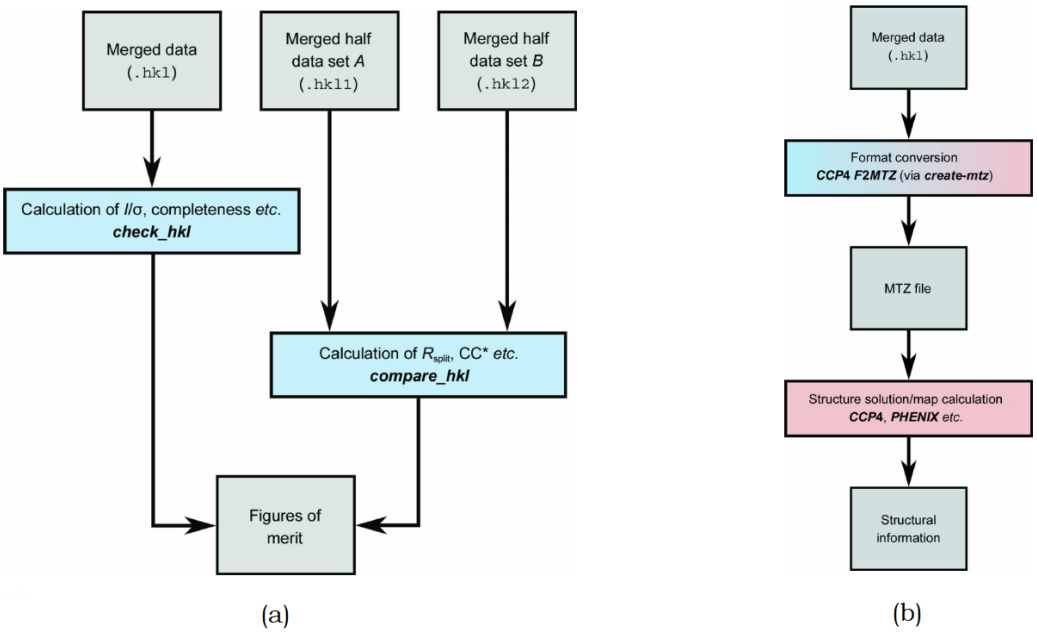


Figure 5.4: *CrystFEL* flowchart [30] (a) figures of merit calculation (b) conversion of the final data to MTZ.

5.3.1 Peak search

CrystFEL has two principal methods available nowadays for peak search: *zaef* and *peakfinder8*. Each one has its low-level parameters that should be adjusted for the user's dataset. The parameters are flexible to avoid noise included in peaks list. In table 5.1 is resumed the most useful parameters concerning SX at synchrotrons.

Method	Options	Low-level parameters	Deafault
zaef	–peaks=zaef	–threshold=thres	800
		–min-squared-gradient=grad	100000
		–min-snr=snr	5
		–peak-radius=inner,middle,outer	4,5,7
		–filter-noise	
		–median-filter=n	
peakfinder8	–peaks=peakfinder8	–threshold=thres	800
		–min-snr=snr	5
		–min-pix-count=n	2
		–max-pix-count=n	200
		–local-bg-radius=n	3
		–min-res=n	0
		–max-res=n	1200
peakfinder9	Not available in current version	–min-snr-biggest-pix=n	
		–min-snr-peak-pix=n	
		–min-sig=n	
		–min-peak-over-neighbour=n	

Table 5.1: Peak search methods in *CrystFEL* and its respective parameters.

The peak search parameters optimization usually follows the same order presented in table 5.1. We implemented an specific routine in Python that can make combinations of parameters, where users can easily adjust numbers that best fit their data (figure 5.5). The automatic pipeline will call `indexamajig` with the input parameters to be tested, with a command similar to:

```
indexamajig -i files.lst -g geometry file .geom -indexing=mosflm -o output.stream -j
number of processors --profile - -norefls-in-stream --peaks=method --options=values.
```

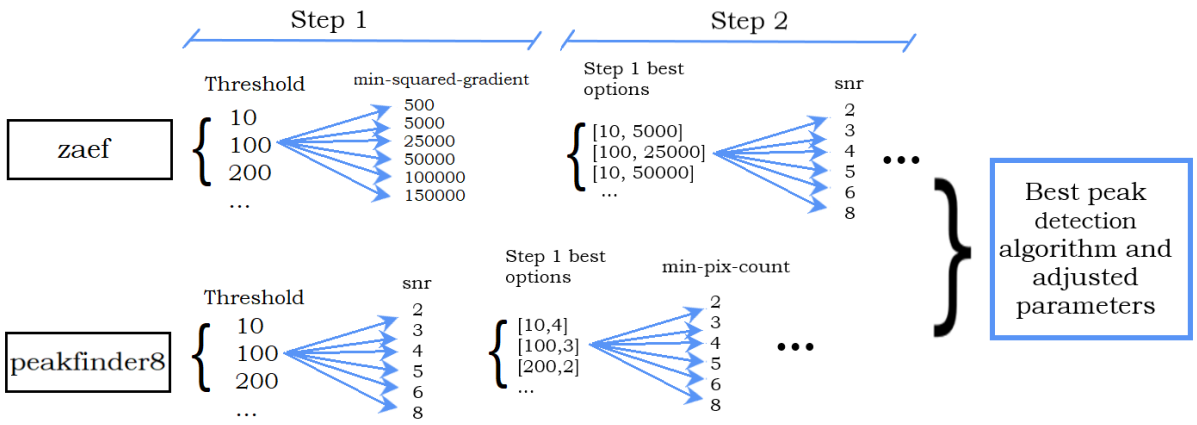


Figure 5.5: Schematic of the peak search optimization script written in Python. The users may test different algorithms available in *CrystFEL*, and adjust their parameters to their data.

The users can evaluate the indexing rate, mean peaks per pattern (MPP), mean ADU-intensity per peak (MAP) and total ADU-intensity per pattern (TAP), according to the parameters tested, from a comparatively plot done by the automatic pipeline (fig. 5.6). From those, the most important metric is to optimize the indexing rate.

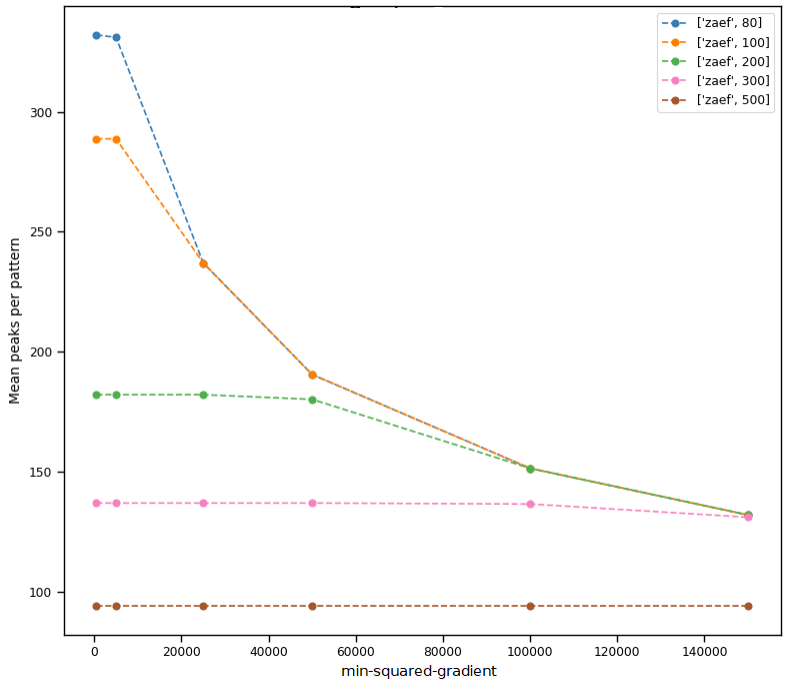


Figure 5.6: Mean peaks per pattern example for different threshold values and minimum squared gradient (step one of peak search optimization for *zaef*).

The objective of this stage is to find real Bragg peaks in all diffraction patterns, avoiding that noise might be included in the peaks list. The metrics listed (indexing rate, MPP, MAP, TAP) might be a clue for better indexing, but it always has to be accompanied by users evaluation. One might check the algorithms performance from the GUI of CrystFEL (in development), or calling a *CrystFEL* internal script called `check-peak-detection`.

The peak parameters optimization might be expensive depending on the number of tests desired, hence it is better to have an initial idea from looking at a part of the dataset and use the peak optimization to make fine adjustments. Also, it is recommended to use as an input a small, but representative, sample of a whole dataset in the initial stages (at least 500 images). Finally, as peak search optimization might call several times `indexamajig`, it is desirable to use the parallelization of *CrystFEL* (option -j number of processors) to use the maximum of processors available. At Sirius, the High Performance Computing (HPC) cluster could be suitable for these jobs.

5.3.2 Geometry file corrections

The geometry file of Pilatus 2M, current Manacá's detector, was based on *CrystFEL* examples and adapted to our detector specifications (appendix A). In order to have a better indexing performance, it is necessary to make some corrections in this file. The main corrections are: detector to sample distance and the beam centre position.

For distance to sample correction we build a function that alter the nominal camera length in the geometry file from an initial to a final point in steps passed by the user. The recommendation is to move at least 500 μm back and forth in steps of 100μm firstly. The unit-cell file here should be triclinic in order to avoid any bias in the detector correction. After seeing an improvement of the indexing rate, one takes a closer interval around this point, with steps of 20 μm.

The detector-to-sample distance has a great impact on the indexing methods performance, resulting on a better Gaussian distribution of the unit-cell parameters [31]. It is also a cumbersome stage, computationally, due to the expansion of input files to a few thousands images, at least, for good detector calibration.

The beam-shift position is also another important aspect to correct in the geometry file. It uses *CrystFEL* predicted spots and reflections position observed to estimate a beam shift for each images. The `detector-shift` script in *CrystFEL* applies directly the mean shifts, selected by user's clicking, in the geometry file changing the `corner_x` and `corner_y` nominal numbers. The detector shift correction was done with lattice type and centring prior information only, following *CrystFEL*'s tutorial.

The automatically pipeline calls `detector-shift` three times, correcting the beam position and reindexing with the new geometry file. It is plotted the shift map for each run, hence users can evaluate whether the correction runs is improving the data distribution or not. Additionally, one should have a look at the indexing rate evolution to decide the best beam shift correction.

5.3.3 Indexing

CrystFEL has some conventional crystallography indexing methods (MOSFLM[32], *Dirax* [33], *XDS*[34]), and some specific algorithms developed in the SX context (asdf[35], *XGANDALF*[36], *TakeTwo*[37] and *Felix*[38]). Their options and low-level parameters are summarized in table 5.2.

Method	Options	Low-level parameters
mosflm	mosflm-nolatt-nocell	
	mosflm-latt-nocell	
	mosflm-nolatt-cell	
	mosflm-latt-cell	
dirax	dirax-nolatt-nocell	
asdf	asdf-nolatt-nocell	
	asdf-nolatt-cell	
xds	xds-nolatt-nocell	
	xds-latt-cell	
xgandalf	xgandalf-nolatt-nocell	–xgandalf-sampling-pitch=n
	xgandalf-nolatt-cell	–xgandalf-grad-desc-iterations=n
		–xgandalf-tolerance=n
		–xgandalf-no-deviation-from-provided-cell
		–xgandalf-max-lattice-vector-length=n
		–xgandalf-min-lattice-vector-length=n
		–xgandalf-max-peaks=n
		–xgandalf-fast-execution
taketwo	taketwo-latt-cell	–taketwo-member-threshold=n
		–taketwo-len-tolerance=n
		–taketwo-angle-tolerance=n
		–taketwo-trace-tolerance=n
felix	Support not available	

Table 5.2: *CrystFEL* indexing methods and their respective low-level parameters.

The option *latt* corresponds to the lattice type prior information, that should be included in the unit-cell file (∗.*cell*). The same for the *cell* option, where it should added unit-cell parameters.

Initially, if there is no clue of target unit-cell, one runs `indexamajig` with a few images (500) and mosflm-nolatt-nocell option. From that, it could be visible a distribution around each unit parameter. From that, users can extract lattice-type and centring information that should be added to the (∗.*cell*) file. Afterwards, indexing runs again with the same number of images, but now with mosflm-latt-nocell option and fit the unit-cell parameters using the internal *CrystFEL* script *cell_explorer*.

Having unit-cell file in hands, it is possible to use all indexing methods of *CrystFEL*. The automatic pipeline implemented has the flexibility to call any method from table 5.2, except for *Felix* that isn't currently available on Manacá's installations. It can be used a combination of selected methods, and a random shuffled order of them. The straightforward process is to use all 12 methods available individually, or combinations of one, keeping the same order for similar datasets. At this stage, one should use a smaller, but representative, part of the whole dataset, e.g. a few thousands images from different points of the total data collection.

Finally, users are able to evaluate indexing options performance, individually, and select the most successful ones. The automatic pipeline plots them comparatively, according to their indexing rate performance, MAP and MPP. If it is desirable a quick answer, one should choose the indexing method with more crystals founded.

Otherwise, one can tune higher indexing rates using the best methods combined, so that if the first one couldn't find a crystal the program will skip to the next one. We opted for a descending order, according to the indexing rate, to optimize the indexing processing time. In this stage it should include the whole dataset, usually tens to hundreds thousands of images, depending on the experimental setup. It is very recommended to have a high performance machine in this step.

5.3.4 Integration

CrystFEL has two different options of integration (table 5.3). The automatic pipeline has a function with similar structure of the peak search optimization (section 5.3.1), where users might want to adjust integration parameters for better final data quality.

Method	Options	Deafult
rings	-cen	–integration=rings-nocen
	-sat	-nosat
	-grad	-nograd
	–int-radius	4,5,7
	–int-diag	none
	–fix-profile-radius, fix-divergence	auto
	–rescut	infinity
prof2d	-cen	–integration=prof2d-cen
	-sat	-nosat
	-grad	-nograd
	–int-radius	4,5,7
	–int-diag	none
	–fix-profile-radius,fix-divergence	auto
	–rescut	infinity

Table 5.3: *CrystFEL* integration methods and their respective options.

5.3.5 Merging, scaling and post refinement

The automatic pipeline has a function, `runmergin`, that calls *CrystFEL* merging scripts (table 5.4, `process_hkl` and `partialator`. User can choose between them and also turn on and off scaling, partialities or post refinement.

Merging	Scaling	Partialities	Post-refinement	Options
partialator	0	0	0	–model=unity –iterations=0
	1	0	0	–model=unity –iterations=1
	0	1	0	–model=xsphere –iterations=0
	1	1	0	–model=xsphere –iterations=1 –no-pr
	1	1	1	–model=xsphere –iterations=1
	0	1	1	–model=xsphere –iterations=1 –no-scale
process_hkl	1	0	0	–scale

Table 5.4: *CrystFEL* internal scripts for merging, scaling and post refinement.

The automatic pipeline runs these two methods: `partialator` with scaling, partialities and post-refinement, and `process_hkl` with scaling. The last one is simpler and faster than the other, which is convenient for quick answer. In this dissertation, all figures of merit were calculated based on the `partialator` with scaling, partialities and post-refinement reflections list (`*.hkl`).

For monochromatic synchrotron radiation there is a a model option in *partialator*, offset, that can be interest to explore the post refinement using *CrystFEL* internal scripts *plot-pr* and *plot-pr-contourmap*. At this stage users must pass the symmetry group according to *CrystFEL* symmetry chart (Annex C).

5.3.6 Figures of merit calculation

The automatic pipeline calls *CrystFEL* script `checl_hkl` to calculate Completeness and signal-to-noise ratio (SNR) over resolution shells for a final dataset. For R_{split}, R_{1f}, R_{1i}, R_2, CC, CC^*, CC_{ano}, CRD_{an}, R_{ano}, R_{ano}/R_{split} e d_{1sig} and d_{2sig} calculation, it calls `compare_hkl`. Our automatic pipeline selects CC, CC^* and R_{split}, and plot them over resolution shells using *matplotlib* Python library.

Finally, users are able to convert the final dataset using functions `convertmtz` or `convertxscale`, setting the unit-cell parameters and point group in the function and passing the chosen final `*.hkl` file from `partialator` or `process_hkl`.

There is still some additional script in CrystFEL that might be interesting for SX at synchrotron as: `render_hkl`, `partial_sim`, `ambigator` and `whirglig`. The `ambigator` is necessary to solve ambiguities in merohedral systems.

Chapter 6

Data analysis of SX first tests on Manacá

In this chapter, we describe the first tests of serial crystallography on Manacá (Sirius). In the first section, we simulate a serial experiment from thousands of patterns obtained from *Ame*GH128 enzyme crystals, collected in oscillation mode. Afterwards, we performed a grid-scan experiment on large lysozyme crystals, cryocooled and at room temperature. There, it is summarized the main data processing strategies used and the data quality of the final datasets.

6.1 Oscillation of multiple *Ame*GH128 cryocooled crystals

*Ame*GH128 is an enzyme from the Carbohydrate-Active Enzymes (CAZymes) family. The CAZymes are natural abundant and are key enzymes for plant cell-wall breaking. It has been of great interest due to their industrial application, as in the production of biofuels. The *Ame*GH128 crystals measured, are well known [39] and were provided from a collaboration group of the Brazilian Biorenewables National Laboratory (LNBR) in CNPEM, Dr. Mário Murakami.

6.1.1 Experimental setup

The experiment was done at the beginning of Manacá's scientific commissioning, in August of 2020. The crystals were kept in liquid nitrogen (77K), using a Cryojet. They were mounted on nylon loops, with a detector distance of 0.0858m from the sample. The beam size varied from 20 to 60 μm, depending on the crystal size (50 to 300 μm). The beam energy was fixed in 9.15 keV.

It was collected 18 complete datasets, e.g. 3600 images with 0.1 degree oscillation around goniometer axis, giving a total of 64800 images. In order to simulate an SX experiment we simulated a loss of information from subsequent patterns, shuffling the same number of random images of each crystal. With the complete datasets, we came up with seven subdatasets containing: 500, 2500, 5000, 10000, 25000, 50000, 64800 images in total.

6.1.2 *CrystFEL* data processing

Unit-cell parameters determination

We began the peak search with small tests that included the *zaef* method: `-threshold=100 and 10 -min-squared-gradient=5000 -min-snr=4 -peak-radius=4,5,7`, and *peakfinder8*: `-threshold = 100 and 10 -min-snr=3 -min-pix-count=2 -max-pix-count=20 -local-bg-radius=6`. The best indexing rate performance obtained was with –peaks=*peakfinder8* `-threshold=10 -min-snr=3 -min-pix-count=2 -max-pix-count=20 -local-bg-radius=6`. The peak search parameters should be optimized for each data collection, as it is implemented on the script. Nevertheless, the combination of peak search parameters might significantly increase the computational cost. One should always consider the processing time when balancing between a quick answer or achieving a higher indexing rate, i.e, number of indexed images and, consequently, data quality.

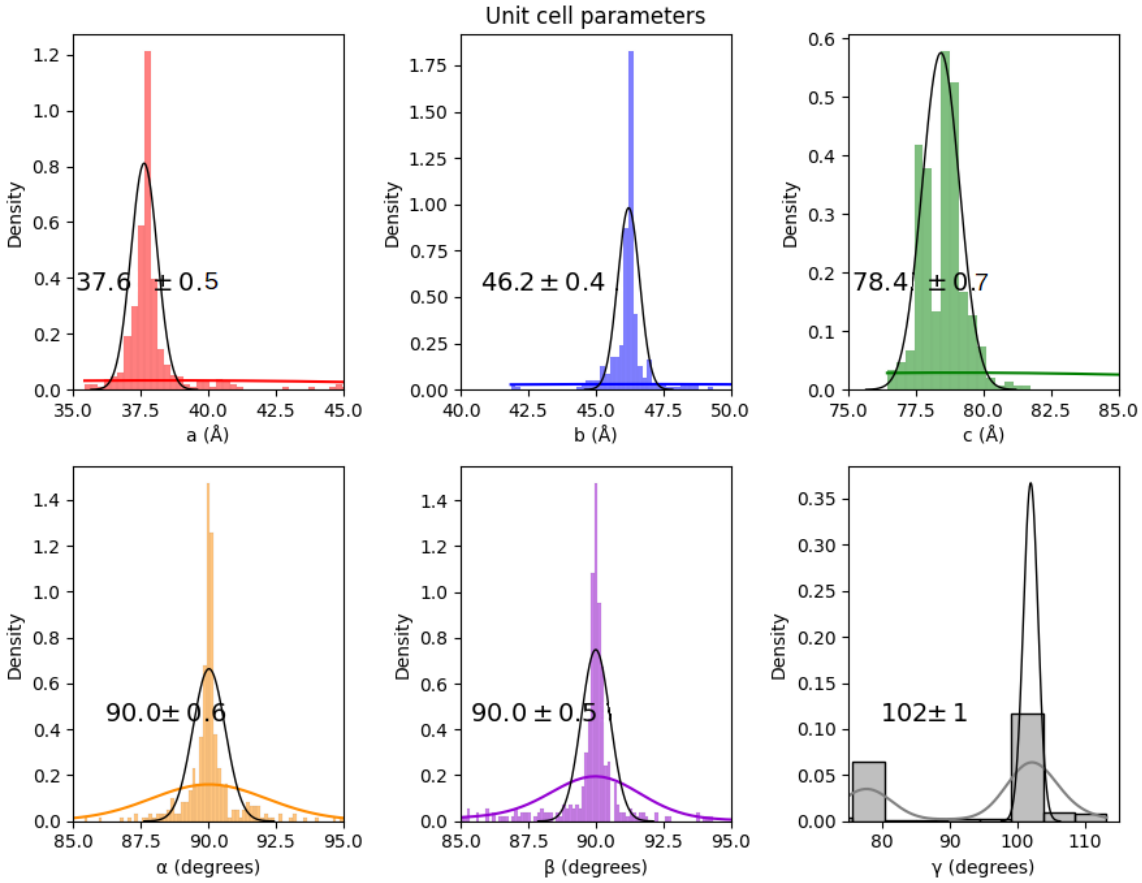


Figure 6.1: Unit-cell parameters distribution from 500 random images of *Ame*GH128, indexed with mosflm-latt-nocell and the simplest prior lattice information (triclinic lattice type, primitive centring): 500 images processed, 492 hits (98.4%), 454 indexable (92.3% of hits, 90.8% overall)

A single indexing process of a few thousand diffraction patterns, using mosflm and no prior information, can take around 20 minutes in a 28 processor machine. Combining m thresholds with n

signal to noise ratio implies n*m indexing process. Usually, in a single peak searching optimization step, it is great to observe at least 3 parameters combined with other 3, totalizing 3 hours per step, which for real time data processing is far from ideal. What we did here was to shorten the input files to 500 hundred images and test only low and high values, 10 and 100 in threshold for example, and look at indexing rate performance and unit-cell parameters distributions.

Initially, we used `mosflm-latt-nocell` for indexing with the simplest lattice type (triclinic and primitive unit-cell). That simplification considerably accelerates the indexing step, which is convenient for a first unit-cell determination. The unit-cell parameters found (figure 6.1) match the known numbers for *Ame*GH128 (a=38 Å b=79 Å c=46 Å α=90°, β=102°, γ=90°), except for an exchange between the b and c axes. If there are many discrepancies between known cell and the indexed, or if data processing is taking too long, one should always certify if the detector distance or beam energy are properly set in the geometry file.

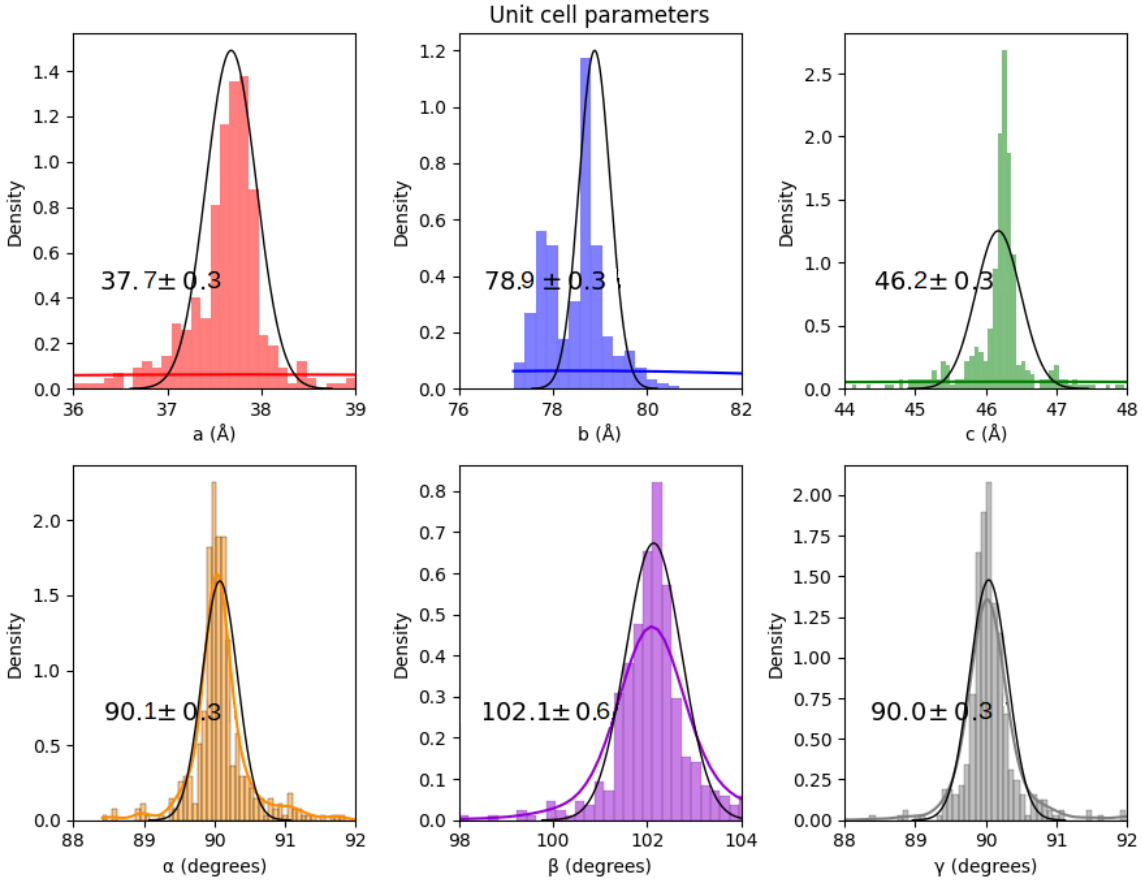


Figure 6.2: Unit-cell parameters distribution from 500 random images of *Ame*GH128 crystals, indexed with mosflm-latt-nocell prior lattice information only (monoclinic lattice type, unique axis b, primitive centring): 500 images processed, 492 hits (98.4%), 410 indexable (83.3% of hits, 82.0% overall).

The unit-cell distribution implies a monoclinic lattice type, and the unique axis (required by

CrystFEL for monoclinic lattices) should be b, to overcome the axis exchange. We follow with a second run of indexing with `mosflm-latt-nocell`, but now specifying the lattice type in the `*.cell` file, for correct unit-cell parameters determination.

From figure 6.2, with input of 500 images, lattice type and centring information only (monoclinic, unique axis b, primitive), we could determine the unit-cell parameters by fitting of the most evident peaks using the `cell_explorer` script build in *CrystFEL*, that can be called from our script too. The unit-cell fitted by cell_explorer was: a= 38±1 b=78.6±0.7 c=46.0±0.7 α=90.0±0.2 β=102.1±0.5 γ=90.0±0.2.

Geometry file corrections

The first correction, that had an important impact in the indexing rate, is the detector distance from the sample. Following a previous procedure[31], we began changing the camera length 500 μm back and forth from the nominal length (0.0858m), with steps of 100 μm (table 6.1). From that, we observed a better indexing rate for -100 μm of camera displacement (0.0857 m) . We tested again in the interval of -200 μm to 0 μm of camera displacement, with steps of 20 μm (table 6.2). Finally, we found an optimal camera distance of 0.0857m (-100μm), with an increase of the indexing rate from 91.3% to 91.4% of hits.

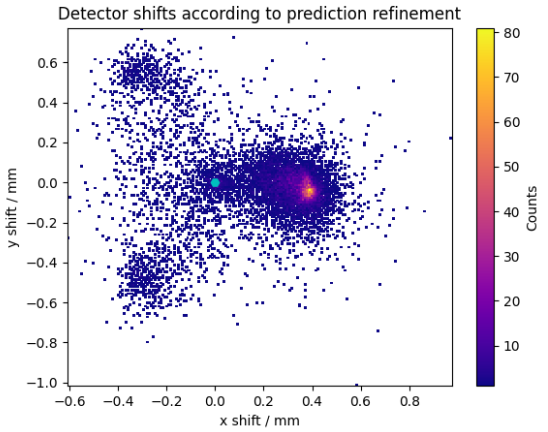
Camera displacement (μm)	Images processed	Hits (%)	Indexable (% of hits, % overall)
-500	14943	14635 (97.9)	13314 (91.0, 89.1)
-400	14956	14648 (97.9)	13328 (91.0, 89.1)
-300	14948	14640 (97.9)	13366 (91.3, 89.4)
-200	14941	14633 (97.9)	13349 (91.2, 89.3)
-100	14930	14622 (97.9)	13364 (91.4, 89.5)
0	14946	14638 (97.9)	13369 (91.3, 89.4)
100	14952	14644 (97.9)	13349 (91.2, 89.3)
200	14964	14656 (97.9)	13364 (91.2, 89.3)
300	14932	14624 (97.9)	13312 (91.0, 89.2)
400	14871	14563 (97.9)	13296 (91.3, 89.4)
500	14905	14597 (97.9)	13289 (91.0, 89.2)

Table 6.1: Detector distance from the sample optimization for 15000 random images of *AmeGH*128 dataset, with steps of 100μm.

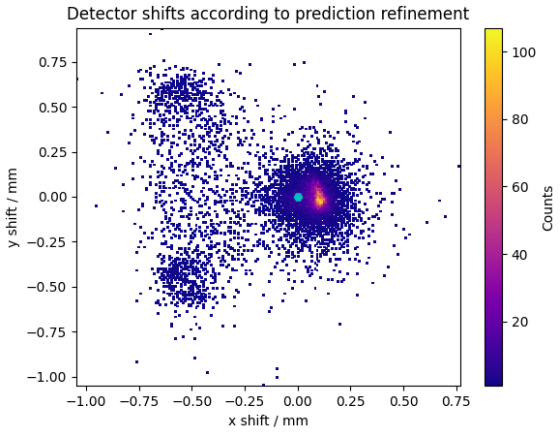
Camera displacement (μm)	Images processed	Hits (%)	Indexable (% of hits, % overall)
-200	14916	14608 (97.9)	13324 (91.2, 89.3)
-180	14947	14639 (97.9)	13344 (91.2, 89.3)
-160	14917	14609 (97.9)	13339 (91.3, 89.4)
-140	14931	14623 (97.9)	13340 (91.2, 89.3)
-120	14907	14599 (97.9)	13292 (91.0, 89.2)
-100	14921	14613 (97.9)	13355 (91.4, 89.5)
-80	14917	14609 (97.9)	13336 (91.3, 89.4)
-60	14788	14480 (97.9)	13214 (91.3, 89.4)
-40	14928	14620 (97.9)	13322 (91.3, 89.4)
-20	14935	14627 (97.9)	13364 (91.4, 89.5)
0	14946	14638 (97.9)	13324 (91.2, 89.3)

Table 6.2: Detector distance from the sample optimization for 15000 random images of *AmeGH*128 dataset, with steps of 20μm.

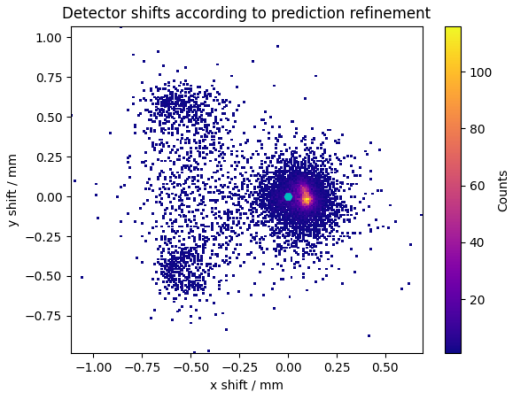
Detector distance to sample, in the way it is implemented here, still is very expensive computationally for real time data processing. During the experiment is better to tune the correct distance using a calibration sample with unit-cell parameters well know, as lysozyme, and look its unit-cell parameters distribution. If a, b, c are higher than expected, the correct detector distance might be lower than the nominal one, written in the geometry file. The same has to be adjusted in case of beam energy uncertainty, which is more often the case in XFELs facilities.



(a) One correction run: 12687 indexable images (86.4% of hits, 84.6% overall).



(b) Two correction runs: 12627 indexable images (86.0% of hits, 84.3% overall).



(c) Three correction runs: 12646 indexable images (86.1% of hits, 84.4% overall).

Figure 6.3: Beam position correction using 15000 random images of *AmeGH*128 dataset: shift maps after n correction runs, initial 12124 indexable images (82.6% of hits, 80.9% overall.)

Secondly, we investigated the beam shift predicted by *CrystFEL*, and applied an average correction of the beam center position in the geometry file. We performed three subsequent automatic corrections (figure 6.3), calling the *CrystFEL* script *detector-shift*. Each point in the heat map is related to the detector shift value, calculated by *CrystFEL* for each crystal comparing reflections position observed and calculated. The visible clusters of detector shifts, selected by clicking, are directly corrected in the geometry file.

The beam-shift correction that had a better indexing rate (first run of figure 6.3) was selected for the next steps. After all detector corrections the unit-cell fitted by `cell_explorer` was: a= 37.4±0.9 b=79±1 c=47±2 α=90.0±0.2 β=102.1±0.5 γ=90.0±0.2. The final unit-cell file is used as reference in the next stages ((Appendix A)). The unit-cell file (`*.cell`), with a tolerance of 5% in a, b and c, and 1.5% on α, β and γ.

ID	Method	Images processed	Hits (%)	Indexable (% of hits, % overall)
0	asdf-nolatt-cell	601	556 (92.5)	0 (0.0, 0.0)
1	mosflm-nolatt-nocell'	601	556 (92.5)	52 (9.4, 8.7)
2	mosflm-latt-cell	601	556 (92.5)	50 (9.0, 8.3)
3	mosflm-latt-nocell	601	556 (92.5)	61 (11.0, 10.1)
4	dirax-nolatt-nocell'	601	556 (92.5)	0 (0.0, 0.0)
5	xds-nolatt-nocell'	601	556 (92.5)	16 (2.9, 2.7)
6	taketwo-latt-cell	601	556 (92.5)	0 (0.0, 0,0)
7	xgandalf-nolatt-cell	601	556 (92.5)	20 (3.6, 3.3)
8	xds-latt-cell	601	556 (92.5)	207 (37.2, 34.4)
9	xgandalf-nolatt-nocell	601	556 (92.5)	0 (0.0, 0.0)
10	mosflm-nolatt-cell	601	556 (92.5)	36 (6.5, 6.0)
11	asdf-nolatt-nocell	601	556 (92.5)	1 (0.2, 0.2)

Table 6.3: *CrystFEL*'s indexing methods evaluation for 5000 random diffraction patterns of *Ame*GH128 enzyme.

Indexing and integration

Five thousand random images (5000) were indexed by all 12 methods available in *CrystFEL*, following the order in table 6.3. Figure 6.4 compares the indexing rate (% of hits), mean peaks per pattern (MPP), mean ADU-intensity per peak (MAP) for each indexing algorithm.

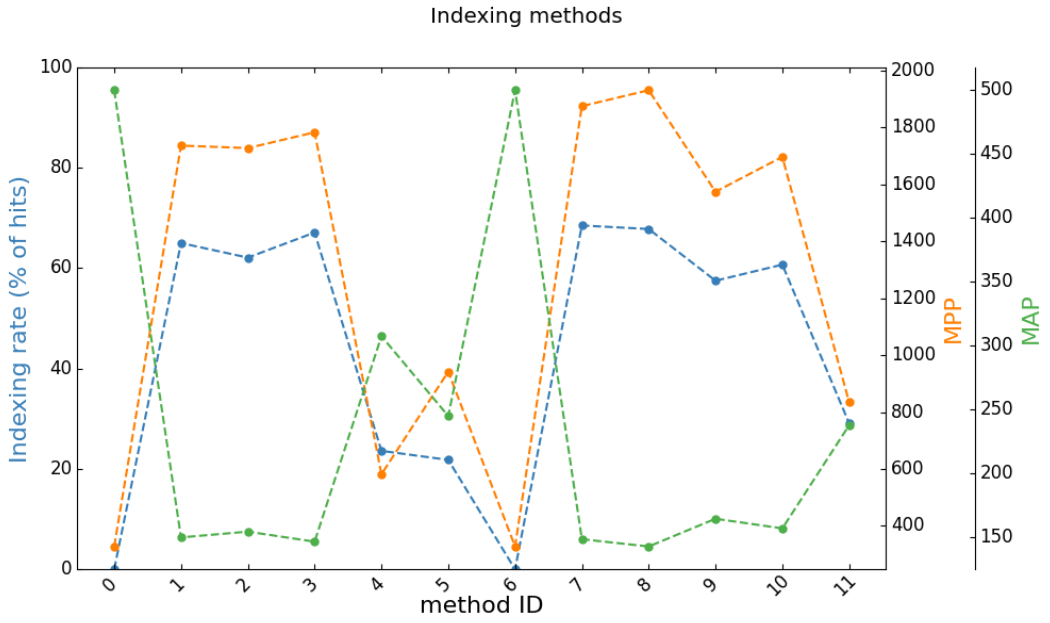


Figure 6.4: Indexing methods comparison for 5000 random images of *Ame*GH128 crystals.

The methods that got better performance, indexing rate >1% of hits, were selected to the next

indexing stage. There, the goal is to achieve the maximum of indexed patterns from the whole dataset (64800). We used a descending order, according to the indexing rate performance, for the last indexing and integration process in order to optimize the processing time. The final sequence of chosen indexing methods were:

1 - xgandalf-nolatt-cell
2- xds-latt-cell
3- mosflm-latt-nocell
4- mosflm-nolatt-nocell
5- mosflm-latt-cell
6- mosflm-nolatt-cell
7- xgandalf-nolatt-nocell
8- asdf-nolatt-nocell
9- dirax-nolatt-nocell
10- xds-nolatt-nocell

The final dataset selected, using the unit-cell as reference, has the distribution of figure 6.5, and has 51946 indexed images (81.8% of hits, 80.2% overall)

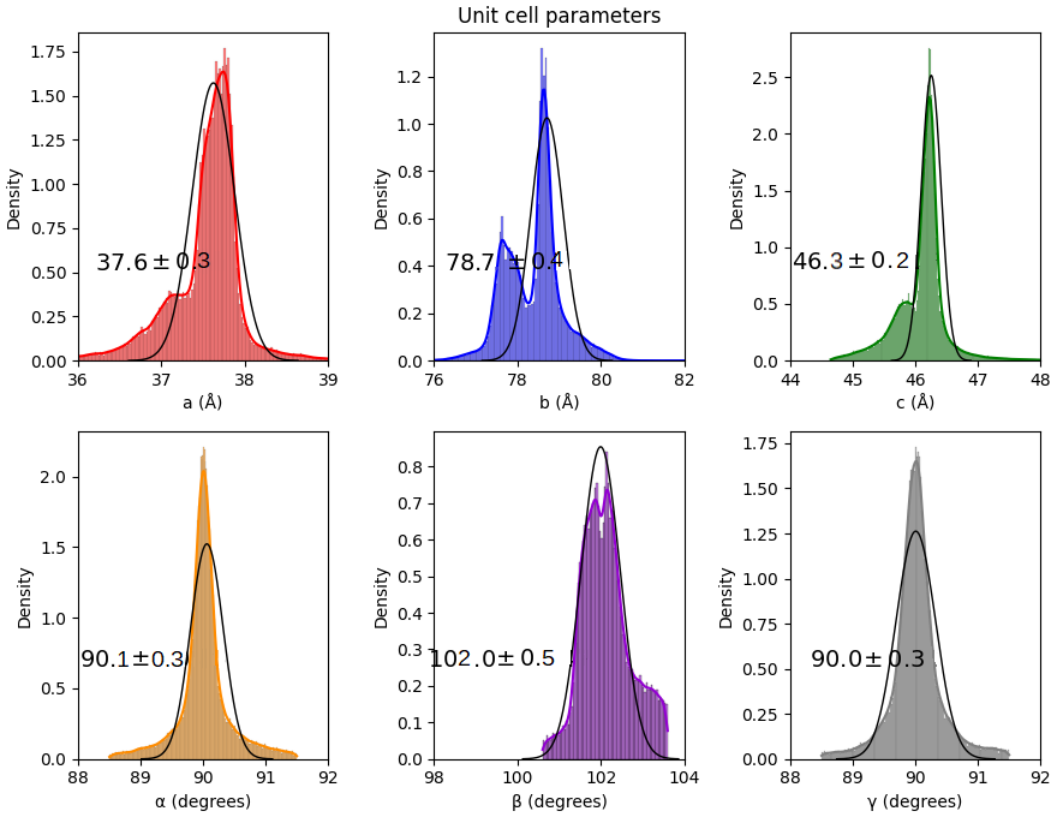


Figure 6.5: Final unit-cell distribution for 64800 diffraction patterns of *Ame*GH128 enzyme, indexed with the sequence of best methods and reference unit-cell check.

Images processed	Hits (%)	Indexable (% of hits, % overall)
500	492 (98.4)	411 (83.5, 82.2)
1000	981 (98.1)	798 (81.3, 79.8)
2500	2459 (98.4)	2020 (82.1, 80.8)
5000	4894 (97.9)	3999 (81.7, 80.0)
10000	9793 (97.9)	7973 (81.4, 79.7)
25000	24513 (98.1)	20061 (81.8, 80.2)
50000	49005 (98.0)	40051 (81.7, 80.1)
64800	63528 (98.0)	51946 (81.8, 80.2)

Table 6.4: Evolution of indexed images as *Ame*GH128 random diffraction patterns were included in the final dataset.

Merging, scaling, post refinement

With the indexing methods selected in the last section, integration on default parameters, geometry file corrected, and unit-cell fitted in the first part, we merged the subdatasets. Then, we analyzed the evolution of the figures of merit and data quality (figures 6.5 to 6.10) as more images were included in the final dataset, (64800, 50000, 25000, 10000, 5000, 2500, 1000, 500 random images).

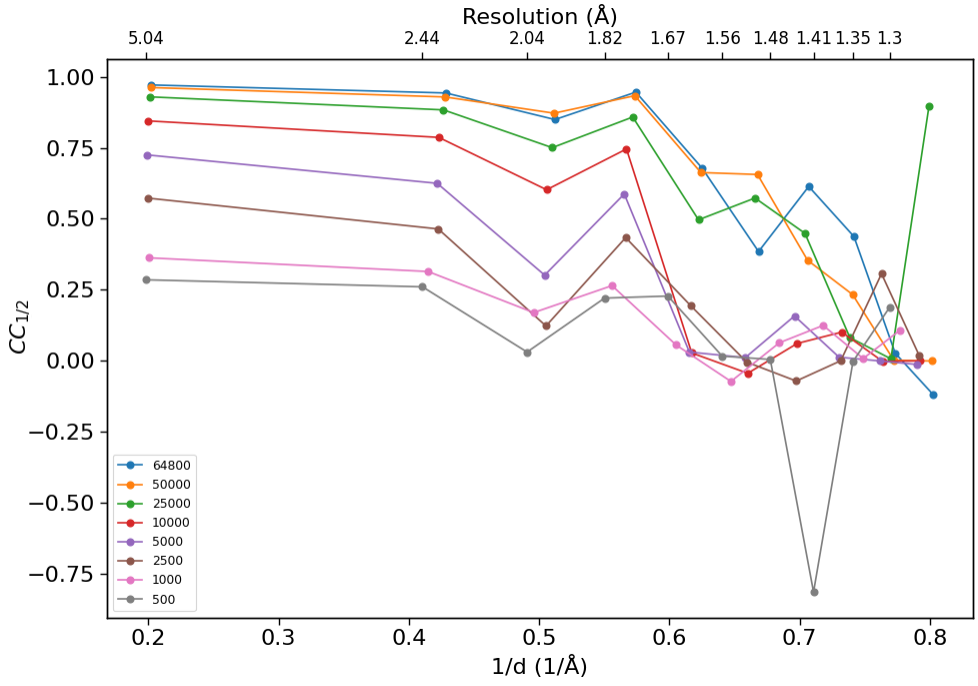


Figure 6.6: $CC_{1/2}$ figure of merit according to the number of *AmeGH*128 diffraction patterns measured.

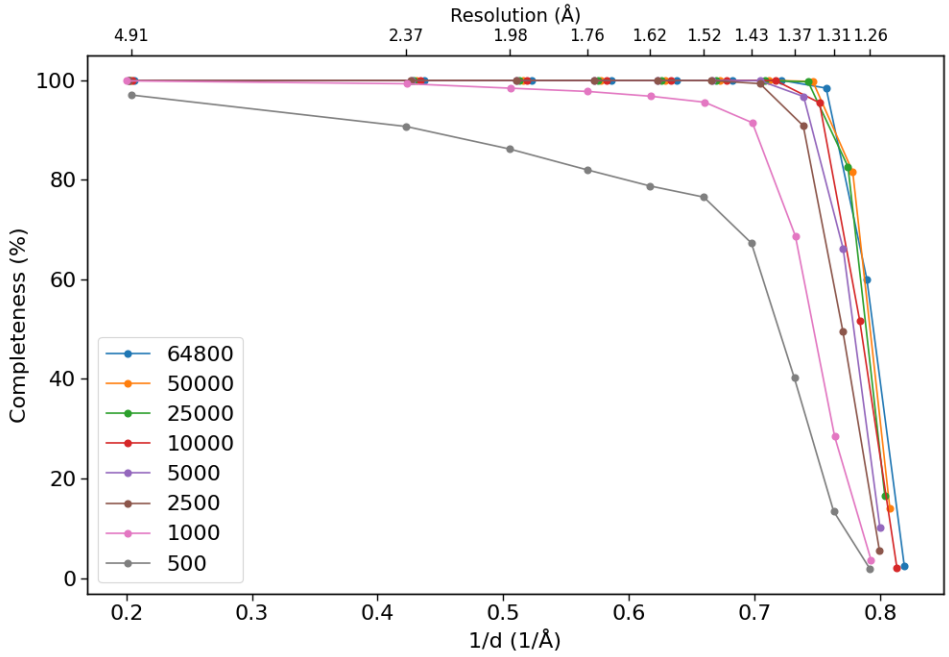


Figure 6.7: Completeness figure of merit according to the number of *AmeGH*128 diffraction patterns measured.

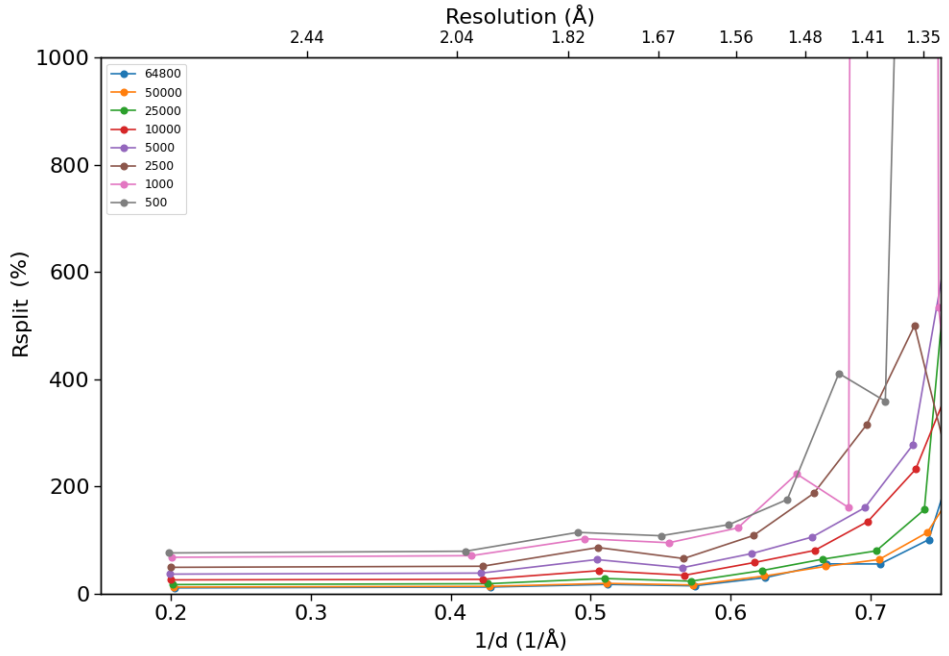


Figure 6.8: R_{split} figure of merit according to the number of *AmeGH*128 diffraction patterns measured.

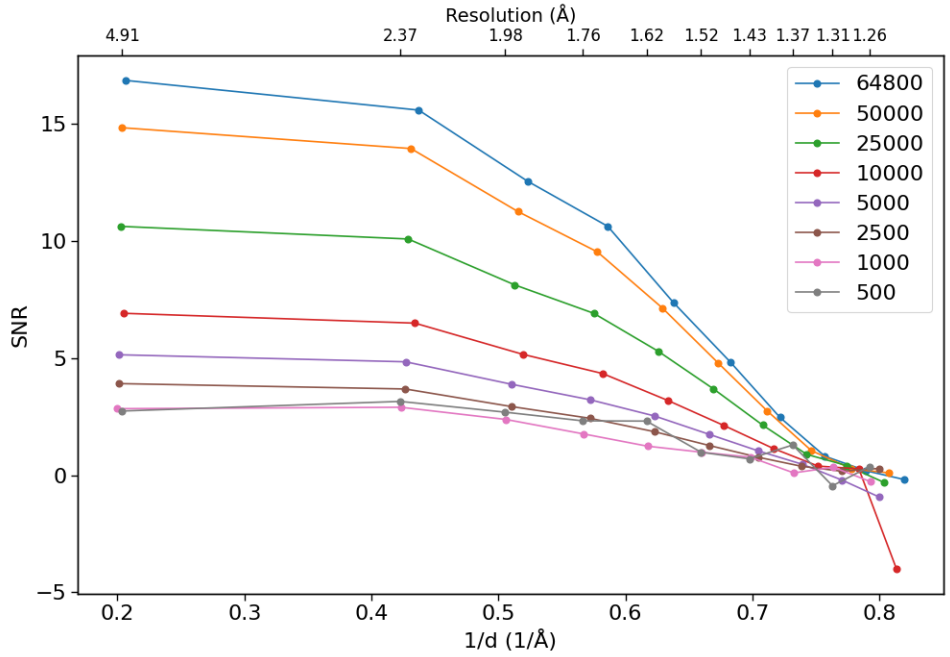


Figure 6.9: SNR figure of merit according to the number of *AmeGH*128 diffraction patterns measured.

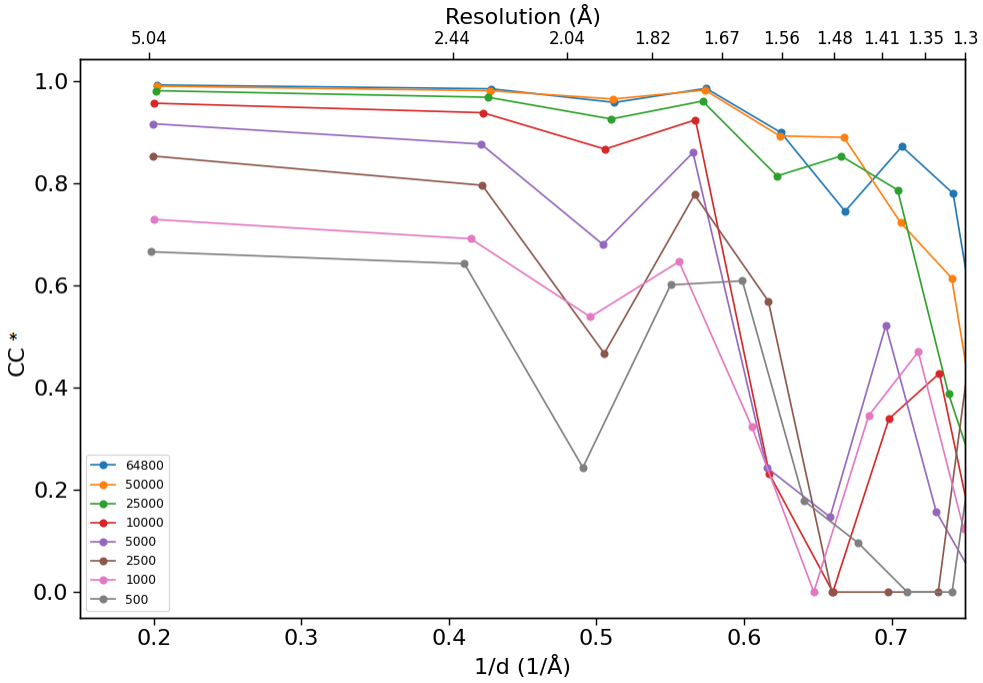


Figure 6.10: CC* figure of merit according to the number of *AmeGH*128 diffraction patterns measured.

6.1.3 *ccCluster* data processing

From each complete dataset, we took the first five degrees and five degrees perpendicular to the first images. The wedges of 5° were merged using `XDS`. Then, the reflections list file (`XDS ASCII.HKL`) were used as an input to *ccCluster*. We came up with 41 subdatasets of *Ame*GH128 for processing with the HCA routine.

Firstly, we tested the hierarchical distance based on the intensity reflections correlation coefficient (cc). From the *ccCluster* dendrogram, we set, in the automatic pipeline, two different thresholds interval: 0.45-0.95 (fig. 6.11) and 0.18-0.98 (fig. 6.12) with five threshold points, equally spaced. Tables 6.5 and 6.6 shows the threshold tested and the number of subdatasets include in the biggest cluster for each threshold.

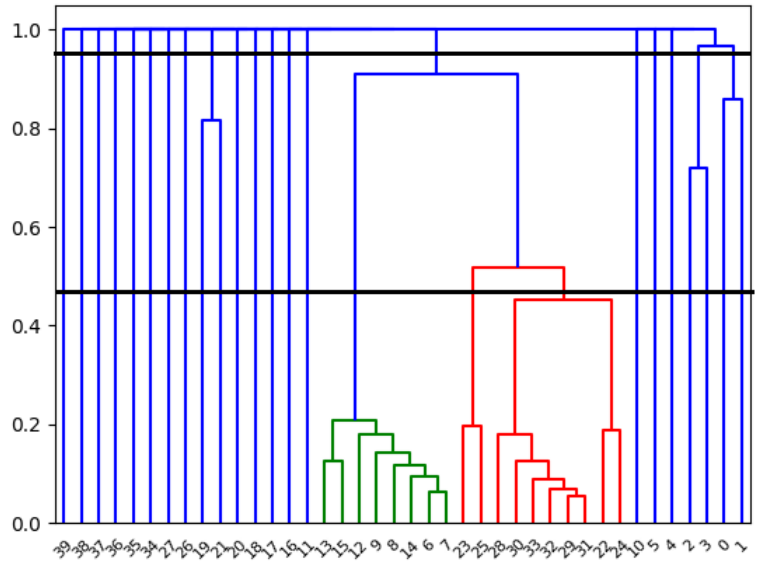


Figure 6.11: Dendrogram of 41 *Ame*GH128 crystals wedges of 5° each. In black, it is the interval (0.45-0.95) from which the biggest cluster was merged for 5 different threshold values.

Threshold	Number of subdatasets in the biggest cluster
0.45	8
0.57	10
0.7	10
0.82	10
0.95	18

Table 6.5: Threshold points tested from 0.45-0.95 for 41 wedges of 5° of *Ame*GH128 crystals and their respective number of subdatasets included in the biggest cluster.

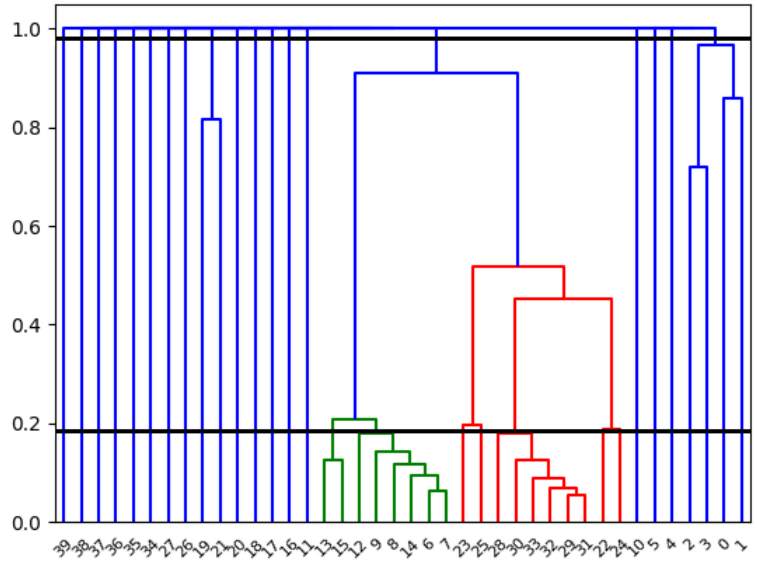


Figure 6.12: Dendrogram of 41 *Ame*GH128 crystals wedges of 5° each. In black, it is the interval (0.18-0.98) from which the biggest cluster was merged for 5 different threshold values.

Threshold	Number of subdatasets in the biggest cluster
0.18	6
0.38	8
0.58	10
0.78	10
0.98	18

Table 6.6: Threshold points tested from 0.18-0.98 for 41 wedges of 5° of *Ame*GH128 crystals and their respective number of subdatasets included in the biggest cluster.

The automatic pipeline takes the `XSCALE` output (`XSCALE.LP`) for each threshold tested and plot them comparatively. Figures 6.13 to 6.15 corresponds to the figures of merit obtained for the first interval. From figure 6.16 to 6.18 corresponds to the same control cards obtained for the second interval.

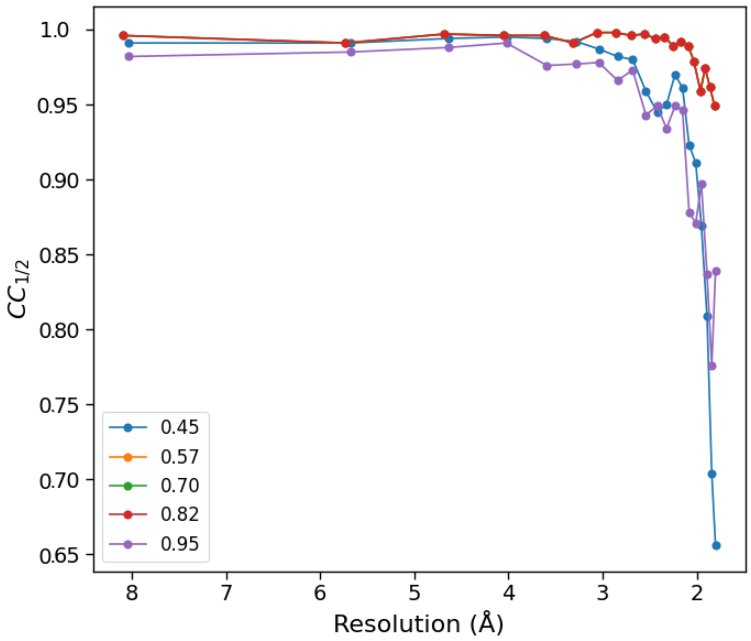


Figure 6.13: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: $CC_{1/2}$ according to the threshold value used to choose the biggest cluster (threshold limits 0.45 to 0.95).

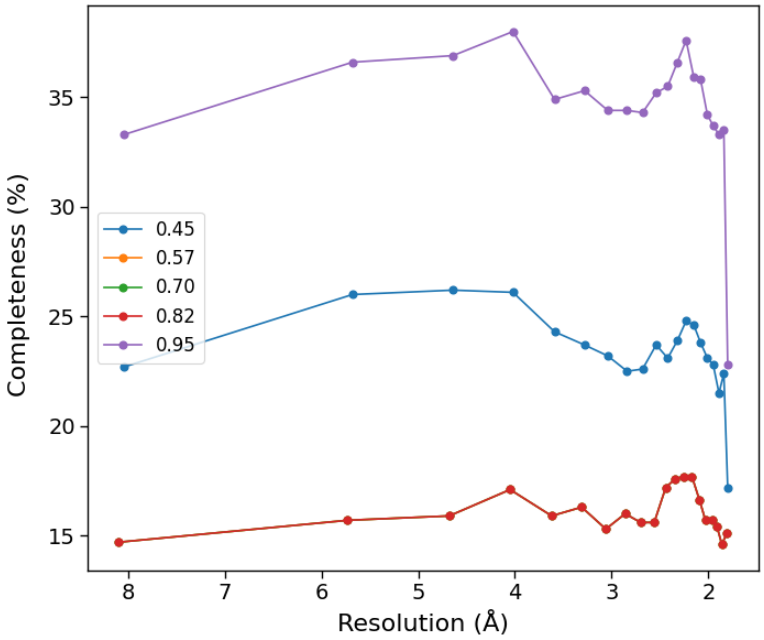


Figure 6.14: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: Completeness according to the threshold value used to choose the biggest cluster (threshold limits 0.45 to 0.95).

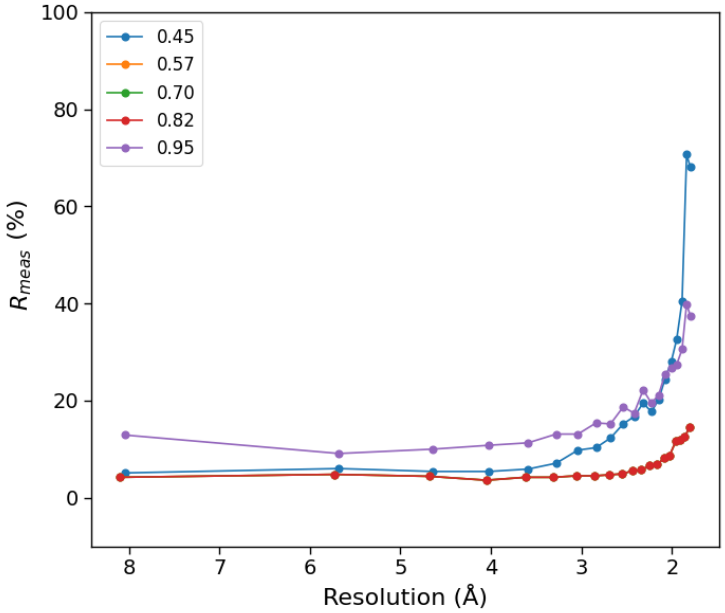


Figure 6.15: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: R_{meas} according to the threshold value used to choose the biggest cluster (threshold limits 0.45 to 0.95).

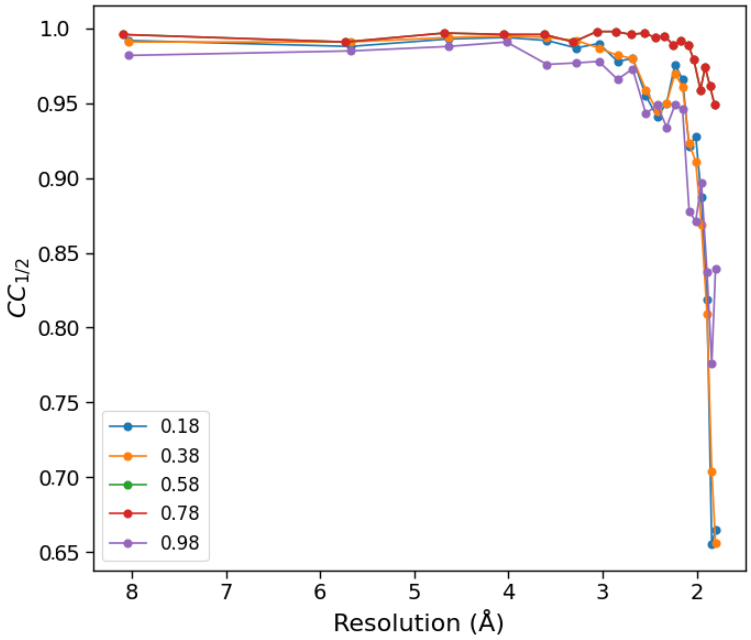


Figure 6.16: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: $CC_{1/2}$ according to the threshold value used to choose the biggest cluster (threshold limits 0.18 to 0.98).

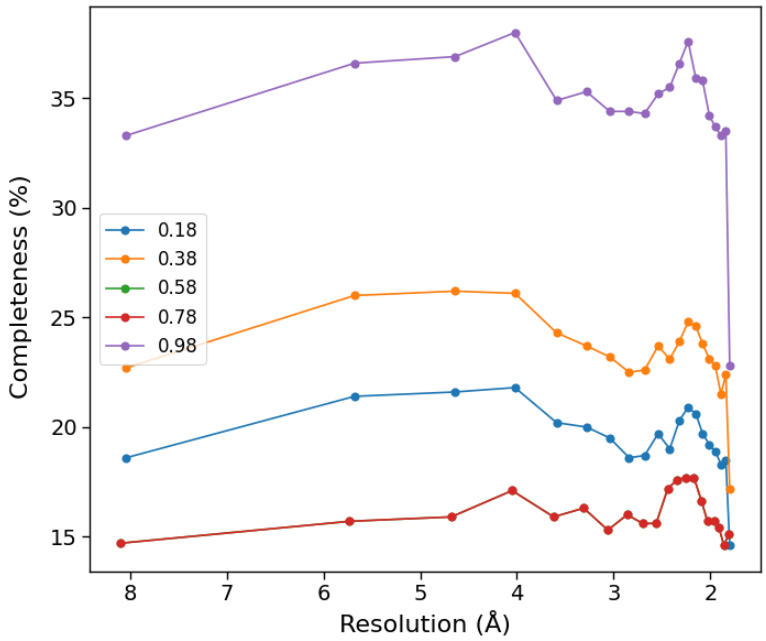


Figure 6.17: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: Completeness according to the threshold value used to choose the biggest cluster (threshold limits 0.18 to 0.98).

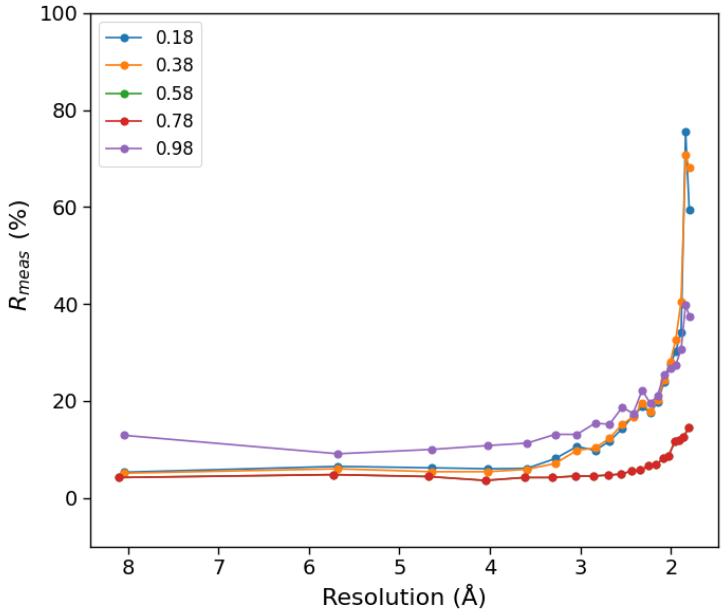


Figure 6.18: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: R_{meas} according to the threshold value used to choose the biggest cluster (threshold limits 0.18 to 0.98).

The *ccCluster* calculated an estimated threshold of 0.95 (figure 6.19). For this cluster, with

its scaled and merged intensity reflection list (`scaled.hkl`), the user is able to solve the structure with the *ccCluster* output files, with their preferred crystallography packages.

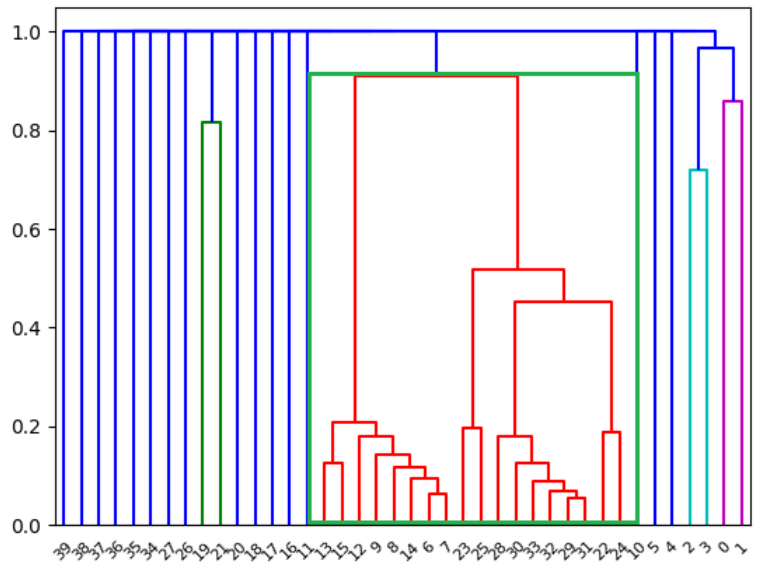


Figure 6.19: Estimated threshold (0.95) for the best cluster (green) suggested by *ccCluster* for 41 *Ame*GH128 crystals subdatasets (wedges of 5°).

We also tested another metric for distance matrix calculation available on *ccCluster*, the unit-cell variation, for the same 41 subdatasets. We tested an interval of 0.05 - 4, with 5 points equally spaced (table 6.7). The best cluster estimated by *ccCluster* was 0.05 (green in figure 6.20).

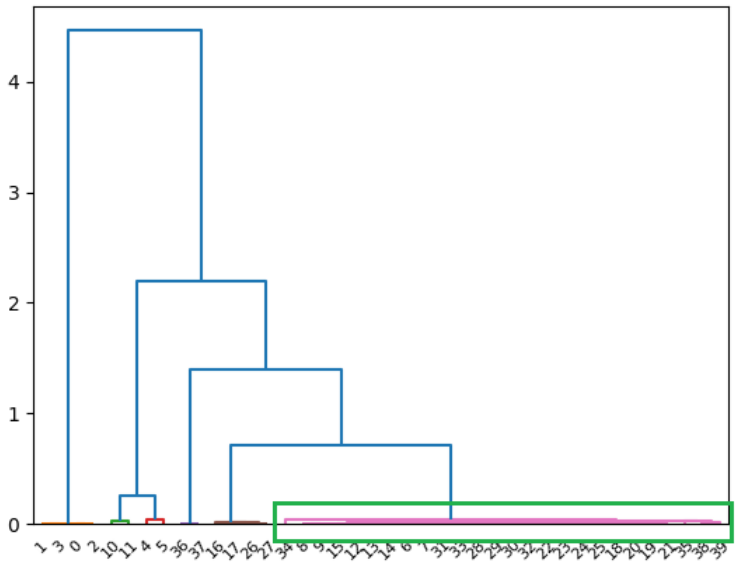


Figure 6.20: Estimated threshold for the best cluster suggested by *ccCluster* for 41 *Ame*GH128 crystals subdatasets (wedges of 5°) with unit-cell variation as hierarchical distance.

Threshold	Number of subdatasets in the biggest cluster
0.05	26
1.04	30
2.02	32
3.01	36
4.00	36

Table 6.7: Threshold points tested from 0.18-0.98 for 41 wedges of 5° of *Ame*GH128 crystals and their respective number of subdatasets included in the biggest cluster.

For each threshold tested, we took the output from *XSCALE*, and plotted them comparatively. Figures 6.21 to 6.23 corresponds to the figures of merit obtained for these clusters.

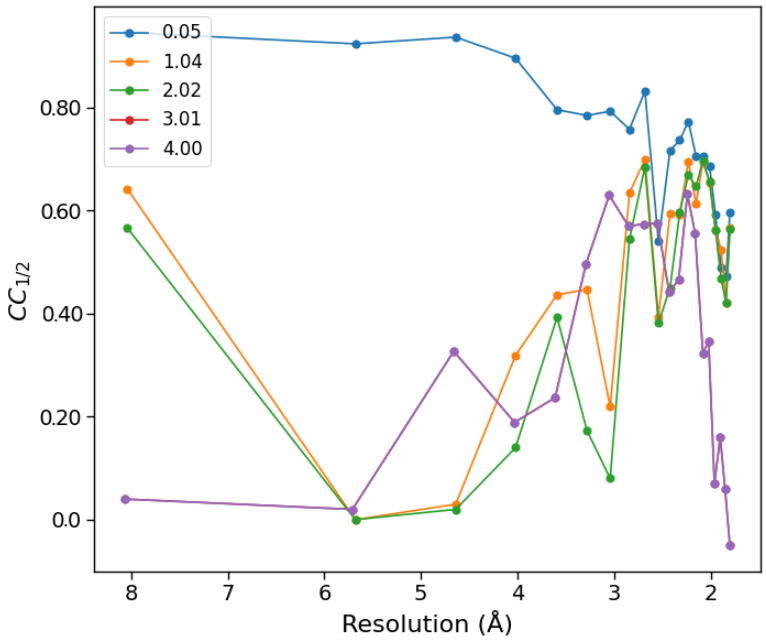


Figure 6.21: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: $CC_{1/2}$ according to the threshold value used to choose the biggest cluster (threshold limits 0.05 to 4, hierarchical distance of unit-cell variation).

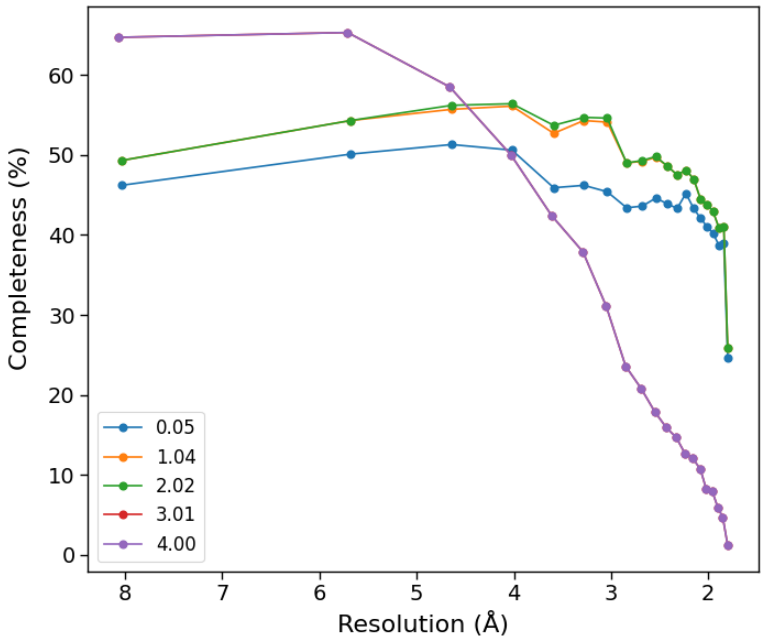


Figure 6.22: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: Completeness according to the threshold value used to choose the biggest cluster (threshold limits 0.05 to 4, hierarchical distance of unit-cell variation).

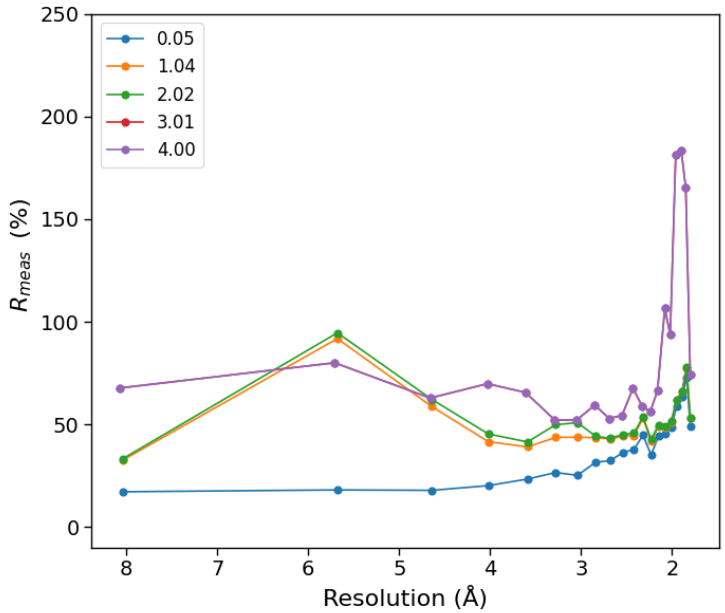


Figure 6.23: Figures of merit for 41 *Ame*GH128 crystals wedges of 5°: R_{meas} according to the threshold value used to choose the biggest cluster (threshold limits 0.05 to 4, hierarchical distance of unit-cell variation).

6.1.4 Final discussions

The *CrystFEL* routine demonstrated to be useful for handling thousands of diffraction patterns in low partiality datasets. The statistics are higher compared to a real SX experiment, as the images were collected in oscillation mode. Even though, it is interesting to notice the number of images range in which the *CrystFEL* data processing is capable to achieve a final dataset that behaves similar as a conventional crystallography data. For a small number of images (500 to 5000) the data quality, most notice by $CC_{1/2}$ (fig. 6.6) and R_{split} (fig. 6.8), is worst than in the others subdatasets. Actually, at the SX regime (thousands of indexed images), the figures of merit indicate that we achieved quite reliable data.

The *ccCluster* routine demonstrated straightforward and less expensive computationally although it restricts the kind of experiments you may perform due to the lack of flexibility in low-level parameters. Both are of great interest to make it available on Manacá, enabling a range of SX experiments. The automatic pipeline can accommodate them and give meaningful results. The main restriction of *ccCluster* may come intrinsically from your subdasets, that might not be well clusterized. To solve that, another clustering algorithms (K-means clustering, for example), or other distance metric inside *ccCluster* might be applied, as it was demonstrated in section 6.1.3.

6.2 Grid-scan of lysozyme cryocooled crystals

6.2.1 Experimental setup

We measured 64 lysozyme crystals manually mounted on nylon loops and kept in Uni-Pucks at cryogenic temperature (77K). The experimental setup is based on previous similar experiments[40] [41]. We used the Manacá automated sample changer (Stäubli robotic arm) to deliver the sample to the goniometer. The data collection was done by running an in-house build Python routine, based on the LNLS (SOL) fly-scan script. We measured the biggest dimensions of the crystal and took equally spaced patterns from a matrix of points in the biggest crystal face. The crystal dimensions varied from 200 to 400 μm.

We began SX tests on Manacá using the fixed-target method for sample-delivery, and larger crystals. The experiment demonstrated suitable due to its high hit-rate and stable sample delivery which severely decreases the number of crystals needed to achieve good data quality.

The beam size used was 25μm, at 12.68 keV, with estimated flux of the order of 10^{12}ph/s, 5% of transmission, 125 mm of distance from detector to sample. Between each measurement point, the goniometer performed a translation of 25 μm and a rotation of 0.1° (fig. 6.24). It was collected around 40 images per crystal, with 30 seconds to measure each point, each crystal took around 30 minutes to be collected. Therefore, the total experiment took 4 shifts (around 40h total), with 16 crystals measured per shift.

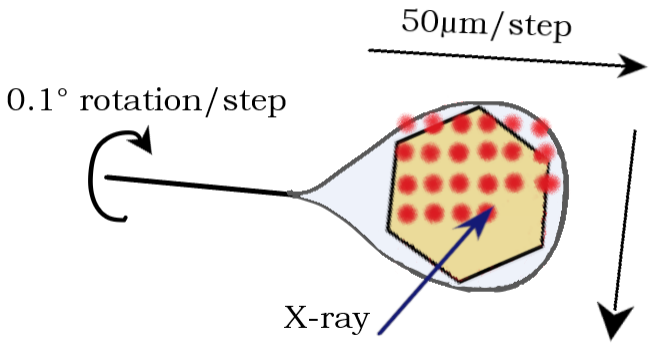


Figure 6.24: Schematic of the grid-scan experimental setup based on previous studies [41] [40]

6.2.2 *CrystFEL* data processing

Unit-cell parameters determination

Initial tests for peak search included the *zaef* method: `-threshold=100 and 10 -min -squared -gradient=5000 -min-snr=4 -peak-radius=4,5,7`, and *peakfinder8*: `-threshold=100 and 10 -min-snr=3 -min-pix-count=2 -max -pix -count=20 -local-bg-radius=6`. The best indexing performance obtained was –peaks=*zaef* `-threshold=10 -min-squared-gradient=5000 -min-snr=4 -peak-radius=4,5,7`. The peak search parameters might be better optimized for this dataset, and it is already implement in our routine. From that, one might notice a correlation between stronger/weaker reflections selection and better indexing rate or data quality.

Firstly, the indexing method used was `mosflm-latt-nocell` with the simplest lattice type triclinic and primitive unit-cell. That simplification considerably accelerates the indexing step. The unit-cell parameters founded (figures 6.25 and 6.26) match with the known numbers for lysozyme (a=79 Å b=79 Å c=37 Å α, β, γ=90°), except for an exchange between the a and c axis.

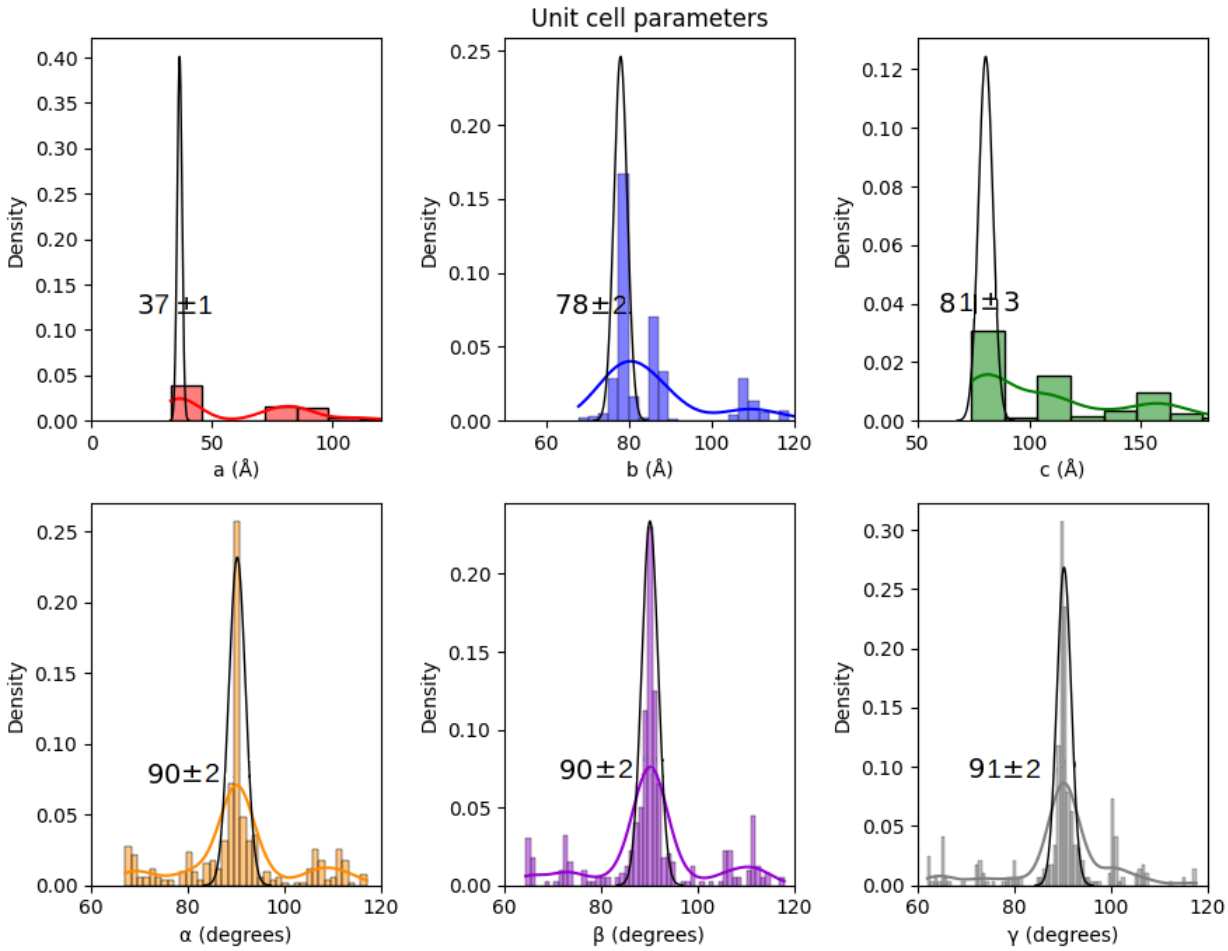


Figure 6.25: Unit-cell parameters distribution from 500 random images of lysozyme cryocooled crystals, indexed with mosflm-latt-nocell and the simplest prior lattice information (triclinic lattice type, primitive centring): 499 images processed, 454 hits (91.0%), 412 indexable (90.7% of hits, 82.6% overall).

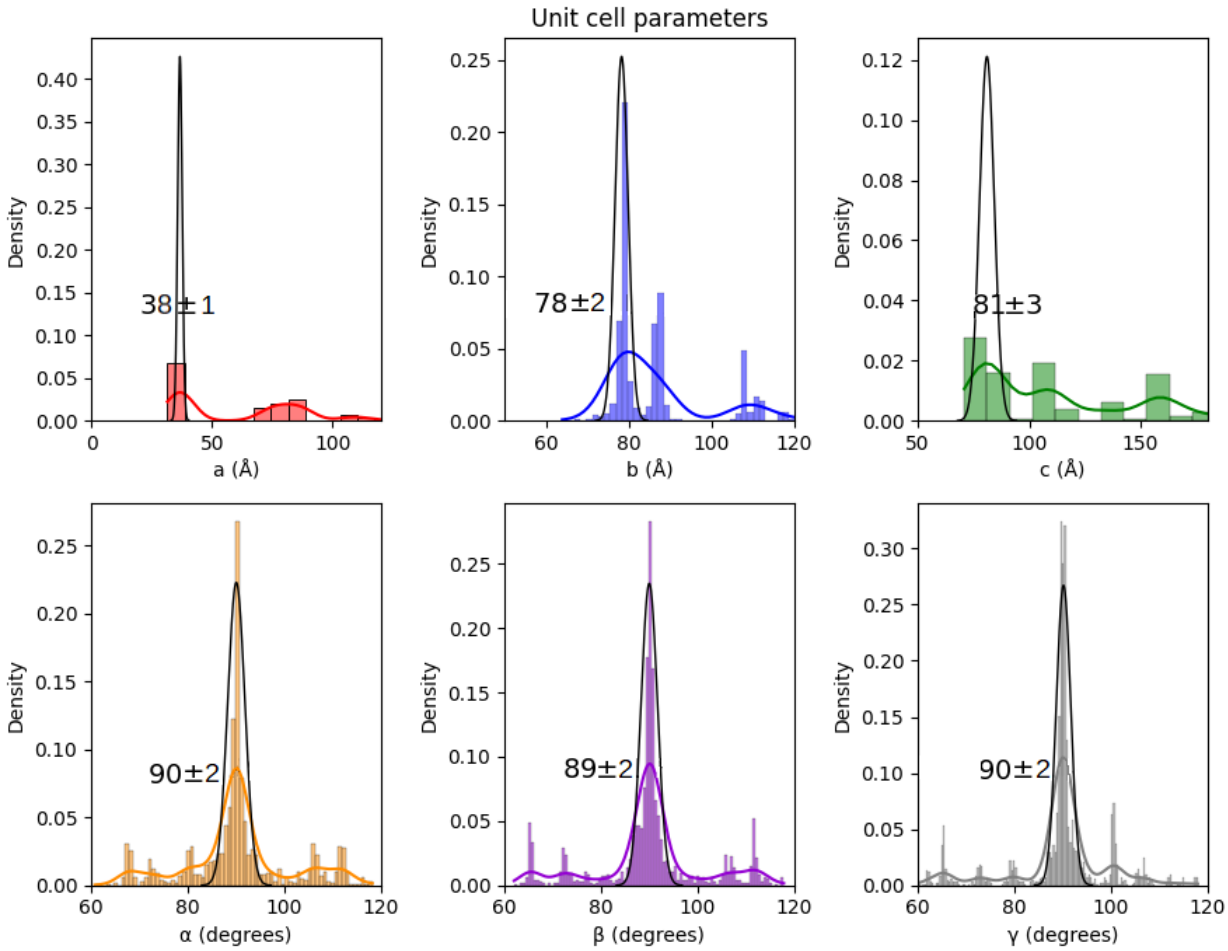


Figure 6.26: Unit-cell parameters distribution from 2910 random images of lysozyme cryocooled crystals, indexed with mosflm-latt-nocell and the simplest prior lattice information (triclinic lattice type, primitive centring): 2907 images processed, 2613 hits (89.9%), 2337 indexable (89.4% of hits, 80.4% overall).

The unit-cell distribution implies a tetragonal lattice type, and the unique axis (required by *CrystFEL* for tetragonal lattices) should be c to overcome the indexing mistake. That is written in the unit-cell file for the correct unit-cell parameters determination with a second run of indexing with mosflm-latt-nocell, but now with these additional information.

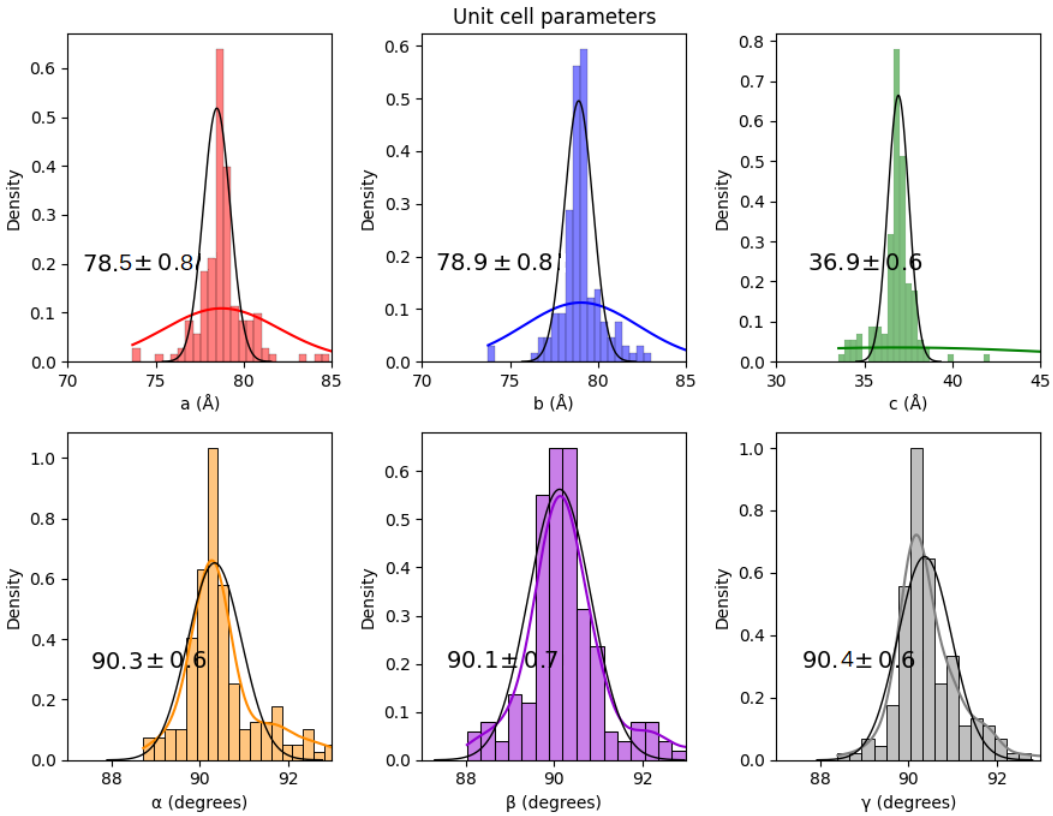


Figure 6.27: Unit-cell parameters distribution from 500 random images of lysozyme cryocooled crystals, indexed with mosflm-latt-nocell prior lattice information only (tetragonal lattice type, unique axis c primitive centring): 500 images processed, 455 hits (91.0%), 165 indexable (36.3% of hits, 33.0% overall).

From figure 6.27, with input of 500 images, lattice type and centring information only (tetragonal, unique axis c, primitive), we could determine the unit-cell parameters by fitting of the most evident peaks using the *cell_explorer* script build in *CrystFEL*, that can be called from our script too. The unit-cell fitted by *cell_explorer* was: a= 78.7±0.4 b=79.0±0.3 c=37.0±0.7 α=90.3±0.3 β=90.1±0.6 γ=90.2±0.3. The cell parameters histogram might be further straightened to a Gaussian distribution with sequential detector distance and beam shift corrections [31].

Geometry file corrections

The first correction is the detector distance from the sample. Following the same procedure as in 6.1.2, we began changing the camera length 500 μm back and forth the nominal length (0.125m), with steps of 100 μm (table 6.8). From that we observe a better indexing rate for camera length of 0.1247μm (camera displacement -300μm). We tested again on the interval of -400 and -100 μm with steps of 20 μm (table 6.9). The final corrected camera distance founded was 0.1247μm (camera displacement -300μm), with an improvement of the indexing rate from 89.4% to 89.6% of hits.

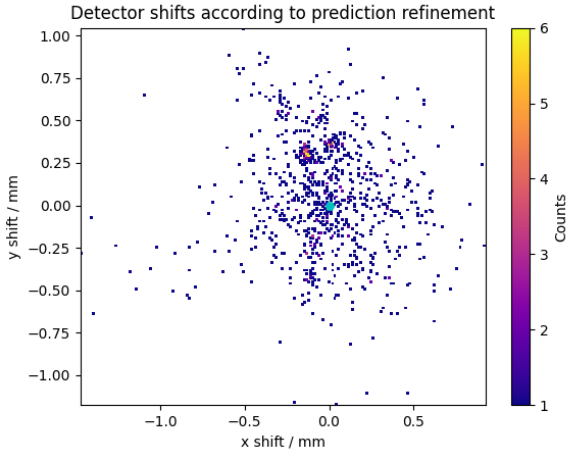
Camera displacement (μm)	Images processed	Hits (%)	Indexable (% of hits, % overall)
-500	2909	2615 (89.9)	2337 (89.4, 80.3)
-400	2909	2615 (89.9)	2332 (89.2, 80.2)
-300	2906	2612 (89.9)	2340 (89.6, 80.5)
-200	2907	2613 (89.9)	2331 (89.2, 80.2)
-100	2909	2615 (89.9)	2335 (89.3,80.3)
0	2903	2609 (89.9)	2333 (89.4, 80.4)
100	2905	2611 (89.9)	2317 (88.7, 79.8)
200	2908	2614 (89.9)	2324 (88.9, 79.9)
300	2903	2609 (89.9)	2322 (89.0, 80.0)
400	2907	2613 (89.9)	2320 (88.8, 79.8)
500	2908	2614 (89.9)	2315 (88.6, 79.6)

Table 6.8: Detector distance from the sample optimization for 2910 images of the cryocooled lysozyme dataset, with steps of 100μm.

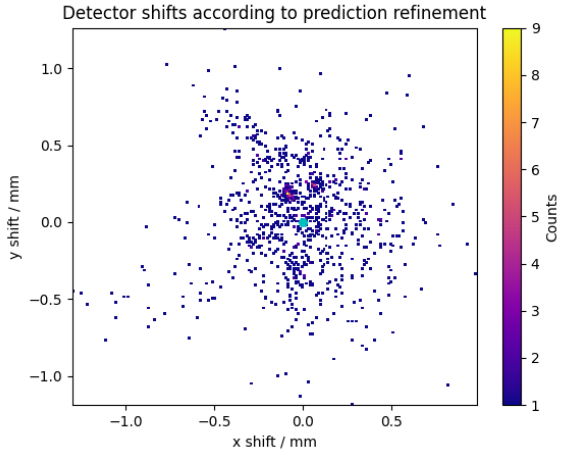
Camera displacement (μm)	Images processed	Hits (%)	Indexable (% of hits, % overall)
-400	2907	2613 (89.9)	2330 (89.2, 80.2)
-380	2904	2610 (89.9)	2330 (89.3, 80.2)
-360	2909	2615 (89.9)	2335 (89.3, 80.3)
-340	2906	2612 (89.9)	2331 (89.2, 80.2)
-320	2906	2612 (89.9)	2333 (89.3, 80.3)
-300	2906	2612 (89.9)	2340 (89.6, 80.5)
-280	2906	2612 (89.9)	2326 (89.1, 80.0)
-260	2903	2609 (89.9)	2329 (89.3, 80.2)
-240	2908	2614 (89.9)	2333 (89.3, 80.2)
-220	2906	2612 (89.9)	2321 (88.9, 79.9)
-200	2907	2613 (89.9)	2331 (89.2, 80.2)
-180	2907	2613 (89.9)	2330 (89.2, 80.2)
-160	2907	2613 (89.9)	2327 (89.1, 80.0)
-140	2909	2615 (89.9)	2330 (89.1, 80.1)
-120	2908	2614 (89.9)	2332 (89.2, 80.2)
-100	2908	2614 (89.9)	2334 (89.3, 80.3)

Table 6.9: Detector distance from the sample optimization for 2910 images of the cryocooled lysozyme dataset, with steps of 20μm.

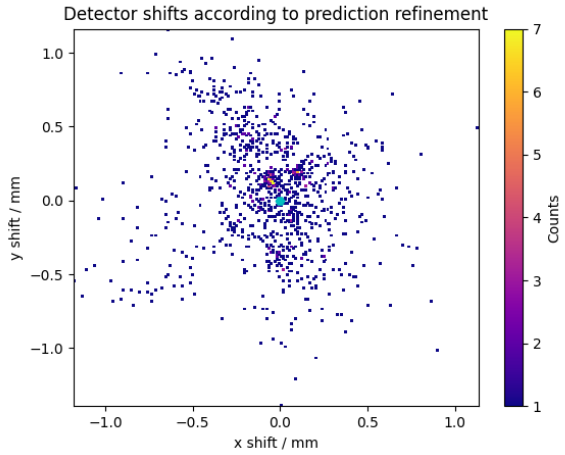
Secondly, we investigated the beam shift predicted by *CrystFEL*, and applied an average correction of the beam center position in the geometry file. We performed three subsequent automatic corrections (figure 6.28), calling the *CrystFEL* script `detector-shift`. The most visible clusters of detector shifts, selected by clicking, were directly corrected in the geometry file. Here, the total of images was considerably smaller than in section 6.1.2, so it is harder to observe a recurrence of shifts in the detector, as in figure 6.3.



(a) One correction run: 983 indexed images (37.6% of hits, 33.8% overall).



(b) Two correction runs: 1001 indexed images (38.3% of hits, 34.4% overall).



(c) Three correction runs: 1086 indexed images (41.5% of hits, 37.3% overall).

Figure 6.28: Beam position correction using 2910 images of lysozyme at room temperature dataset: shift maps after n automatic correction runs using the *detector-shift* *CrystFEL*'s script. Initial indexed images 949 (36.3% of hits, 32.6% overall)

The beam-shift correction that had a better indexing rate (third run of figure 6.28) was selected for the next steps. After all detector corrections (camera length and beam shift) the unit-cell fitted by `cell_explorer` was: a= 78.6±0.6 b=78.8±0.6 c=36±1 α=90.3±0.4 β=90.2±0.6 γ=90.2±0.3. The final unit-cell file is used as reference in the next stages ((Appendix A)).

Indexing and integration

The whole dataset (2910 images) were indexed, individually, by all 12 methods available in *CrystFEL*, following the order in table 6.10. Figure 6.29 compares the indexing rate (% of hits),

mean peaks per pattern (MPP), mean ADU-intensity per peak (MAP) for each indexing algorithm.

ID	Method	Images processed	Hits (%)	Indexable (% of hits, % overall)
0	asdf-nolatt-cell	2909	2614 (89.9)	0 (0.0, 0.0)
1	mosflm-nolatt-nocell'	2910	2615 (89.9)	555 (21.2, 19.1)
2	mosflm-latt-cell	2910	2615 (89.9)	782 (29.9, 26.9)
3	mosflm-latt-nocell	2910	2615 (89.9)	691 (26.4, 23.7)
4	dirax-nolatt-nocell'	2909	2614 (89.9)	217 (8.3, 7.5)
5	xds-nolatt-nocell'	2910	2615 (89.9)	108 (4.1, 3.7)
6	taketwo-latt-cell	2910	2615 (89.9)	0 (0.0, 0,0)
7	xgandalf-nolatt-cell	2910	2615 (89.9)	704 (26.9, 24.2)
8	xds-latt-cell	2910	2615 (89.9)	1294 (49.5, 44.5)
9	xgandalf-nolatt-nocell	2910	2615 (89.9)	199 (7.6, 6.8)
10	mosflm-nolatt-cell	2910	2615 (89.9)	588 (22.5, 20.2)
11	asdf-nolatt-nocell	2910	2615 (89.9)	167 (6.4, 5.7)

Table 6.10: *CrystFEL*'s indexing methods evaluation for 2910 diffraction patterns of lysozyme cryocooled.)

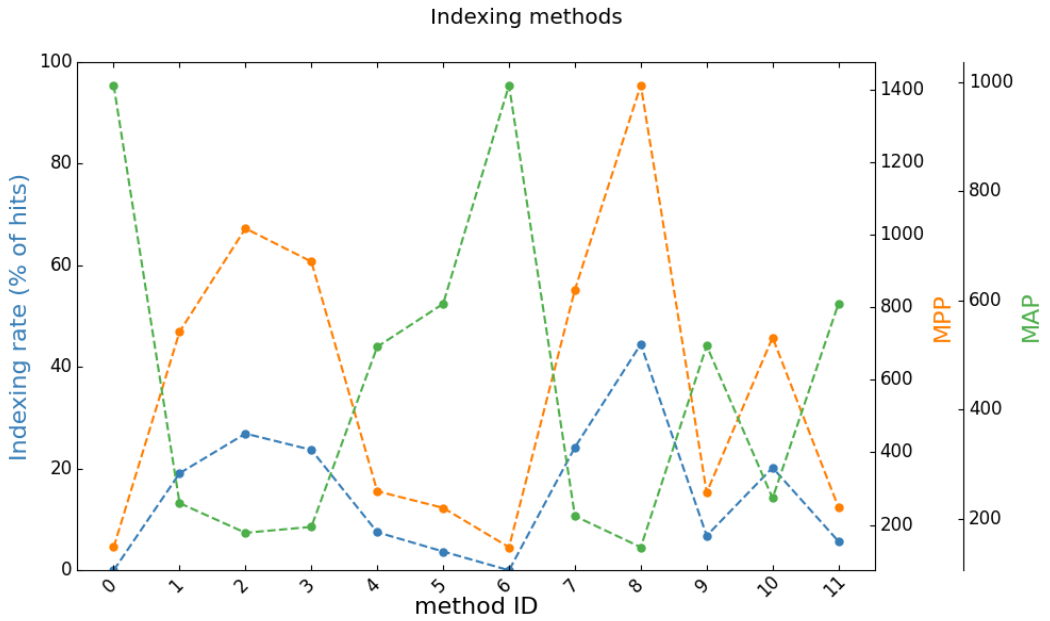


Figure 6.29: Indexing methods comparison for 2910 images of lysozyme cryocooled crystals.

The methods that got better performance, indexing rate >1% of hits, were selected to the next indexing stage. There, the goal is to achieve the maximum of indexed patterns from the whole dataset (2910). We used a descending order, according to the indexing rate performance, for the last indexing and integration process in order to optimize the processing time. The final sequence of chosen indexing methods were:

1 - xds-latt-cell
2- mosflm-latt-cell
3- xgandalf-nolatt-cell
4- mosflm-latt-nocell
5- mosflm-nolatt-cell
6- mosflm-nolatt-nocell
7- dirax-nolatt-nocell
8- xgandalf-nolatt-nocell
9- asdf-nolatt-nocell
10- xds-nolatt-nocell

The final dataset selected, using the unit-cell as reference, has the distribution of figure 6.30, and has 1685 indexed images (64.5% of hits, 57.9% overall).

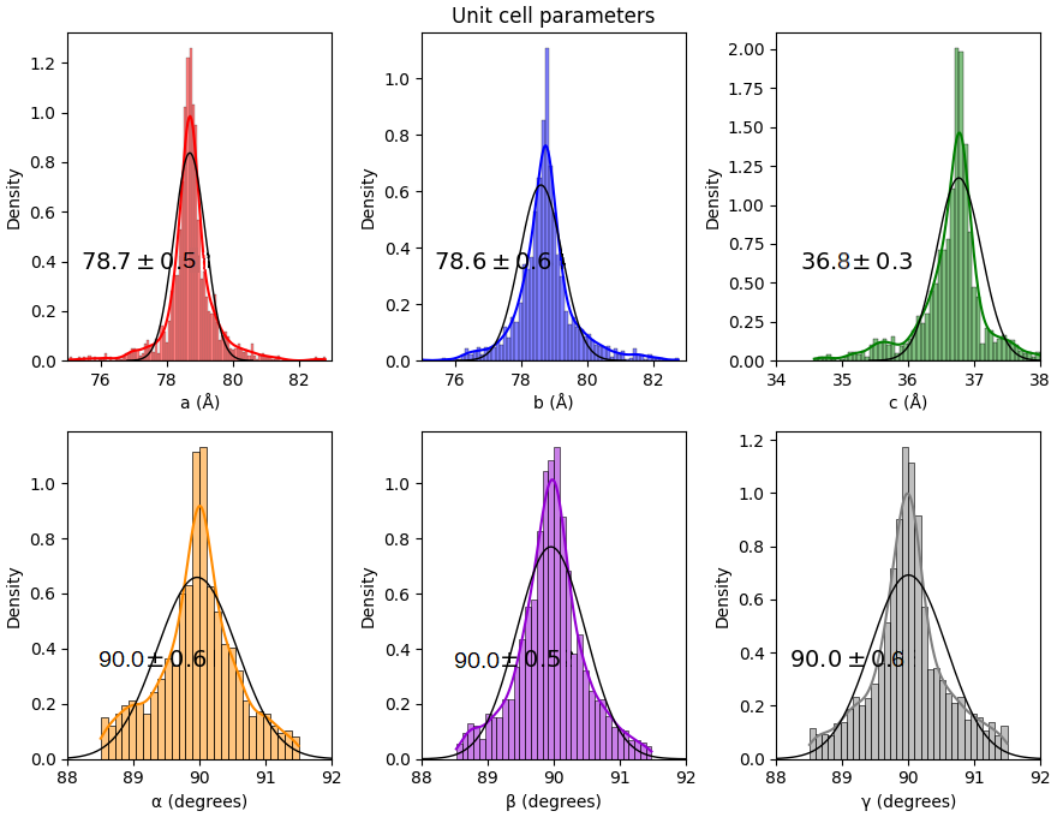


Figure 6.30: Final unit-cell distribution for 2910 diffraction patterns of cryocooled lysozyme, indexed with the sequence of best methods and reference unit-cell check.

Merging, scaling, post refinement and figures of merit calculation

With the indexing methods selected in the last section, integration on default parameters, geometry file corrected and unit-cell fitted in the first part, we merged the subdatasets. Then,

we analyzed the evolution of the figures of merit and data quality (figures 6.31 to 6.35) as more images were included in the final dataset, (16 crystals - 741 images, 32 crystals - 1544 images, 48 - 2297 images, 64 crystals - 2910 images).

Crystals	Images processed	Hits (%)	Indexable (% of hits, % overall)
16	741	587 (79.2)	467 (79.6, 63.0)
32	1544	1330 (86.1)	1008 (75.8, 65.3)
48	2297	2050 (89.2)	1331 (64.9, 57.9)
64	2909	2614 (89.9)	1685 (64.5, 57.9)

Table 6.11: Evolution of indexed images as lysozyme cryocooled crystals were measured and included in the final dataset.

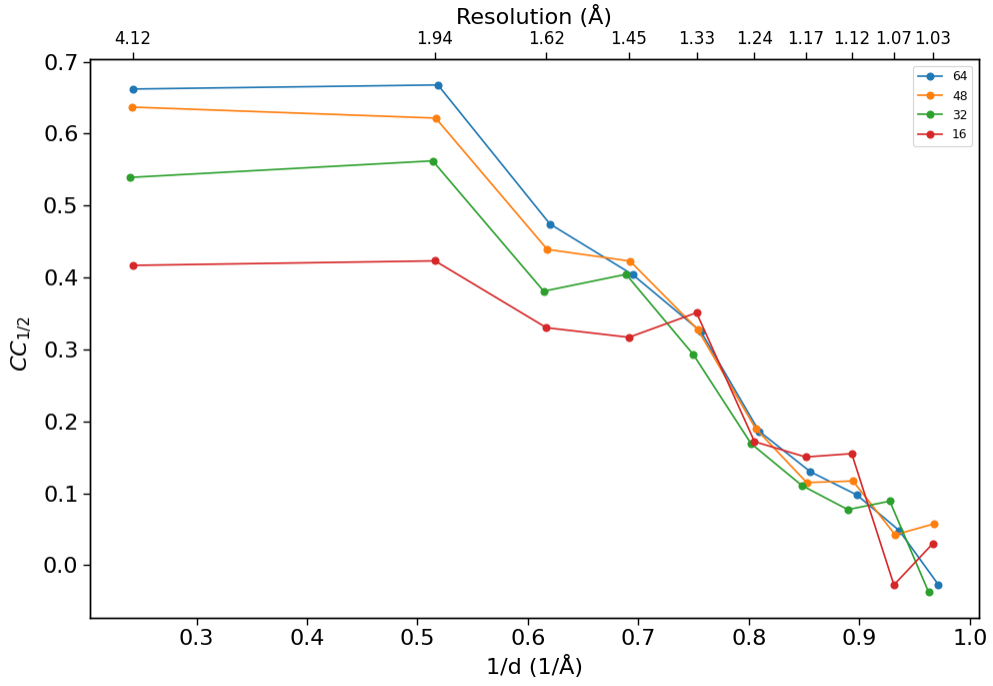


Figure 6.31: $CC_{1/2}$ figure of merit according to the number of lysozyme cryocooled crystals measured in grid-scan mode.

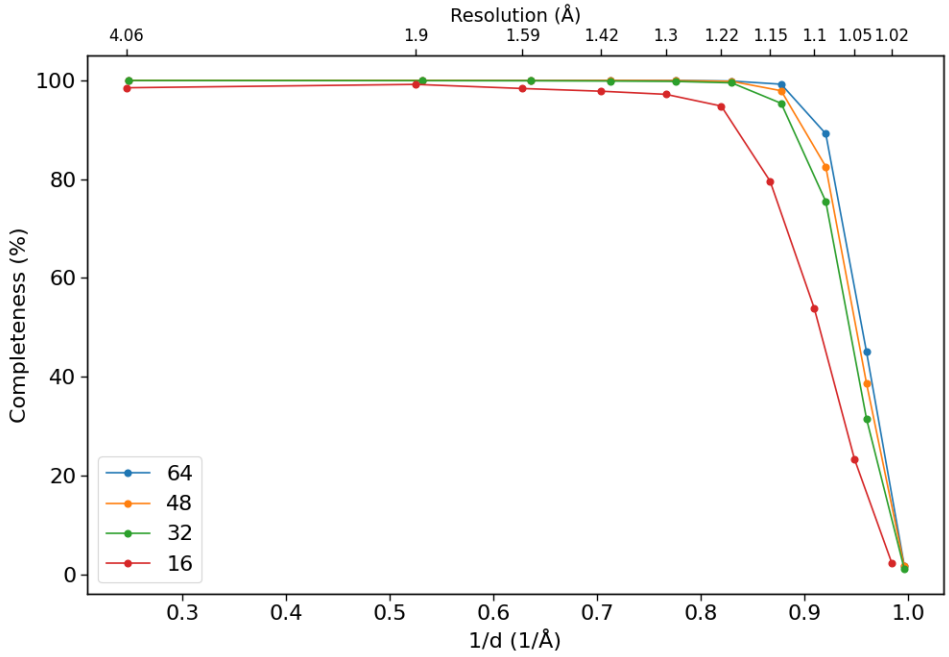


Figure 6.32: Completeness figure of merit according to the number of lysozyme cryocooled crystals measured in grid-scan mode.

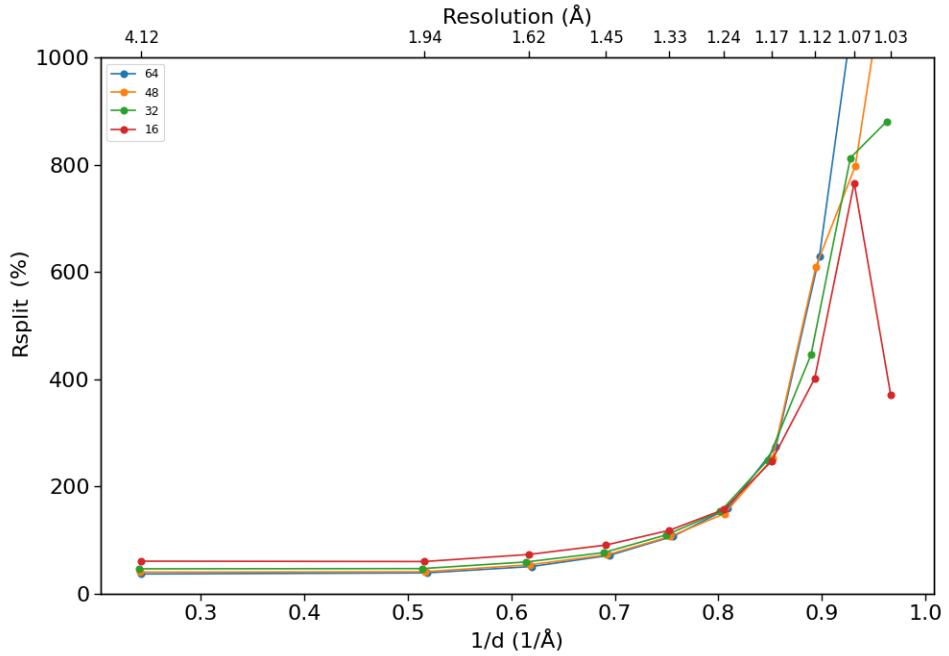


Figure 6.33: R_{split} figure of merit according to the number of lysozyme cryocooled crystals measured in grid-scan mode.

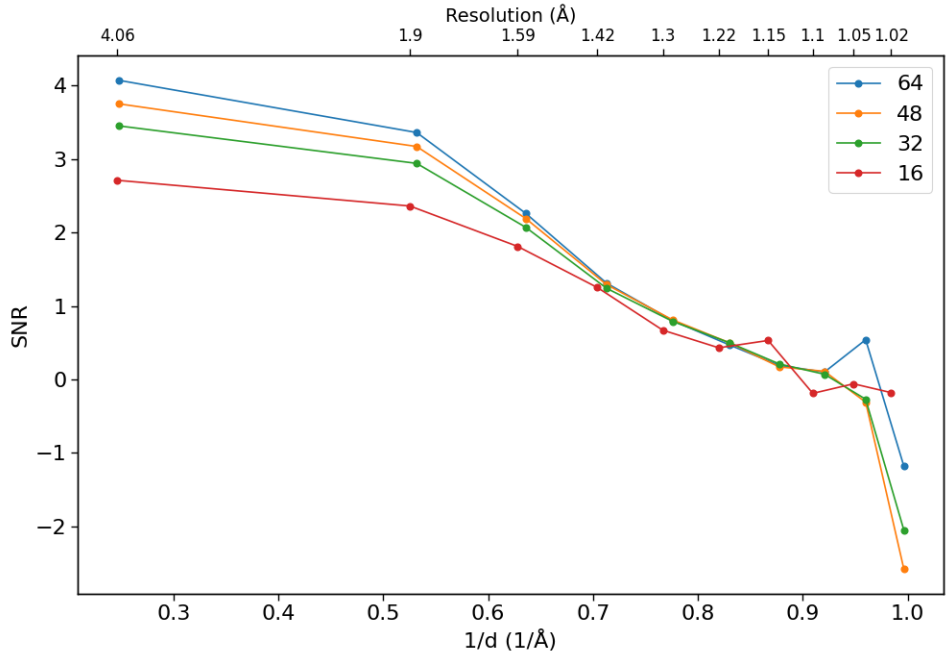


Figure 6.34: SNR figure of merit according to the number of lysozyme cryocooled crystals measured in grid-scan mode.

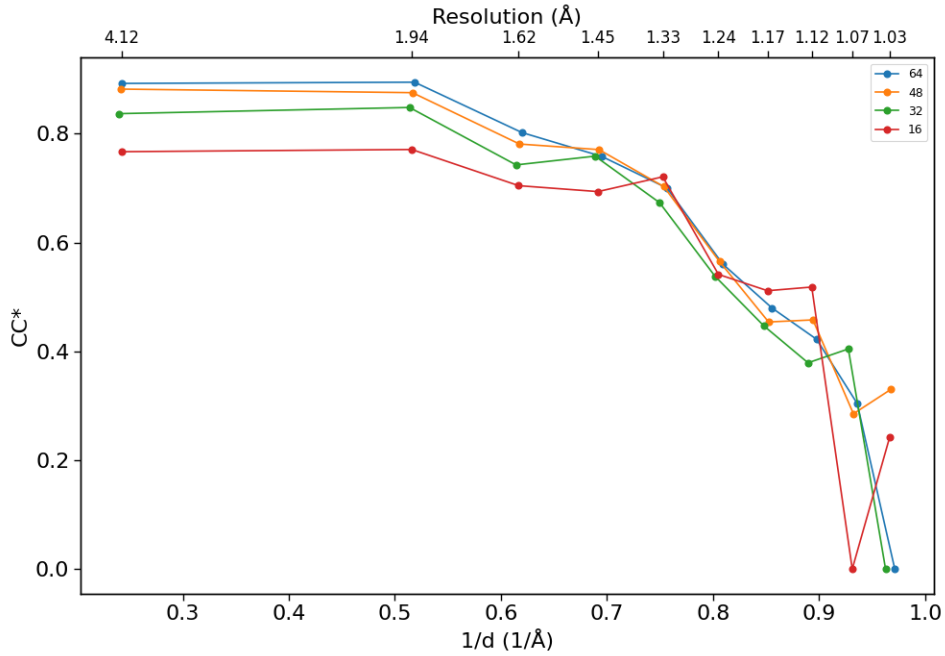


Figure 6.35: CC* figure of merit according to the number of lysozyme cryocooled crystals measured in grid-scan mode.

According to figure 6.35, an approximation for resolution cut should be close to 1.2Å, with CC*>0.5[42]. Beyond this, the signal observed is not reliable. One should increase the number of images correctly indexed in order to obtain a higher resolution data, with good statistics in higher resolution shells.

6.2.3 *ccCluster* data processing

In the regime measured, we could solve around 4º from each crystal as a simple rotation using XDS, and then use the HCA routine to evaluate different combinations of subdatasets and their final data quality [41]. From the 64 crystals measured, we generated 59 integrated reflection lists (`XDS_ASCII.HKL`).

The hierarchical distance was based on the intensity reflections correlation coefficient (cc). Based on the *ccCluster* dendrogram, we set in the automatic pipeline two different thresholds interval: 0.45-0.95 (fig. 6.36) with five threshold points equally spaced (table 6.12), and 0.6-0.8 with 9 threshold points equally spaced.

Threshold	Number of subdatasets in the biggest cluster
0.45	3
0.57	4
0.7	17
0.82	53
0.95	58

Table 6.12: Threshold points tested from 0.45-0.95 and the number of subdatasets included in the respective biggest cluster.

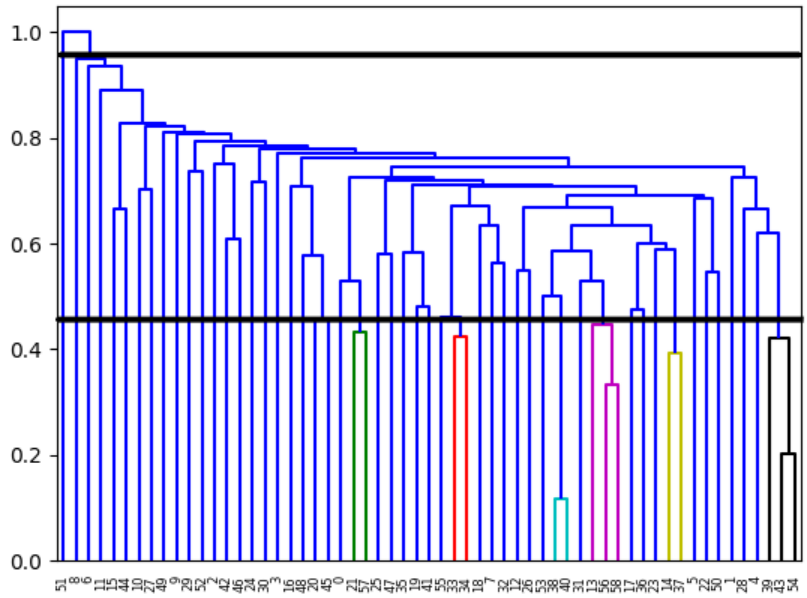


Figure 6.36: Dendrogram of 59 lysozyme cryocooled crystals wedges of around 4° each. In black, it is the interval (0.45-0.95) from which the biggest cluster was merged for 5 different threshold values.

Threshold	Number of subdatasets in the biggest cluster
0.6	7
0.62	7
0.65	12
0.68	14
0.7	17
0.72	31
0.75	37
0.78	42
0.8	49

Table 6.13: Threshold points tested from 0.6-0.8 and the number of subdatasets included in the respective biggest cluster.

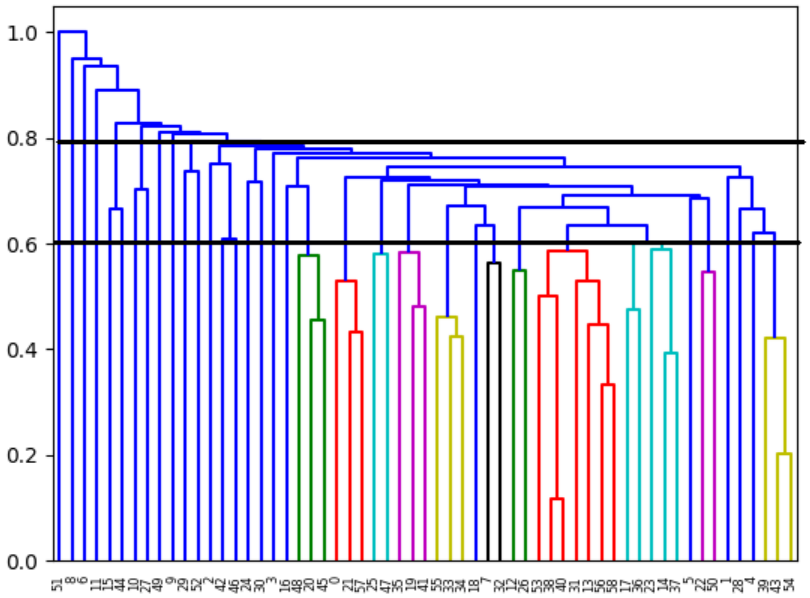


Figure 6.37: Dendrogram of 59 lysozyme cryocooled crystals wedges of around 4° each.In black, it is the interval (0.6-0.8) from which the biggest cluster was merged for 9 different threshold values.

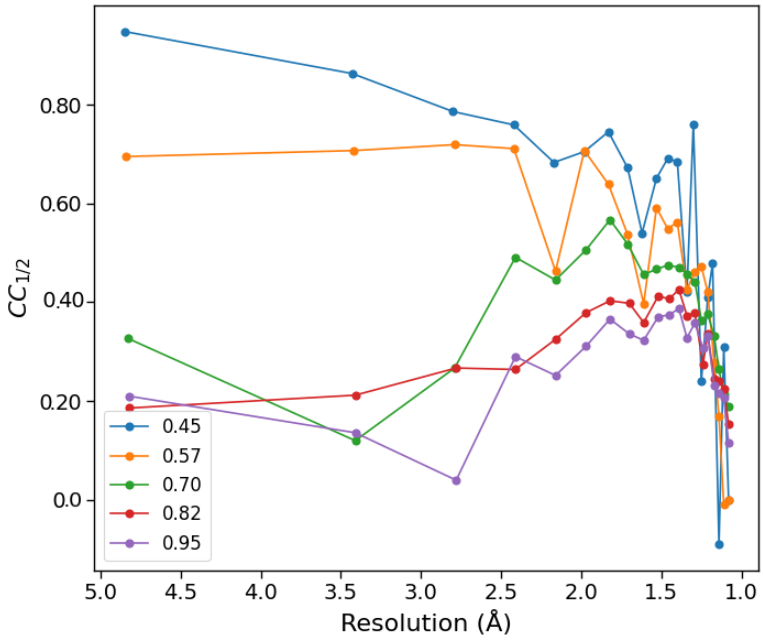


Figure 6.38: Figures of merit for 59 lysozyme cryocooled crystals wedges of around 4°: $CC_{1/2}$ according to the threshold value used to choose the biggest cluster (threshold limits 0.45 to 0.95).

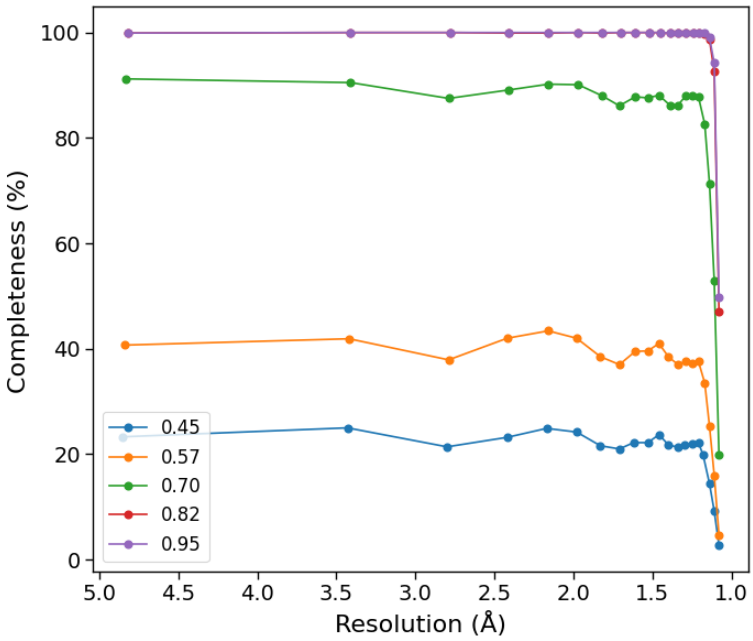


Figure 6.39: Figures of merit for 59 lysozyme cryocooled crystals wedges of around 4°: Completeness according to the threshold value used to choose the biggest cluster (threshold limits 0.45 to 0.95).

The `XSCALE` outputs were plotted comparatively for every cluster merged. Figures 6.38 to 6.40 corresponds to the figures of merit obtained for the first interval. From figure 6.41 to 6.44 corresponds to the same control cards obtained for the second interval.

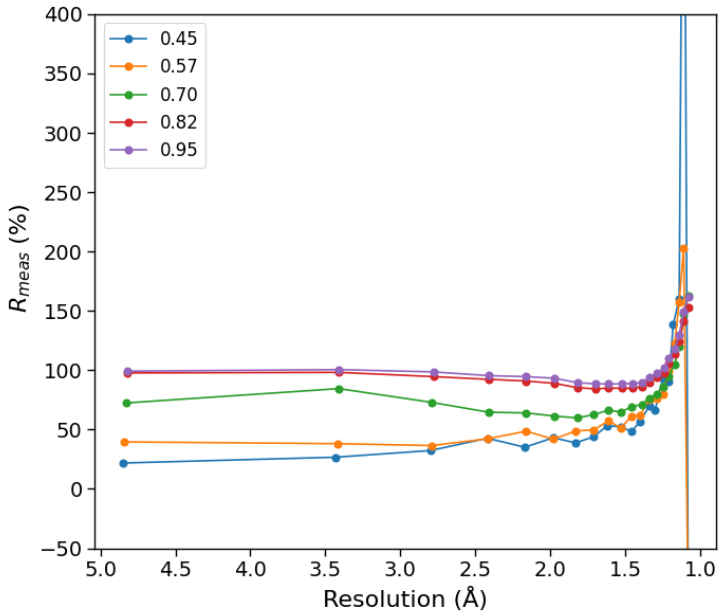


Figure 6.40: Figures of merit for 59 lysozyme cryocooled crystals wedges of around 4°: R_{meas} according to the threshold value used to choose the biggest cluster (threshold limits 0.45 to 0.95).

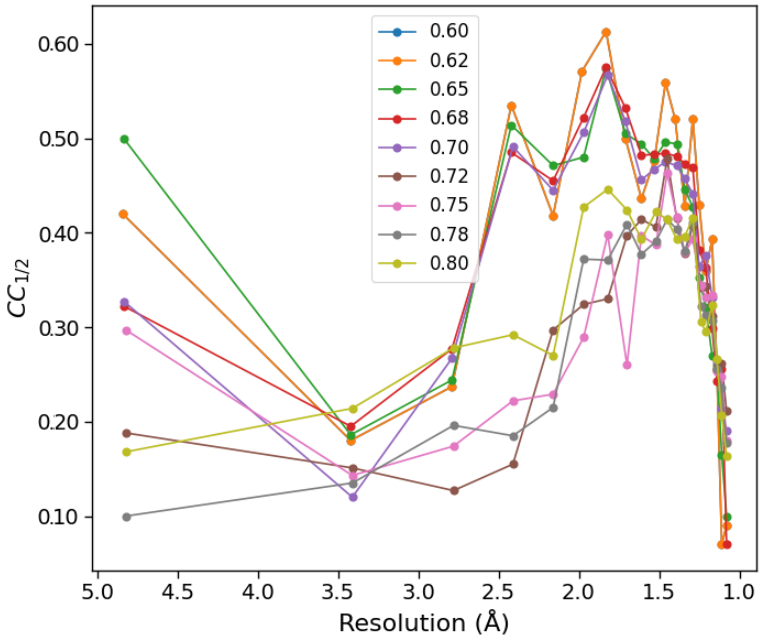


Figure 6.41: Figures of merit for 59 lysozyme cryocooled crystals wedges of around 4°: $CC_{1/2}$ according to the threshold value used to choose the biggest cluster (threshold limits 0.6 to 0.8).

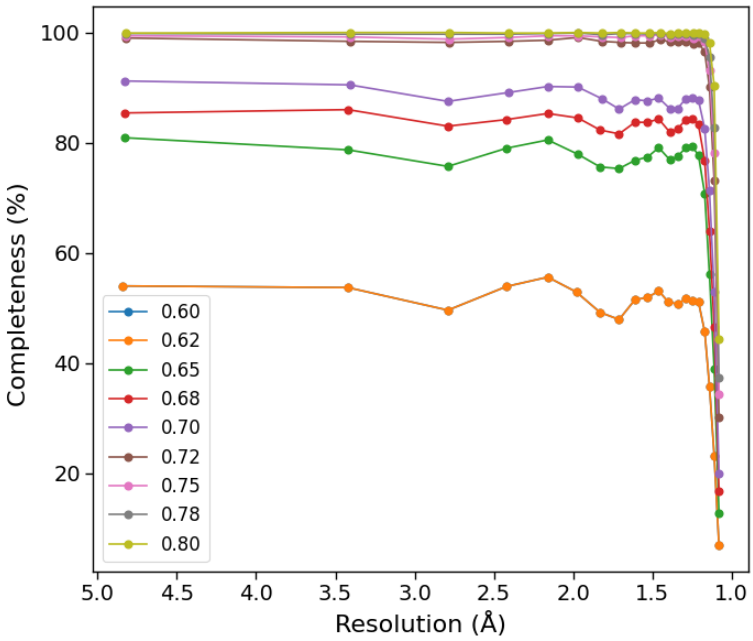


Figure 6.42: Figures of merit for 59 lysozyme cryocooled crystals wedges of around 4°: Completeness according to the threshold value used to choose the biggest cluster (threshold limits 0.6 to 0.8).

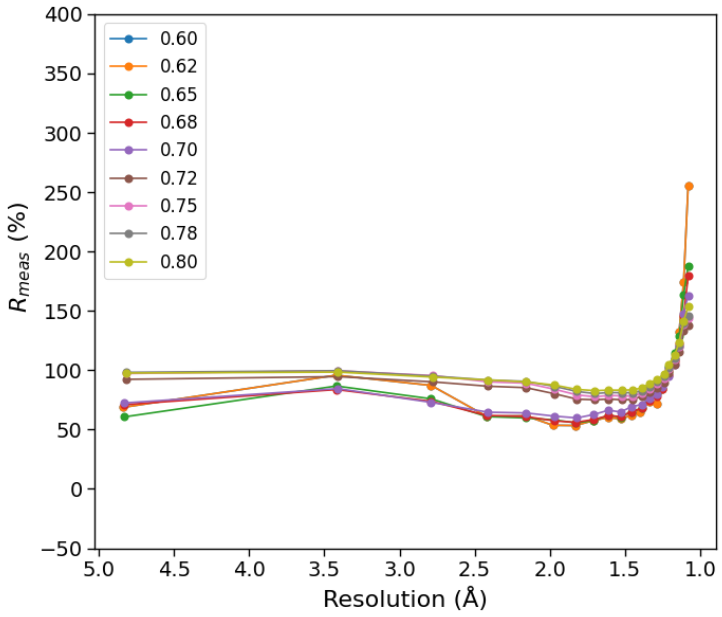


Figure 6.43: Figures of merit for 59 lysozyme cryocooled crystals wedges of around 4°: R_{meas} according to the threshold value used to choose the biggest cluster (threshold limits 0.6 to 0.8).

The *ccCluster* calculated an estimated threshold of 0.75 (magenta in figure 6.44). For this cluster, with its scaled and merged intensity reflections list (`scaled.hkl`), the user is able to solve the structure with the *ccCluster* output files, with their preferred crystallography packages.

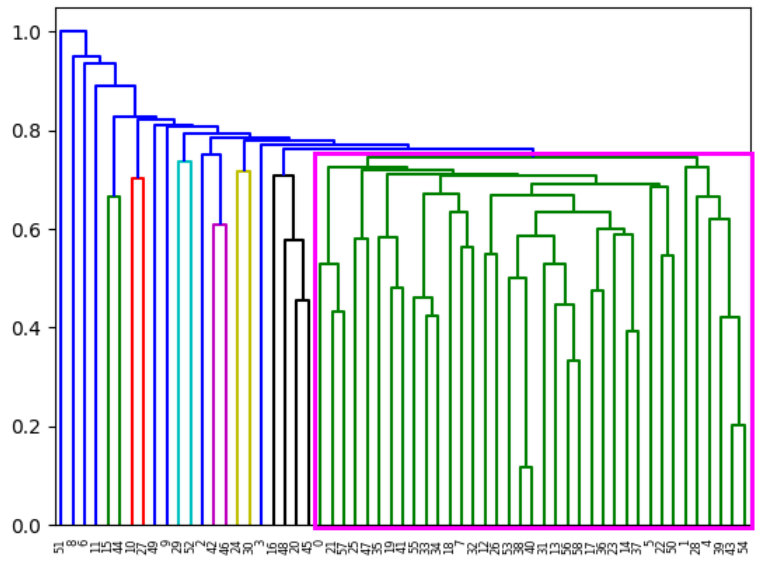


Figure 6.44: Estimated threshold (0.75) for the best cluster (magenta) suggested by *ccCluster* for 59 lysozyme cryocooled crystals subdatasets (wedges of 5°).

6.2.4 Final discussions

From both routines you can see an improvement in data quality as more indexed images (*CrystFEL*) or subdatasets in the cluster (*ccCluster*) are merged. However the figures of merit, for the number of images collected, are still bellow the acceptable for a final dataset. In order to improve the data we should continue collecting crystals and, in parallel, better optimize software parameters (peak search, geometry file and low-level indexing parameters), in order to achieve a better indexing rate.

This was the first grid-scan experiment performed at the Manacá beamline, so there are several software developments to be done in grid-scan data collection, for example. That should be solved with the MXCuBE complete integration with the beamline. Also, the goniometer stepper motors are planned to be soon upgraded for faster ones. Added to Sirius' beam stabilization with future accelerators optimization, it should increase the number of crystals measured per shift for this kind of experiment.

6.3 Grid-scan of lysozyme crystals at room temperature (RT)

6.3.1 Experimental setup

Following the same procedure of section 6.2.1, we measured 20 lysozyme crystals, at RT, with an in-house build sample holder, sealed with Kapton. The chip might accommodate 2 to 4 lysozyme crystals, depending on the crystals size. The same grid-scan script, based on LNLS (SOL) fly-scan script, were used for data collection. The crystals were manually mounted on the goniometer, which implied a time optimization loss compared to the previous experiment. We collected 20 crystals in two shifts (around 16h in total).

6.3.2 *CrystFEL* data processing

Unit-cell parameters determination

Following the same procedure of 6.2.2, initial tests for peak search included the *zaef* method: `-threshold=100 and 10 -min-squared-gradient=5000 -min-snr=4 -peak-radius=4,5,7`, and *peakfinder8*: `-threshold=100 and 10 -min-snr=3 -min-pix-count=2 -max-pix-count=20 -local -bg-radius=6`. The best indexing performance obtained was –peaks=*peakfinder8* `-threshold=10 -min-snr=3 -min-pix-count=2 -max-pix-count=20 -local-bg-radius=6`.

Firstly, the indexing method used was mosflm-latt-nocell with triclinic lattice type and primitive unit-cell. The unit-cell parameters founded (figure 6.45) match with the known numbers for lysozyme (a=79 Å b=79 Å c=37 Å α, β, γ=90°), except for an exchange between the a and c axis. The unit-cell distribution implies a tetragonal lattice type, and the unique axis (required by *CrystFEL* for tetragonal lattices) should be c to overcome the indexing mistake. That is written in the unit-cell file for the correct unit-cell parameters determination.

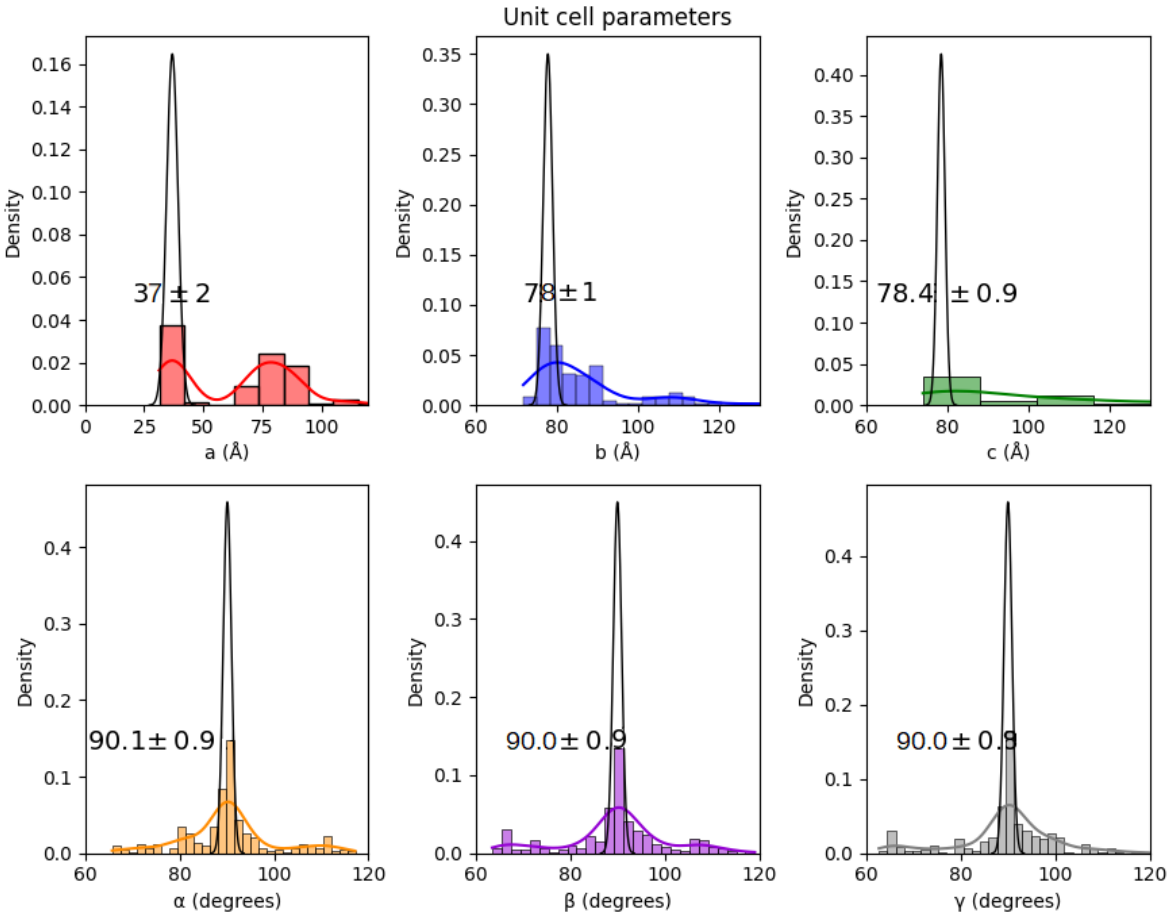


Figure 6.45: Unit-cell parameters distribution 601 images of lysozyme at RT, indexed with mosflm-latt-nocell and the simplest prior lattice information (triclinic lattice type, primitive centring): 600 images processed, 556 hits (92.7%), 393 indexable (70.7% of hits, 65.5% overall).

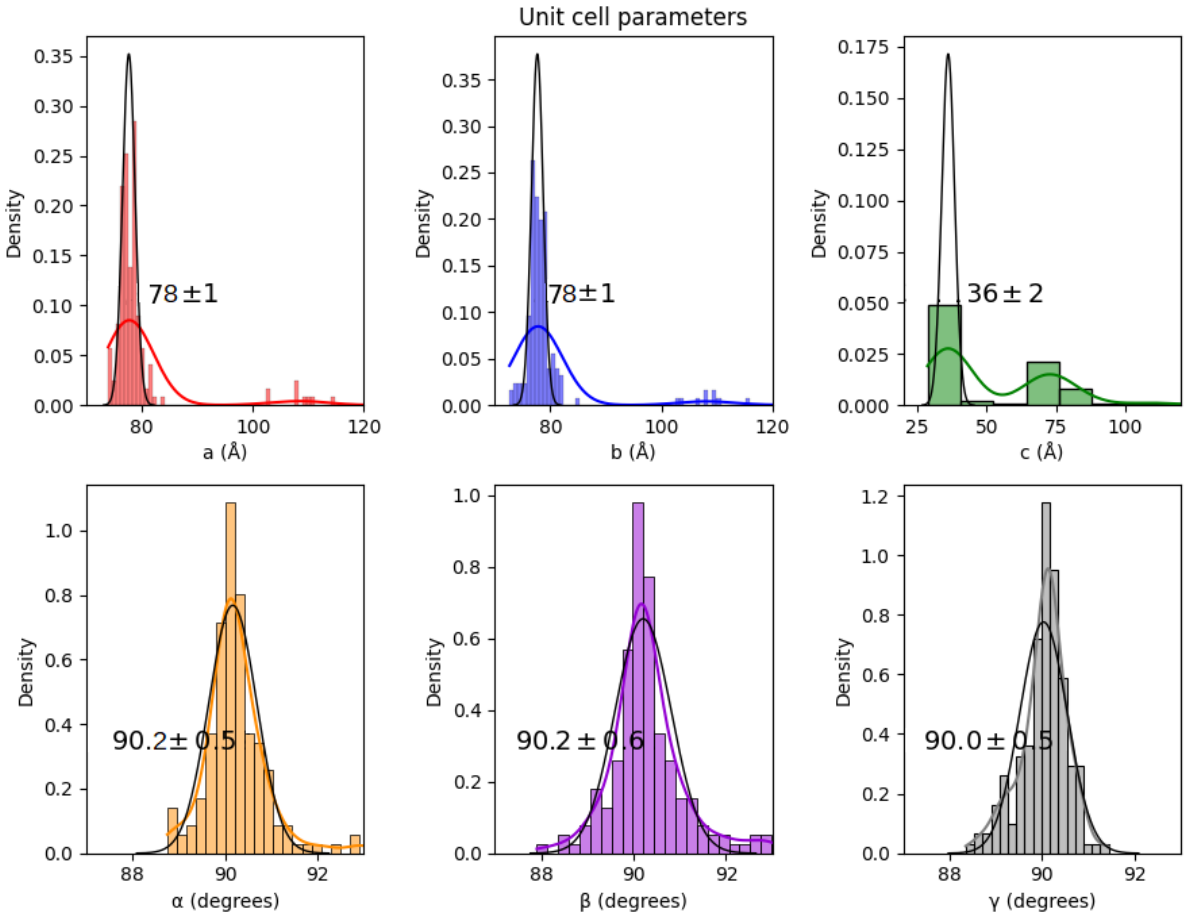


Figure 6.46: Unit-cell parameters distribution 601 images of lysozyme at RT, indexed with mosflm-latt-nocell prior lattice information only (tetragonal lattice type, c unique axis, primitive centring): 601 images processed, 557 hits (92.7%), 169 indexable (30.3% of hits, 28.1% overall)

From figure 6.46, we could determine the unit-cell parameters by fitting the most populated peaks using the `cell_explorer` script build in *CrystFEL*, that can be called from our script too. The unit-cell fitted by cell_explorer was: a= 78±1 b=78±2 c=37±2 α=90.2±0.4 β=90.2±0.5 γ=90.1±0.2. This unit-cell file (`*.cell`) was used on the next steps as reference for the indexing, with a tolerance of 5% in a, b and c, and 1.5% on α, β and γ.

Geometry file corrections

Following the same procedure of 6.1.2, we began changing the camera length 500 μm back and forth the nominal length (0.125m), with steps of 100 μm (table 6.14). From that we observe a better indexing rate for camera length of 0.1255μm (camera displacement +500μm). We tested again on the interval of +400 and +600 μm with steps of 20 μm (table 6.15). The final corrected camera distance founded was 0.1255μm (camera displacement +500μm), with an improvement of the indexing rate from 70.7% to 71.3% of hits.

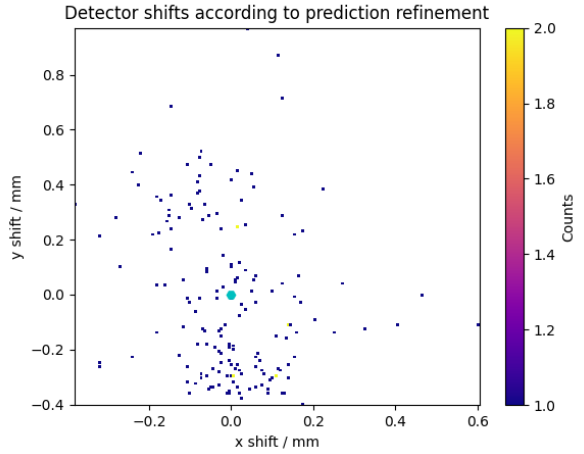
Camera displacement (μm)	Images processed	Hits (%)	Indexable (% of hits, % overall)
-500	600	556 (92.7)	389 (70.0, 64.8)
-400	601	557 (92.7)	382 (68.6, 63.6)
-300	601	557 (92.7)	380 (98.2, 63.2)
-200	601	557 (92.7)	388 (69.7, 64.6)
-100	601	557 (92.7)	384 (68.9, 63.9)
0	601	557 (92.7)	393 (70.7, 65.5)
100	601	557 (92.7)	389 (69.8, 64.7)
200	601	557 (92.7)	384 (68.9, 63.9)
300	601	557 (92.7)	393 (70.6, 65.4)
400	601	557 (92.7)	395 (70.9, 65.7)
500	601	557 (92.7)	395 (71.3, 66.1)

Table 6.14: Detector distance from the sample optimization for 601 images of the lysozyme room temperature dataset, with steps of 100μm.

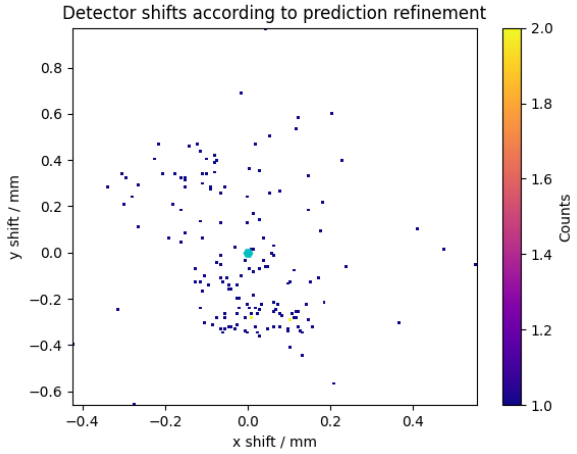
Camera displacement (μm)	Images processed	Hits (%)	Indexable (% of hits, % overall)
400	601	557 (92.7)	395 (70.9, 65.7)
420	601	557 (92.7)	391 (70.2, 65.1)
440	601	557 (92.7)	391 (70.2, 65.1)
460	601	557 (92.7)	392 (70.4, 65.22)
480	601	557 (92.7)	395 (70.9, 65.7)
500	601	557 (92.7)	397 (71.3,66.1)
520	601	557 (92.7)	391 (70.2, 65.1)
540	601	557 (92.7)	387 (69.5, 64.4)
560	601	557 (92.7)	394 (70.7, 65.6)
580	601	557 (92.7)	395 (70.9, 65.7)
600	601	557 (92.7)	389 (69.8, 64.7)

Table 6.15: Detector distance from the sample optimization for 601 images of the lysozyme at room temperature dataset, with steps of 20μm.

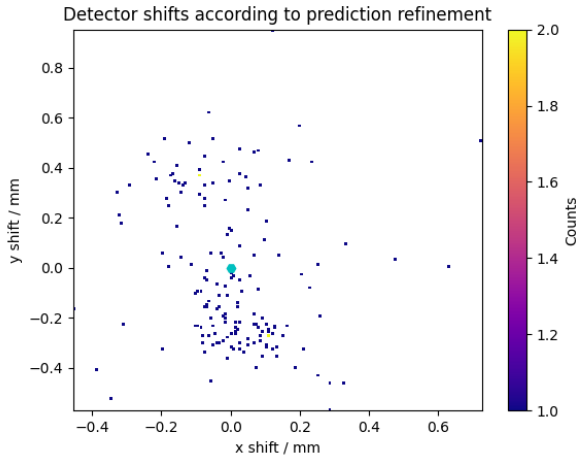
Secondly, we investigated the beam shift predicted by *CrystFEL*, and applied an average correction of the beam center position in the geometry file. We performed three subsequent automatic corrections (fig. 6.47), calling the *CrystFEL* script `detector-shift`. The most visible clusters of detector shifts, selected by clicking, were directly corrected in the geometry file. Here, the total of images was even smaller than in section 6.1.2, so it is also hard to observe a recurrence of shifts in the detector, as in figure 6.3.



(a) One correction run: 171 indexed images (30.8% of hits, 28.5% overall).



(b) Two correction runs: 164 indexed images (29.5% of hits, 27.3% overall).



(c) Three correction runs: 167 indexed images (30.0% of hits, 27.8% overall).

Figure 6.47: Beam position correction using 601 images of lysozyme at room temperature dataset: shift maps after n automatic correction runs using the *detector-shift CrystFEL*'s script. Initial indexed images 169 (30.4% of hits, 28.1% overall

The beam-shift correction that had a better indexing rate (first run of figure 6.47) was selected for the next steps. After all detector corrections the unit-cell fitted by cell_explorer was: a= 78±2 b=78±2 c=37±2 α=90.0±0.3 β=90.2±0.5 γ=90.0±0.3. The final unit-cell file is used as reference in the next stages (Appendix A).

Indexing and integration

The whole dataset (601 images) were indexed, individually, by all 12 methods available in *CrystFEL*, following the order in table 6.16. Figure 6.48 compares the indexing rate (% of hits), mean peaks per pattern (MPP), mean ADU-intensity per peak (MAP) for each indexing algorithm.

ID	Method	Images processed	Hits (%)	Indexable (% of hits, % overall)
0	asdf-nolatt-cell	601	556 (92.5)	0 (0.0, 0.0)
1	mosflm-nolatt-nocell'	601	556 (92.5)	52 (9.4, 8.7)
2	mosflm-latt-cell	601	556 (92.5)	50 (9.0, 8.3)
3	mosflm-latt-nocell	601	556 (92.5)	61 (11.0, 10.1)
4	dirax-nolatt-nocell'	601	556 (92.5)	0 (0.0, 0.0)
5	xds-nolatt-nocell'	601	556 (92.5)	16 (2.9, 2.7)
6	taketwo-latt-cell	601	556 (92.5)	0 (0.0, 0,0)
7	xgandalf-nolatt-cell	601	556 (92.5)	20 (3.6, 3.3)
8	xds-latt-cell	601	556 (92.5)	207 (37.2, 34.4)
9	xgandalf-nolatt-nocell	601	556 (92.5)	0 (0.0, 0.0)
10	mosflm-nolatt-cell	601	556 (92.5)	36 (6.5, 6.0)
11	asdf-nolatt-nocell	601	556 (92.5)	1 (0.2, 0.2)

Table 6.16: *CrystFEL*'s indexing methods evaluation for 601 diffraction patterns of lysozyme at room temperature.)

The methods that got better performance, indexing rate >1% of hits, were selected to the next step, where the objective is to achieve the maximum of indexed patterns. The order in the final indexing and integration step was descending according to the indexing rate. The final sequence of chosen indexing methods were:

1- xds-latt-cell

2- mosflm-latt-nocell

3- mosflm-nolatt-nocell

4- mosflm-latt-cell

5- mosflm-nolatt-cell

6- xgandalf-nolatt-cell

7- xds-nolatt-nocell

The final dataset selected, using the unit-cell as reference, has the distribution of figure 6.49, and has 242 indexed images (43.5% of hits, 40.3% overall).

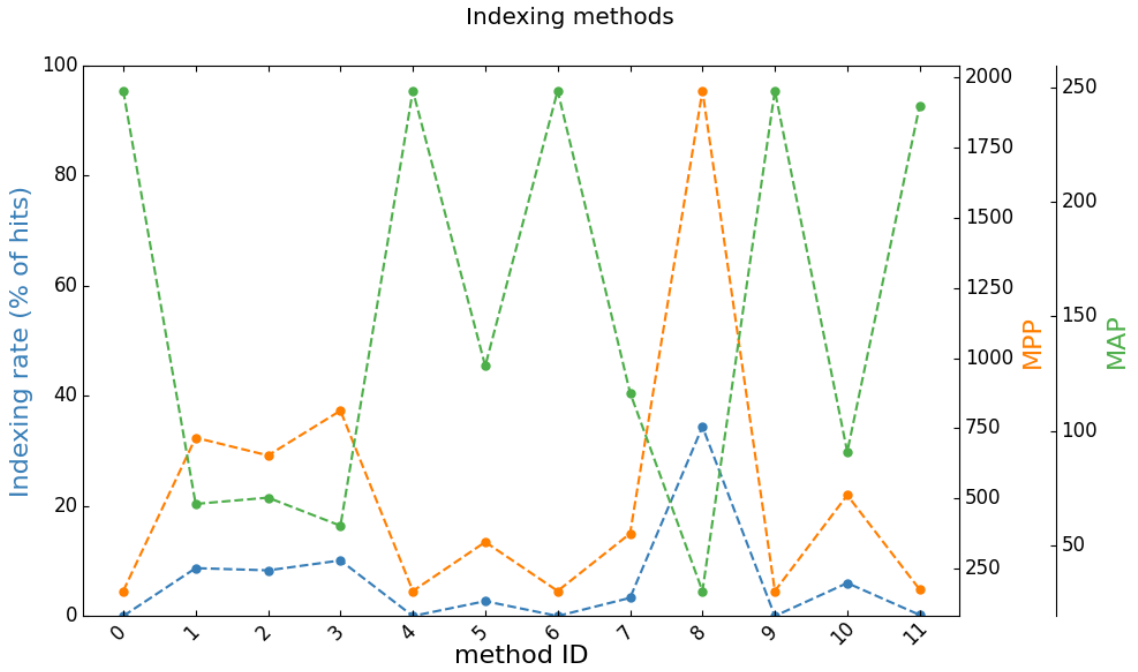


Figure 6.48: Indexing methods comparison for for 2910 images of lysozyme crystals at room temperature.

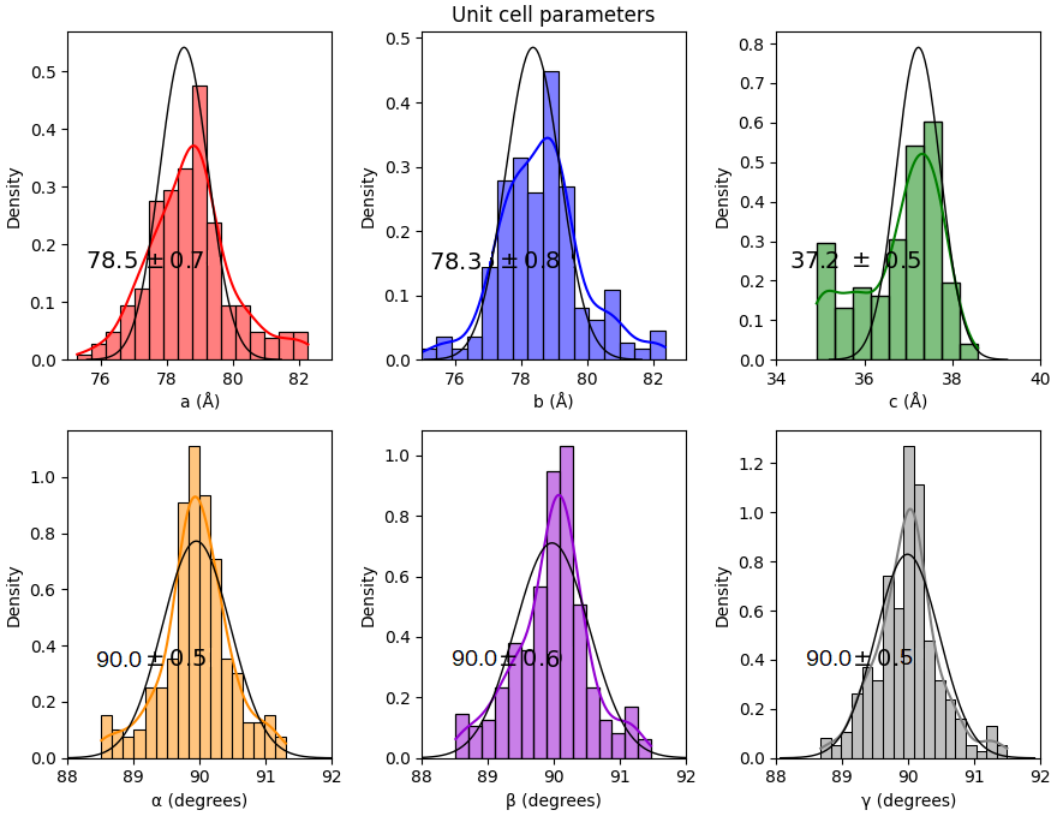


Figure 6.49: Final unit-cell distribution for 601 diffraction patterns of lysozyme at room temperature, indexed with the sequence of best methods and reference unit-cell check.

Merging, scaling, post refinement

With the indexing methods selected in the last section, integration on default parameters, geometry file corrected and unit-cell fitted in the first part, we merged the subdatasets. Then, we analyzed the evolution of the figures of merit and data quality (figures 6.50 to 6.54) as more images were included in the final dataset, (16 crystals - 741 images, 32 crystals - 1544 images, 48 - 2297 images, 64 crystals - 2910 images).

Crystals	Images processed	Hits (%)	Indexable (% of hits, % overall)
9	279	266 (95.3)	111 (41.7, 39.8)
20	601	556 (92.5)	242 (43.5, 40.3)

Table 6.17: Evolution of the number of indexed images as lysozyme crystals at room temperature were measured and included in the final dataset.

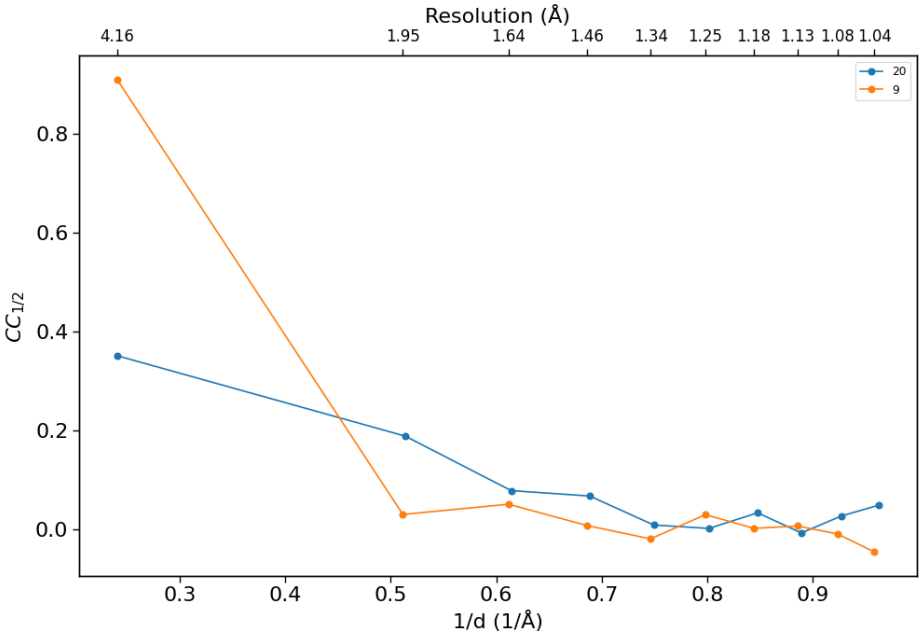


Figure 6.50: $CC_{1/2}$ figure of merit according to the number of lysozyme crystals at room-temperature measured in grid-scan mode.

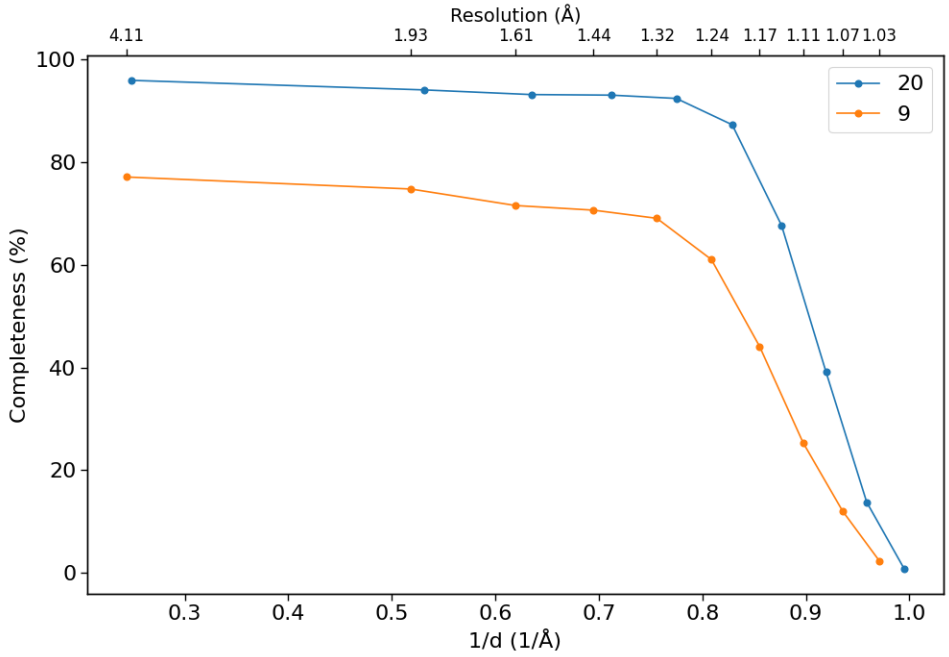


Figure 6.51: Completeness figure of merit according to the number of lysozyme crystals at room-temperature measured in grid-scan mode.

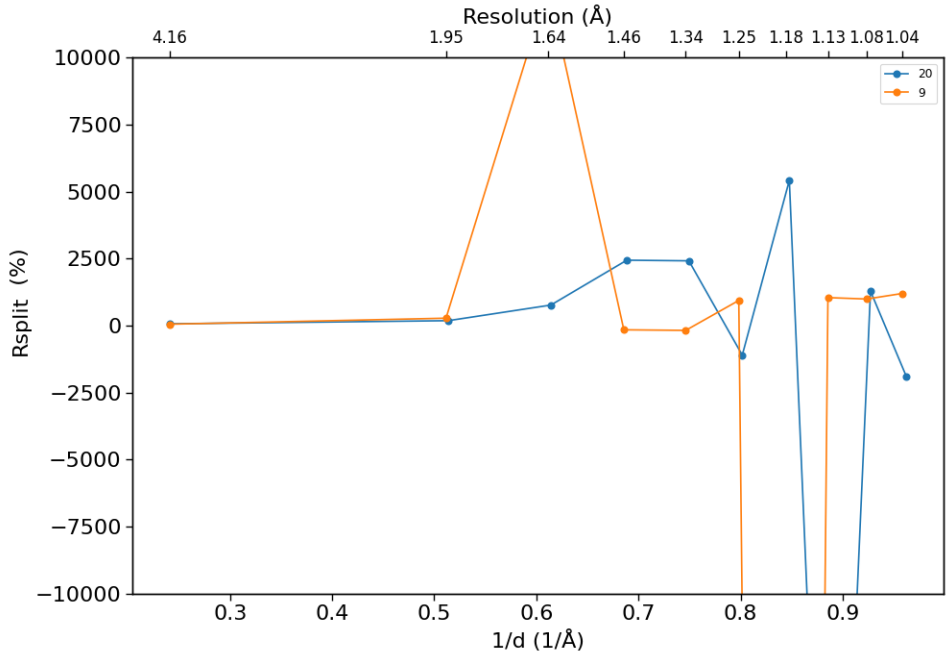


Figure 6.52: R_{split} figure of merit according to the number of lysozyme crystals at room-temperature measured in grid-scan mode.

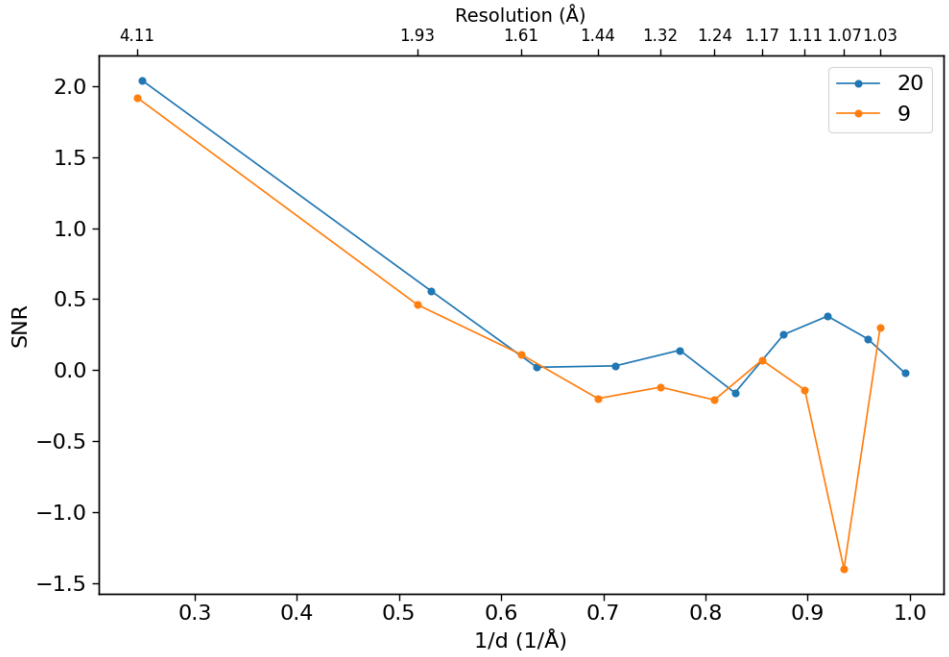


Figure 6.53: SNR figure of merit according to the number of lysozyme crystals at room-temperature measured in grid-scan mode.

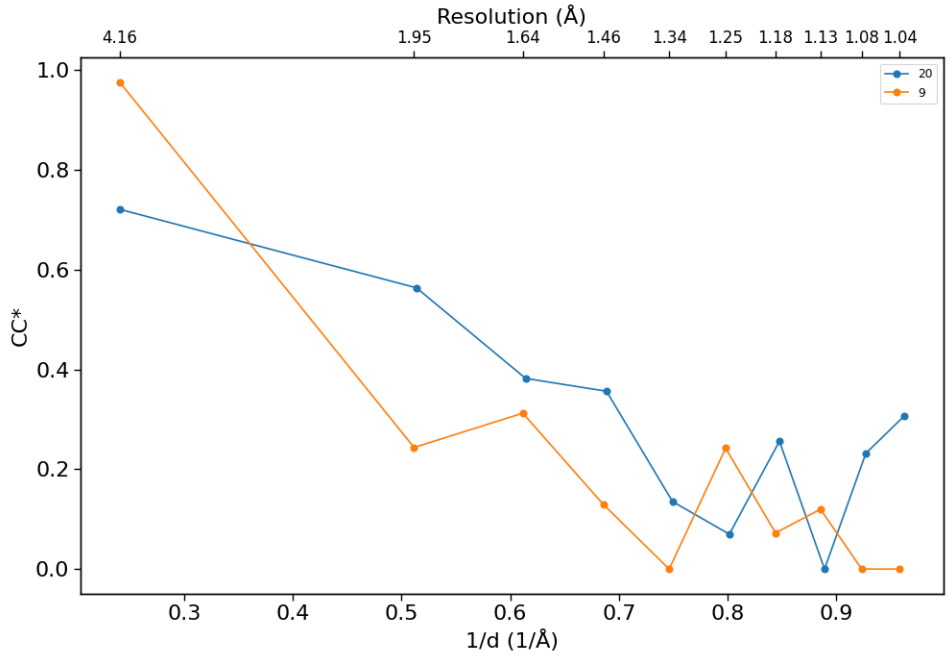


Figure 6.54: CC* figure of merit according to the number of lysozyme crystals at room-temperature measured in grid-scan mode.

6.3.3 *ccCluster* data processing

The lysozyme crystals at RT couldn't be solved as small rotations by conventional crystallography packages (XDS) in general. For those who XDS could solved, they were not reliable as their $CC_{1/2}$ was close to zero for most resolution shells.

6.3.4 Final discussions

The crystals at RT mainly diffract less, which makes it difficult for conventional packages to find common reflections and increase the correlation between reflections. Also, the crystals were, in general, smaller than the cryocooled ones, and more sensitive to radiation. All of those factors may have decreased the number of strong diffraction patterns. Therefore, for the experiment at room temperature, snapshot images routine (*CrystFEL*) demonstrated to be more promising in order to solve the structure and have a good data quality. Our group is already working on the adaption of the automatic sample changer tool to a variety of room-temperature sample holders. After that, data collection at RT should be much improved.

Chapter 7

Summary and outlook

The aim of this dissertation was to develop an automatic data processing pipeline for SX experiments at Manacá. The beamline is, currently, open for users in scientific commissioning mode. In order to test our pipeline, we simulated SX with thousands of images collected from multiple *Ame*GH128 crystals. After, we proposed a grid-scan experiment on lysozyme crystals (cryocooled and at room temperature) to obtain subdatasets of low partiality as a simple test for our script.

The snapshot images (*CrystFEL* routine) demonstrated itself to be versatile and useful for automatic testing, optimizing SX data processing steps. We successfully determined the unit-cell parameters for the tested crystals. Automatic geometric corrections, as shift beam and detector distance to the sample, could be done straightforwardly, having an HPC available. The merit comparison plots between datasets might guide users decisions according to the number of crystals collected. The routine still has to be better optimized for higher indexing rate, which is low compared to same conditions experiments[40]. For that we improve upon the peak search optimization, geometry file refinement, and low-level parameters of the indexing methods, as in *XGANDALF* and *TakeTwo*.

With the HCA routine (ccCluster package) we automatically tested different thresholds and plotted them comparatively for the users. It can alternate between different distances (unit-cell variation, LCV, others) as it is being available at *ccCluster*. It would be interesting to have an option to look at different clusters in the same threshold automatically. Also, we should better evaluate if the estimated threshold, in a simple algorithm, is trustworthy for users to quickly obtain answers for their data.

Furthermore, it is great to include alternative packages on our automatic pipeline (*nXDS*, *cctbx.xfel*, for *CrystFEL* and *BLEND*, *xscale_isocluster*, *xds_nonisomorphism* for *ccCluster*). Thus, users can change between algorithm approaches and compare the variant results. Usually, they don't agree with final answers, since they have different philosophies that complicates their direct comparison. On the other hand, they are also powerful and might give different perspectives for the same problem, which could be interesting to have a more detailed view of the searching target.

Manacá's SX first tests aggregated to our data collection and real time data processing knowledge. The grid-scan experiment with lysozyme still lacks a few more thousands of collected patterns, the ideal regime for SX is around 5000 indexed images. This can be observed in quite low

statistics in final dataset. Nevertheless, it brought huge knowledge for the group concerning not only the data processing pipeline improvement, but also software implementations, hardware and sample delivery developments that will be regarded in the near future for Manacá beamline.

In conclusion, our automatic pipeline for serial crystallography at Manacá is ready and functional. It can be soon available for future Manacá users who want to perform serial crystallography experiments, even though the routine still have room for refining and becoming more user-friendly, requiring less previous programming knowledge.

Codes are available at:
https://github.com/anananacr/HCAmanaca.git (*ccCluster*)
https://github.com/anananacr/SSXmanaca.git (*CrystFEL*).

References

1. White, T. A. *et al.* CrystFEL: a software suite for snapshot serial crystallography. *J. Appl. Cryst.* **45,** 335–341 (2012).
2. Santoni, G. *et al.* Hierarchical clustering for multiple-crystal macromolecular crystallography experiments: the ccCluster program. *J. Appl. Cryst.* **50,** 1844–1851 (2017).
3. Foadi, J. *et al.* Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. *Acta Crystallographica Section D* **69,** 1617–1632 (2013).
4. Jumper, J., Evans, R., Pritzel & et al., A. High Accuracy Protein Structure Prediction Using Deep Learning. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)* (2020).
5. Giacovazzo, C. *Fundamentals of Crystallography* (Oxford University Press, Oxford, 2011).
6. Ibach, H. & Lüth, H. *Solid-State Physics: an introduction to principles of materials science* (2009).
7. Diederichs, K. & Wang, M. *Serial Synchrotron X-Ray Crystallography (SSX)* (2017).
8. Gati, C. *et al.* Serial crystallography on *in vivo* grown microcrystals using synchrotron radiation. *IUCrJ* **1,** 87–94 (2014).
9. Tolstikova, A. *Development of diffraction analysis methods for serial crystallography* Dissertation, University of Hamburg, 2020. Dissertation (University of Hamburg, 2020), 1–153. https://bib-pubdb1.desy.de/record/441104.
10. Wang, J., Brudvig, G. W., Batista, V. S. & Moore, P. B. On the relationship between cumulative correlation coefficients and the quality of crystallographic data sets. *Protein Science* **26,** 2410–2416 (2017).
11. Karplus, P. A. & Diederichs, K. Linking Crystallographic Model and Data Quality. *Science* **336,** 1030–1033 (2012).
12. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Current Opinion in Structural Biology* **34,** 60–68 (2015).
13. Castelvecchi, D. Next-generation X-ray source fires up. *Nature* **525,** 15–16 (2015).
14. Meents, A. *et al.* Pink-beam serial crystallography. *Nature Communications* **8** (Nov. 2017).

15. Chapman, H., Caleman, C. & Timneanu, N. Diffraction before destruction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **369,** 2014 (July 2014).

16. Stellato, F. *et al.* Room-temperature macromolecular serial crystallography using synchrotron radiation. *IUCrJ* **1,** 204–212 (2014).

17. Owen, R. L. *et al.* Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallographica Section D* **73,** 373–378 (2017).

18. Mehrabi, P. *et al.* Time-resolved crystallography reveals allosteric communication aligned with molecular breathing. *Science* **365,** 1167–1170 (2019).

19. Pearson, A. R. & Mehrabi, P. Serial synchrotron crystallography for time-resolved structural biology. *Current Opinion in Structural Biology* **65,** 168–174. ISSN: 0959-440X (2020).

20. DePonte, D. P. *et al.* Gas dynamic virtual nozzle for generation of microscopic droplet streams. **41,** 195505 (2008).

21. Weierstall, U., Spence, J. C. H. & Doak, R. B. Injector for scattering measurements on fully solvated biospecies. *Review of Scientific Instruments* **83,** 035108 (2012).

22. Botha, S. *et al.* Room-temperature serial crystallography at synchrotron X-ray sources using slowly flowing free-standing high-viscosity microstreams. *Acta crystallographica. Section D, Biological crystallography* **71,** 387–97 (2015).

23. White, T. A. *et al.* Serial femtosecond crystallography datasets from G protein-coupled receptors. *Scientific Data* **3,** 160057 (2016).

24. Kovácsová, G. *et al.* Viscous hydrophilic injection matrices for serial crystallography. *IUCrJ* **4,** 400–410 (2017).

25. LNLS-CNPEM. *Sirius: acelerando o futuro da ciência* https://www.lnls.cnpem.br/wp-content/uploads/2016/08/Livro-do-Projeto-Sirius-2014.pdf (2020).

26. LNLS-CNPEM. *Projeto Sirius: a nova fonte d eluz síncrotron brasileira* https://www.lnls.cnpem.br/wp-content/uploads/2016/08/Livro-do-Projeto-Sirius-2014.pdf (2014).

27. Geraldes, R. *et al. The New High Dynamics DCM for Sirius* in *Proc. MEDSI'16* Barcelona, Spain (JACoW Publishing, Geneva, Switzerland, 2017), 141–146. ISBN: 978-3-95450-188-5.

28. Gao, Y. *et al.* High-speed raster-scanning synchrotron serial microcrystallography with a high-precision piezo-scanner. *Journal of Synchrotron Radiation* **25,** 1362–1370 (2018).

29. Zander, U. *et al. MeshAndCollect*: an automated multi-crystal data-collection workflow for synchrotron macromolecular crystallography beamlines. *Acta Crystallographica Section D* **71,** 2328–2343 (2015).

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

31. Nass, K. *et al.* Protein structure determination by single-wavelength anomalous diffraction phasing of X-ray free-electron laser data. *IUCrJ* **3,** 180–191 (2016).

32. Leslie, A. G. W. & Powell, H. R. in, 41–51 (July 2007). ISBN: 978-1-4020-6314-5.

33. Duisenberg, A. J. M. Indexing in single-crystal diffractometry with an obstinate list of reflections. *Journal of Applied Crystallography* **25,** 92–96 (1992).

34. Kabsch, W. *XDS*. *Acta Crystallographica Section D* **66,** 125–132 (2010).

35. White, T. A. *et al.* Recent developments in *CrystFEL*. *Journal of Applied Crystallography* **49,** 680–689 (2016).

36. Gevorkov, Y. *et al.* XGANDALF – extended gradient descent algorithm for lattice finding. *Acta Crystallographica Section A Foundations and Advances* **75,** 694–704 (Sept. 2019).

37. Ginn, H. M. *et al. TakeTwo*: an indexing algorithm suited to still images with known crystal parameters. *Acta Crystallographica Section D* **72,** 956–965 (2016).

38. Beyerlein, K. R. *et al.* FELIX: an algorithm for indexing multiple crystallites in X-ray free-electron laser snapshot diffraction images. *Journal of Applied Crystallography* **50,** 1075–1083 (2017).

39. Santos, C., Costa, P. & Vieira, P. e. a. Structural insights into β-1,3-glucan cleavage by a glycoside hydrolase family. *Nature Chemical Biology* **16,** 1–10 (2020).

40. Halsted, T. P. *et al.* An unprecedented dioxygen species revealed by serial femtosecond rotation crystallography in copper nitrite reductase. *IUCrJ* **5** (2018).

41. Hirata, K., Shinzawa-Itoh, K. & Yano, N. e. a. Determination of damage-free crystal structure of an X-ray-sensitive protein using an XFEL. *Nature methods* **11** (2014).

42. Bücker, R. *et al.* Serial protein crystallography in an electron microscope. *Nature Communications* **11** (2020).

Appendix A

Pilatus 2M geometry file

```
;Pilatus 2M

;set proper photon energy
photon_energy = 12688
adu_per_eV = 0.0001
;set nominal detector distance in meters
clen = 0.125
coffset=0
res = 5814.0  ; 172 micron pixel size

; Define rigid group quadrant for a single panel, asic group and
collections for geoptimiser
rigid_group_q0 = 0
rigid_group_a0 = 0
rigid_group_collection_quadrants = q0
rigid_group_collection_asics = a0


; corner_{x,y} set the position of the corner of the detector (in
pixels)
; relative to the beam

0/min_fs = 0
0/max_fs = 1474
0/min_ss = 0
0/max_ss = 1678
0/corner_x = -736
0/corner_y = -858
0/fs = x
0/ss = y

bad_beamstop/min_x = -736
bad_beamstop/max_x = 22
bad_beamstop/min_y = -66
bad_beamstop/max_y = 24
```

Appendix B

CrystFEL unit-cell files

It follows the final unit-cell used as reference for indexing in data processing of *Ame*GH128, lysozyme cryocooled and lysozyme at room temperature datasets.

B.0.1 *Ame*GH128 unit-cell file

```
CrystFEL unit cell file version 1.0

lattice_type = monoclinic
unique_axis = b
centering = P
a = 37.53 A
b = 78.63 A
c = 46.04 A
al = 90.00 deg
be = 102.11 deg
ga = 90.00 deg
```

B.0.2 Lysozyme cryocooled unit-cell file

```
CrystFEL unit cell file version 1.0

lattice_type = tetragonal
unique_axis = c
centering = P
a = 78.82 A
b = 78.82 A
c = 37.03 A
al = 90.00 deg
be = 90.00 deg
ga = 90.00 deg
```

B.0.3 Lysozyme at room temperature unit-cell file

```
CrystFEL unit cell file version 1.0

lattice_type = tetragonal
unique_axis = c
centering = P
a = 78.27 A
b = 78.27 A
c = 36.73 A
al = 90.00 deg
be = 90.00 deg
ga = 90.00 deg
```

Anexo I

Symmetry Classification for Serial Crystallography Experiments

The symmetry chart that follows, created by Thomas White, is available at CrystFEL page.

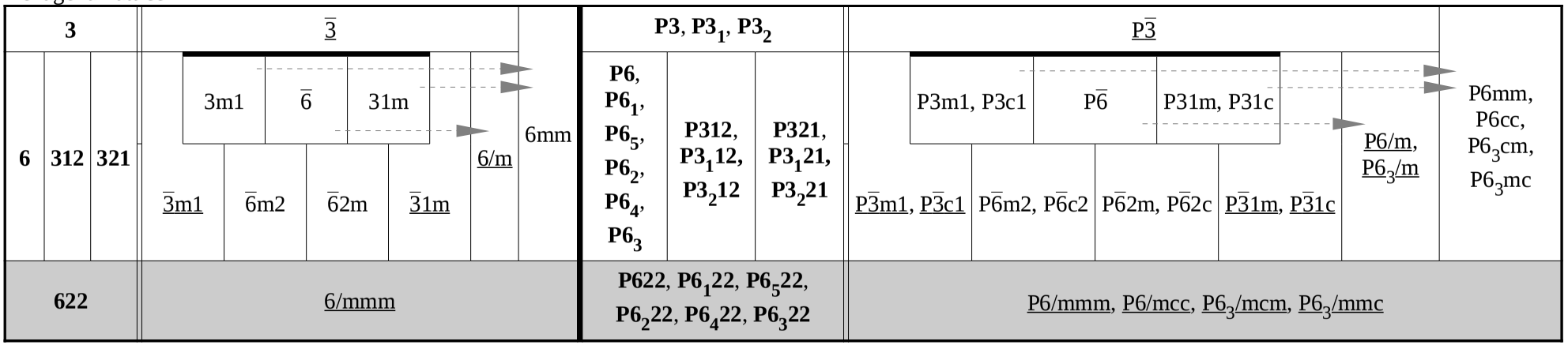
Symmetry Classification for Serial Crystallography Experiments

Groups with white backgrounds are merohedral and will exhibit indexing ambiguities. Chiral groups are shown in bold, centrosymmetric groups are underlined.

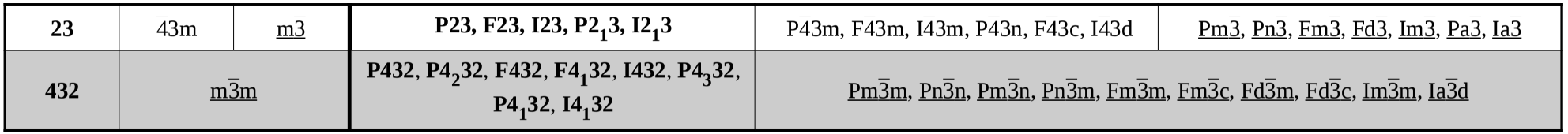
Move downwards or follow grey arrows to find supergroups which can be accessed with only rotation operations. Do not cross vertical or thick black horizontal lines unless following a grey arrow. When you reach a cell with a shaded background, you have found the corresponding "source symmetry". A partial ambiguity resolution could be attempted into any intermediate group you can reach.

Point Groups		**Space Groups**	
Triclinic lattice			
1	$\underline{\bar{1}}$	**P1**	$\underline{P\bar{1}}$
Monoclinic lattice			
	m		Pm, Pc, Cm, Cc
2	$\underline{2/m}$	**P2, P2$_1$, C2**	$\underline{P2/m}$, $\underline{P2_1/m}$, $\underline{C2/m}$, $\underline{P2/c}$, $\underline{P2_1/c}$, $\underline{C2/c}$
Orthorhombic lattice			
	mm2		Pmm2, Pmc2$_1$, Pcc2, Pma2, Pca2$_1$, Pnc2, Pmn2$_1$, Pba2, Pna2$_1$, Pnn2, Cmm2, Cmc2$_1$, Ccc2, Amm2, Aem2, Ama2, Aea2, Fmm2, Fdd2, Imm2, Iba2, Ima2
222	$\underline{mmm}$	**P222, P222$_1$, P2$_1$2$_1$2, P2$_1$2$_1$2$_1$, C222$_1$, C222, F222, I222, I2$_1$2$_1$2$_1$**	$\underline{Pmmm}$, $\underline{Pnnn}$, $\underline{Pccm}$, $\underline{Pban}$, $\underline{Pmma}$, $\underline{Pnna}$, $\underline{Pmna}$, $\underline{Pcca}$, $\underline{Pbam}$, $\underline{Pccn}$, $\underline{Pbcm}$, $\underline{Pnnm}$, $\underline{Pmmn}$, $\underline{Pbcn}$, $\underline{Pbca}$, $\underline{Pnma}$, $\underline{Cmcm}$, $\underline{Cmce}$, $\underline{Cmmm}$, $\underline{Cccm}$, $\underline{Cmme}$, $\underline{Ccce}$, $\underline{Fmmm}$, $\underline{Fddd}$, $\underline{Immm}$, $\underline{Ibam}$, $\underline{Ibca}$, $\underline{Imma}$
Tetragonal lattice			
4	$\bar{4}$	**P4, P4$_1$, P4$_2$, P4$_3$, I4, I4$_1$**	P$\bar{4}$, I$\bar{4}$
	$\bar{4}$2m, $\bar{4}$m2, $\underline{4/m}$		P$\bar{4}$2m, P$\bar{4}$2c, P$\bar{4}$2$_1$m, P$\bar{4}$2$_1$c, I$\bar{4}$2m, I$\bar{4}$2d; P$\bar{4}$m2, P$\bar{4}$c2, P$\bar{4}$b2, P$\bar{4}$n2, I$\bar{4}$m2, I$\bar{4}$c2; $\underline{P4/m}$, $\underline{P4_2/m}$, $\underline{P4/n}$, $\underline{P4_2/n}$, $\underline{I4/m}$, $\underline{I4_1/a}$
	4mm		P4mm, P4bm, P4$_2$cm, P4$_2$nm, P4cc, P4nc, P4$_2$mc, P4$_2$bc, I4mm, I4cm, I4$_1$md, I4$_1$cd
422	$\underline{4/mmm}$	**P422, P42$_1$2, P4$_1$22, P4$_1$2$_1$2, P4$_2$22, P4$_2$2$_1$2, P4$_3$22, P4$_3$2$_1$2, I422, I4$_1$22**	$\underline{P4/mmm}$, $\underline{P4/mcc}$, $\underline{P4/nbm}$, $\underline{P4/nnc}$, $\underline{P4/mbm}$, $\underline{P4/mnc}$, $\underline{P4/nmm}$, $\underline{P4/ncc}$, $\underline{P4_2/mmc}$, $\underline{P4_2/mcm}$, $\underline{P4_2/nbc}$, $\underline{P4_2/nnm}$, $\underline{P4_2/mbc}$, $\underline{P4_2/mnm}$, $\underline{P4_2/nmc}$, $\underline{P4_2/ncm}$, $\underline{I4/mmm}$, $\underline{I4/mcm}$, $\underline{I4_1/amd}$, $\underline{I4_1/acd}$
Rhombohedral lattice			
3	$\underline{\bar{3}}$	**R3 (H3)**	$\underline{R\bar{3}}$ ($\underline{H\bar{3}}$); R3m (H3m), R3c (H3c)
32	$\underline{\bar{3}m}$	**R32 (H32)**	$\underline{R\bar{3}m}$ ($\underline{H\bar{3}m}$), $\underline{R\bar{3}c}$ ($\underline{H\bar{3}c}$)

Hexagonal lattice



Cubic lattice



Laue Classes

$\bar{1}$	$\bar{1}$	
2/m	**2**	m
mmm	**222**	mm2
4/m	**4**	$\bar{4}$
4/mmm	**422**	$\bar{4}$2m, $\bar{4}$m2, 4mm

$\bar{3}$	**3**	
$\bar{3}$m	**32**	3m
$\bar{3}$m1	**321**	3m1
$\bar{3}$1m	**312**	31m

6/m	**6**	$\bar{6}$
6/mmm	**622**	$\bar{6}$m2, $\bar{6}$2m, 6mm
m$\bar{3}$	**23**	
m$\bar{3}$m	**432**	$\bar{4}$3m