Absorption correction for long-wavelength macro-molecular crystallography

Supervised by

Professor Wesley Armour

Yishun Lu

St Anne's College

Department of Engineering Science

University of Oxford

Trinity Term, 2024

A thesis submitted for the degree of

Doctor of Philosophy

Declaration

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Yishun Lu
St Anne's College

Acknowledgements

Firstly, I am profoundly grateful to my wonderful supervisor, Professor **Wesley Armour**, for his unwavering support, patience, and motivation. His immense knowledge and guidance have been invaluable in helping me explore the unknown and develop as a critical thinker. I could not have asked for a better advisor and mentor during my DPhil journey. Beyond academics, his sense of humour and warmth have made daily life more enjoyable, turning even the toughest days into moments of laughter and learning.

I would like to express my sincere thanks to my collaborators **Karel Adámek** from Oxford e-Research Centre, **Armin Wagner**, **Ramona Duman**, and **Tihana Štefanić**, from Beamline I23, Diamond Light Source, and **James Beilsten-Edmands** and **Graeme Winter** from DIALS team, Diamond Light Source. The stimulating discussions we have shared, the teamwork, and the joy of working together over the past three years have been an invaluable part of this journey. I am also grateful to **Yifu Tao**, **Zongyao Zhang**, and **Xianqi Jiang** for their invaluable contributions to the public software testing phase of this project. Their efforts in thoroughly testing the software and providing critical feedback were essential in enhancing the quality and reliability of the results.

I wish to convey my heartfelt appreciation to **Graeme Smith**, **Steven Williams**, **Isobel (Izzy) Griffin Morris**, and **Judy Dendy** for their invaluable administrative support throughout this work. Their assistance has been essential in ensuring the smooth progression of my research. I am also deeply thankful to **Xiaotong Li**, **Jack White**, and **Radostin Stoyanov** for the insightful discussions and constructive feedback they provided, which greatly enriched the development of this project.

Lastly, I extend my deepest gratitude to my mum, **T.H. Pan**, and my dad, **G.H. Lu**, for their warm spiritual support throughout my DPhil life. Their encouragement has been a constant source of strength for me.

Abstract

This thesis presents novel methods for improving long-wavelength macromolecular crystallography (MX), focusing on analytical absorption correction through segmented tomography reconstruction. While the absorption effect is only a minor factor in standard macromolecular crystallography, it can become the largest source of uncertainty for experiments performed at long wavelengths. Current software packages for macromolecular crystallography typically employ empirical models to correct for the effects of absorption, with corrections determined by minimizing the differences in intensities between symmetry-equivalent reflections. These models are well-suited to capture smoothly varying experimental effects.

However, for very long wavelengths, empirical methods become an unreliable approach for modelling strong absorption effects with high fidelity. This issue is particularly acute when data multiplicity is low. This thesis addresses key challenges in absorption correction by introducing AnACor1.0, a ray-tracing analytical absorption correction method that utilizes segmented 3D models of crystal samples, including mounting loops and mother liquor. The accuracy of absorption correction is significantly improved compared to traditional empirical models, reducing systematic errors and enhancing data quality.

To further improve computational efficiency, AnACor2.0 was developed as a GPU-accelerated version of the absorption correction algorithm, leveraging CUDA-based implementation to achieve significant reductions in processing time. Additionally, CPU-based acceleration methods are introduced for cases with limited NVIDIA GPU resources. This acceleration makes the technique more practical for application to large-scale datasets, significantly reducing computational time and broadening its potential use in crystallography.

The combined contributions of this thesis significantly advance the field of long-wavelength macromolecular crystallography by improving the accuracy and efficiency of absorption correction methods. The development of GPU-accelerated algorithms, along with the

integration of machine learning-based segmentation techniques, establishes a strong foundation for automated and rapid long-wavelength crystallographic data analysis. These advancements enable more precise experimental structural determinations in crystallography. Furthermore, when combined with AlphaFold (a protein structure predictive model that earned the Nobel Chemistry Prize 2024), these methods applied on real crystallography experiments provide opportunities for deeper insights into molecular structures, driving progress in structural biology and related fields.

List of publications

The following is a list of articles that were published or written as a result of the research carried out for this thesis.

Articles

- **Lu, Y.**, Duman, R., Beilsten-Edmands, J., Winter, G., Basham, M., Evans, G., Kamps, J.J., Orville, A.M., Kwong, H.S., Beis, K., and Armour, W., 2024. Ray-tracing analytical absorption correction for X-ray crystallography based on tomographic reconstructions. *Journal of Applied Crystallography*, **57**(3), 649–658. `https://doi.org/10.1107/S1600576724002243`.

- **Lu, Y.**, Adámek, K., Stefanic, T., Duman, R., Wagner, A., and Armour, W. (2024). *AnACor2.0: A GPU-accelerated open-source software package for analytical absorption corrections in X-ray crystallography*. **(Accepted in Journal of Applied Crystallography but not in print)**

- **Lu, Y.**, Duman, R., Wagner, A., and Armour, W. (2024). *Enhancing X-Ray Crystallography Segmentation with AI: Leveraging Simulation Data for Improved Accuracy* **(in preparation for Conference on Computer Vision and Pattern Recognition (CVPR))**

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Macromolecular X-ray Crystallography

Since the discovery of atoms, scientists have persistently engaged in the investigation of the microscopic world. The dimensions of atoms and molecules are too small to be discerned by using a conventional microscope. This is due to the fact that their typical size is of the order of angstroms (Å) (10^{-10} m), while visible light in nature has wavelengths ranging from 400 nanometers (nm) to 700 nanometers (nm). The resolution of a microscope is limited by the wavelength of light it uses, as described below in the Abbe diffraction limit for a microscope:

$$d = \frac{\lambda}{2NA} \tag{1.1}$$

where NA is the numerical aperture (maximum of 1.6 in modern optics) [1], d is the resolution and λ is the wavelength. Therefore, a microscope cannot resolve structures smaller than about $\frac{1}{2*16}$ the wavelength of the light used. Thus, with visible light, the smallest detail that can be resolved is about 125 nm, which is much larger than the size of an atom. Hence, an alternative way is needed to view the atomic world.

X-ray crystallography is a powerful technique used to determine the atomic and molecular structures of biological molecules and materials. Crystallographers crystallize the samples to form an ordered solid where the molecules are arranged in a repeating pattern, referred to as crystal lattices. When X-rays are directed at a crystal, if their wavelength is similar to the spacing between the crystal planes, the coherent X-rays constructively interfere to produce diffraction, as described by Bragg's law. By measuring the angles and intensities of these diffracted beams, a three-dimensional model of the electron density within the crystal can

be created. This electron density map is then used to determine the positions of the atoms, allowing scientists to infer the detailed structure of the samples. A synchrotron, a type of particle accelerator, can produce X-rays of various wavelengths to enable high-resolution diffraction studies of molecules of varying sizes. Thus, crystallographers can acquire more comprehensive and clear structures of proteins or materials through the utilization of X-rays with diverse wavelengths.

Macromolecular Crystallography (MX), also known as Protein Crystallography (PX), is well recognized as a very effective and widely utilized method for determining the atomic three-dimensional structure of massive biological molecules. Elucidating the three-dimensional structure of biological molecules provides researchers with critical insights into the underlying mechanisms of various biological processes, which is essential for medication development, understanding disease pathways (like analysis of the structure of Covid-19 [2]), and designing industrial enzymes. For instance, understanding the specific substructure of an enzyme allows researchers to determine how certain chemicals might interact with the enzyme within the human body, potentially influencing its activity and the chemical reactions it catalyzes. This knowledge is vital for designing drugs that can effectively target and modulate enzyme function. Additionally, by analyzing the detailed anatomical structures of bacteria or viruses, scientists can gain a deeper understanding of how these pathogens invade and infect human cells. This information is crucial for developing strategies to prevent or treat infections. For example, identifying the structural components that bacteria or viruses use to attach to and penetrate host cells can lead to the creation of drugs or vaccines that block these critical steps in the infection process, thereby enhancing our ability to combat infectious diseases.

The process of macromolecular crystallography typically involves a series of sequential processes in order to determine the three-dimensional structures of the sample. The procedure involves four main steps:

- **Crystallization:** (in section 2.1.1) The process of forming an ordered, periodic arrangement of biological molecules, such as proteins, from a supersaturated solution.

In a crystal, the basic building block is the asymmetric unit, which contains the unique atomic arrangement that, through symmetry operations defined by the crystal's space group, generates the entire crystal lattice. Protein crystals are typically non-centrosymmetric, meaning they lack a center of symmetry, which is important for phase determination in X-ray crystallography.

- **Data acquisition** (in section 2.1.2) The collection of raw data from crystallographic experiments is based on Bragg diffraction theory [3], typically using X-ray diffraction to gather information about the crystal. The diffracted spots on the detector are also called reflections.

- **Data correction** (in section 2.1.3) The corrections of raw crystallographic data to account for systematic and experimental errors and improve accuracy before building the 3D model.

- **Model building** (in section 2.1.4) The construction of a three-dimensional model of the biological molecule based on the corrected crystallographic data.

1.2 Long-wavelength macromolecular X-ray Crystallography

Single wavelength anomalous diffraction (SAD) is a technique in crystallography that can be utilized in experimental phasing and light-atom localization during model building. The anomalous diffraction effect occurs when the wavelength of incident X-rays is near the K absorption edge of the atom and this incurs a rapid increase and then decreases before and after the K absorption edge, as illustrated in Figure 1.1. These anomalous differences can be used to determine the position of the atom performing the anomalous diffraction. Based on the information about the position of those atoms and the collected diffraction, an accurate 3D structure of the protein can be constructed. However, light atoms, such as sulfur and potassium, which commonly exist in biological molecules, have high absorption K edges. For example, the absorption K edges of sulfur and potassium are 2.472 keV (5.017 Å) and 3.605 keV (3.439 Å), respectively. To better observe anomalous

signals from those atoms, the wavelengths of the incident X-rays should be as close to the absorption K edge as possible.

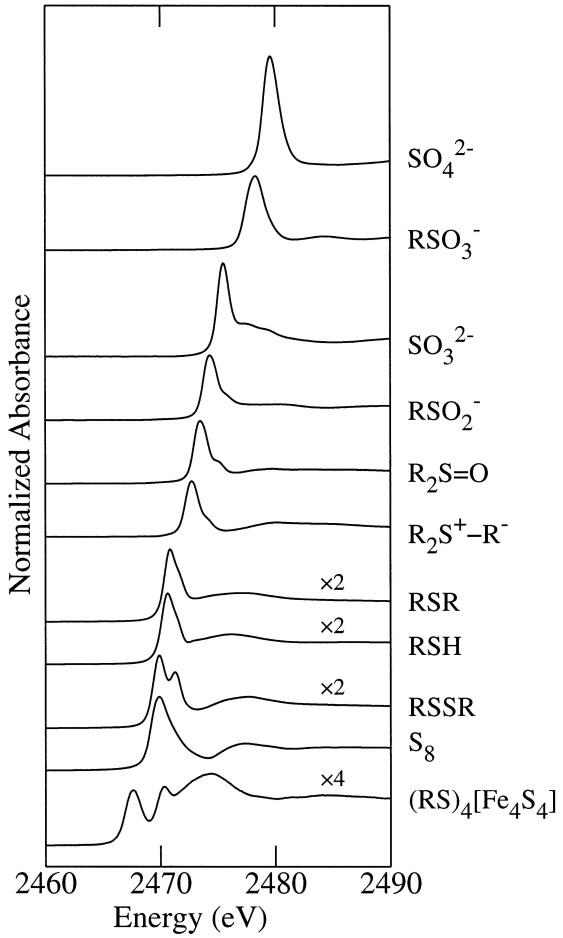


Figure 1.1: Normalized sulfur K-edge X-ray absorption spectra of biologically relevant compounds [4]. It is clear that there are huge absorption differences before and after the absorption K edges, which are about 2.47-2.48 keV.

Long-wavelength crystallography [5, 6] can help significantly enhance the anomalous signal. For example, for sulfur single-anomalous diffraction (S-SAD) experiment, a typical usage case of SAD, the anomalous signal from sulfur atoms at the common X-ray wavelengths ($1\text{Å}-2\text{Å}$) is very low. This is because the absorption edge for sulfur is at about 5Å aforementioned above. This requires very accurate measurement of the anomalous differences [7] to obtain accurate results. However, the precision of the equipment used during the measurement setup can constrain the ability to accurately detect small anomalous differences, which are critical for initial phase determination. Once phases are obtained,

the quality of the structure factors becomes the dominant factor influencing the final 3D structure.

Acquiring small anomalous differences accurately in long-wavelength crystallography is a challenging task. This is because although long-wavelength X-rays enhance anomalous scattering, they also lead to increased absorption effect, requiring careful data collection and processing strategies to minimize data deterioration. This might manifest as reduced resolution and increased ambiguity in determining molecular boundaries. For example, increased absorption effect leads to higher radiation damage, which can disrupt the ordered lattice of delicate crystals by breaking chemical bonds and inducing structural disorder. This damage may cause partial collapse or distortion of the crystalline lattice, which in turn leads to a loss of diffraction quality, reduced resolution, and increased background noise in the collected data. This can be mitigated by cryo-cooling the sample [6], but the absorption correction in data-processing still needs to be investigated.

By using long-wavelength X-ray crystallography, the quality of the anomalous signal is significantly improved, leading to accurate determination of the position of light atoms, such as sulfur, potassium and even sodium atoms. Thereafter, the researchers can gain insights into the placement of specific amino acids, the formation of disulfide bridges and the gating mechanism of cell membranes, which are critical for understanding the protein's structural and functional properties. Also, including light atoms like sulfur in the model improves the accuracy of the electron density map and the overall structural model. The electron density of sulfur can be distinguished more easily compared to other light atoms due to its higher electron count. The positions of sulfur atoms can serve as checkpoints for validating the correctness of the protein model during the refinement process or from AlphaFold [8]. Their anomalous signals provide an independent method to confirm the presence and position of these atoms, ensuring the reliability of the structural model.

1.3 Motivation

1.3.1 Absorption correction in long wavelengths

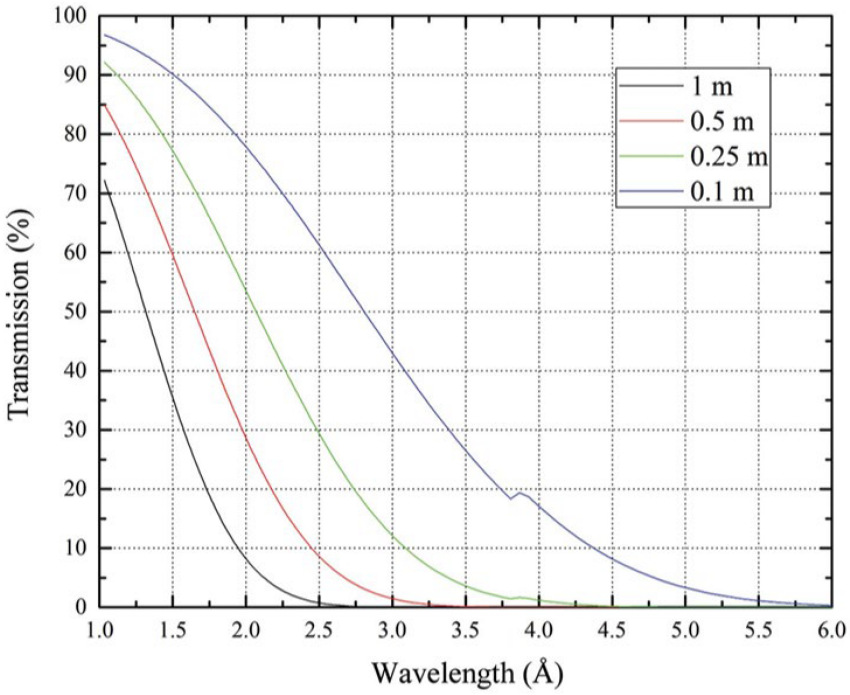


Figure 1.2: X-ray transmission through different lengths of air as a function of wavelength [6]

Utilizing long wavelengths in X-ray crystallography can mitigate issues related to insufficient anomalous signal. However, this approach introduces a significant increase in the absorption effect, which escalates with a higher atomic number or longer wavelength. Specifically, the absorption coefficient increases approximately cubically as a function of wavelength, as illustrated in Figure 1.2. This pronounced absorption effect poses a considerable challenge for accurate absorption correction.

X-ray absorption by the crystal profoundly impacts data collection and compromises data quality when using long-wavelength X-ray beams. Consequently, numerous intensities measured by the detector are significantly reduced during data collection. Without additional processing, data processing software might even disregard some of these low-value intensities. Nonetheless, these intensities remain beneficial and often crucial for three-dimensional structure determination. Researchers typically focus on specific parts of a molecule's structure rather than the overall structure. Therefore, omitting low-magnitude

integrated diffraction intensities can result in the loss of critical molecular substructures, essential for understanding biological compounds or verifying structural models. Furthermore, the anomalous difference may be too weak to observe accurately due to the substantial absorption. In single-wavelength anomalous diffraction (SAD) experiments, the anomalous signal is intertwined with the absorption effect, both contributing to the measured intensities. Inaccurate absorption correction can lead to either an overestimation or underestimation of diffraction spot intensities, complicating the distinction between normal absorption effects and anomalous differences. Consequently, absorption correction is a significant challenge in long-wavelength X-ray crystallography that must be addressed to ensure precise data interpretation.

$$T = \frac{1}{V}\int_z\int_y\int_x e^{-\mu L}\, dx\, dy\, dz \tag{1.2}$$

Figure 1.3: Flat-fielded corrected projection images of Thermolysin. The crystal is glued on the mounting loop by mother liquor.

Equation 1.2 represents the methodology to calculate the absorption factor necessary for performing absorption correction in X-ray crystallography [9]. In this equation, T is the transmission ratio $\frac{I}{I_0}$, V is the volume of the crystal, and μ is the absorption coefficient,

expressed in units of μm^{-1}. Equation 1.2 represents the mean attenuation over every possible X-ray path. The key inputs for this calculation are the absorption coefficients and the volumetric representation of the crystal. For crystals with regular shapes, the integral in Equation 1.2 can be solved either analytically or numerically, and the results are tabulated in the International Tables for Crystallography [10]. However, in practical experimental setups, the sample includes not only the crystal but also the mounting loop (used to hold the crystal in the X-ray beam) and the mother liquor (used to attach the crystal to the loop). These components are depicted in a flat-fielded corrected projection image in Figure 1.3. While the loop and the liquor do not typically produce diffraction patterns, they do absorb a portion of the X-ray beam's energy, which consequently reduces the intensities of the reflections. As indicated in Table 1.1, the absorption coefficients of the loop and mother liquor in the Thermolysin sample, measuring $230 \times 70 \times 70$ μm^3, are non-negligible and must be considered in the absorption correction. However, the absorption factors provided in the International Tables, which are derived based on the crystal's shape, do not account for the absorption effects of the loop and the mother liquor. This oversight can lead to inaccuracies in the absorption correction if these additional absorptive elements are ignored.

Thermolysin Crystal Size (μm^3)	Absorption Coefficient (μm^{-1})		
	Crystal	Liquor	Loop
$230 \times 70 \times 70$	0.01312	0.01583	0.01171

Table 1.1: Absorption coefficients for Thermolysin sample at 3.5 keV (3.53 Å)

The current cutting-edge technique for implementing absorption corrections, known as spherical harmonics correction, exhibits diminished effectiveness as the wavelength of the experiment increases, specifically at long wavelengths. Therefore, an accurate absorption correction method for long-wavelength crystallography is needed.

1.3.2 Research statement

It is more feasible to conduct Single wavelength anomalous diffraction (SAD) studies and achieve accurate localization of light atoms at longer wavelengths. The motivation for this research stems from a collaborative partnership with Diamond Light Source (DLS) and scientists conducting tests on the I23 Beamline [11], specifically focusing on long wavelengths. This study introduces the use of analytical absorption correction via a ray-tracing method, demonstrating its efficacy in long-wavelength experiments.

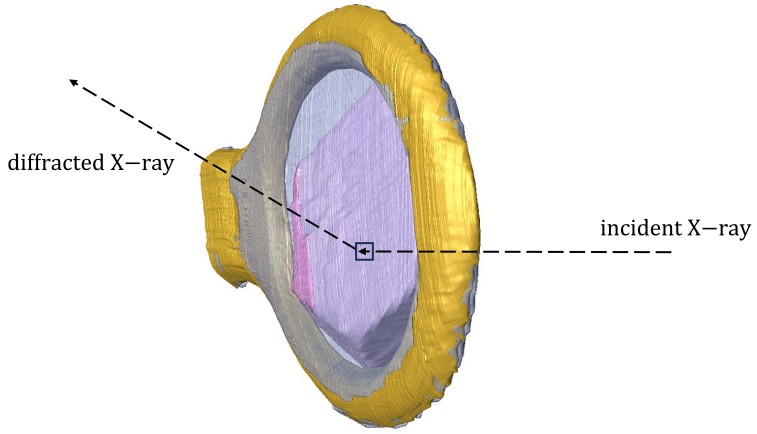


Figure 1.4: Illustration of the ray-tracing method on a 3D tomographic reconstruction model. In the ray-tracing method, the incident X-rays are assumed to be separated into many independent rays.

Accurate analytical absorption corrections require not only the absorption coefficients but also an accurate 3D model of the sample, encompassing the crystal, mounting loop, and mother liquor. In this study, segmented X-ray tomography reconstruction is employed to extract a high-resolution 3D model of the sample. It is important to note that a high-resolution 3D model often consists of a substantial number of crystal voxels, often reaching tens of millions. To perform ray-tracing calculations, it is necessary to identify and analyze thousands of voxels along both the incident and diffracted X-ray paths, as illustrated in Figure 1.4. Consequently, the computational cost of determining absorption correction factors for samples in protein crystallography is significant. A full analytical absorption correction can take more than one week to complete, whereas the other steps of data

processing, such as indexing, integration, and building 3D density map, typically require minutes or hours. This stark imbalance severely limits the overall efficiency and timely analysis of crystallographic data.

Previously, determining absorption coefficients and the segmentation of the tomography reconstruction were completed manually. To automate these processes, machine learning and deep learning techniques are employed, facilitating the automatic determination of absorption coefficients and the segmentation and refinement of segmentation models. All these advancements for achieving absorption correction in long-wavelength crystallography are integrated into a comprehensive and open-source software called AnACor.

1.4 Research contribution of this thesis

1.4.1 My Contributions

1. The effectiveness of analytical absorption correction for X-ray crystallography at long wavelengths has been examined. The findings have been published in the Journal of Applied Crystallography (ranking 1/33 in the Q1 quartile of the crystallography category, in Journal Citation Reports (JCR) in both 2022 and 2023).

2. The computational speed in previous publications was sub-optimal. Acceleration techniques have been applied and a CUDA version with a Python API to parallelize the ray-tracing algorithm has been developed. The computational efficiency has been significantly improved and optimized. A related paper *AnACor2.0: A GPU-accelerated open-source software package for analytical absorption corrections in X-ray crystallography* has been accepted by the Journal of Applied Crystallography.

3. Manual calculation of absorption coefficients can introduce errors. Automated determination methods with statistical tests have been implemented to reduce human effort and increase accuracy. Furthermore, manual segmentation is labour-intensive and can take up to a day to complete. A segmentation model with an inference time of about 2 minutes, trained on a synthetic simulation dataset, has been provided to

address this issue. A research paper *CrystalSeg: Automating Synchrotron Tomographic Reconstruction Segmentation for Crystallography with Physically Guided Simulations* was reviewed and rejected by Computer Vision and Pattern Recognition Conference (CVPR), will submit the manuscript to Nature communications.

1.4.2 Contributions by collaborators

- Crystal sample preparations mentioned in the thesis were performed by collaborators at Beamline I23, Diamond Light Source.
- Diffraction and tomography data collection for Ompk36, Cld, Thaumatin, Thermolysin, and Insulin samples mentioned in the thesis were performed by collaborators at Beamline I23, Diamond Light Source.
- Manual segmentation of tomography reconstructions for Ompk36, Cld, Thaumatin, Thermolysin, and Insulin samples mentioned in the thesis was also performed by collaborators at Beamline I23, Diamond Light Source.

Chapter 2

Background and literature review

2.1 Steps in macromolecular X-ray crystallography

2.1.1 Crystallization

Crystallography requires crystals, as the name suggests. Before performing a macromolecular X-ray crystallography experiment, chemists must crystallize the protein molecules to ensure significant X-ray diffraction and successful data collection on the detector. This process involves carefully preparing a solution containing the target molecules and allowing them to crystallize by gradually adjusting conditions such as temperature, pH, and concentration, as illustrated in Figure 2.1.

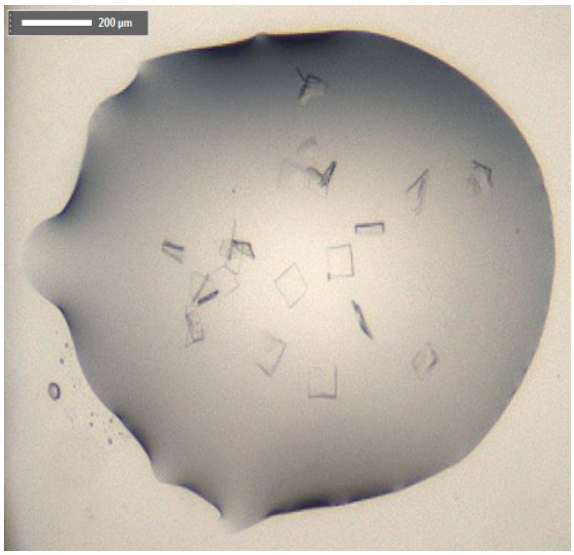


Figure 2.1: Example image of crystals of SARS-CoV-2 main protease (CoVID-19) in mother liquor [2]

In crystallization, biological molecules are arranged into an ordered and periodic three-dimensional array. The fundamental building block is the asymmetric unit, which contains one or more molecules depending on the symmetry and packing of the crystal. Symmetry

operations defined by the crystal's space group replicate the asymmetric unit throughout the unit cell. This orderly three-dimensional array of molecules facilitates high-quality data collection. An example of a crystal lattice and its unit cells is shown in Figure 2.2.

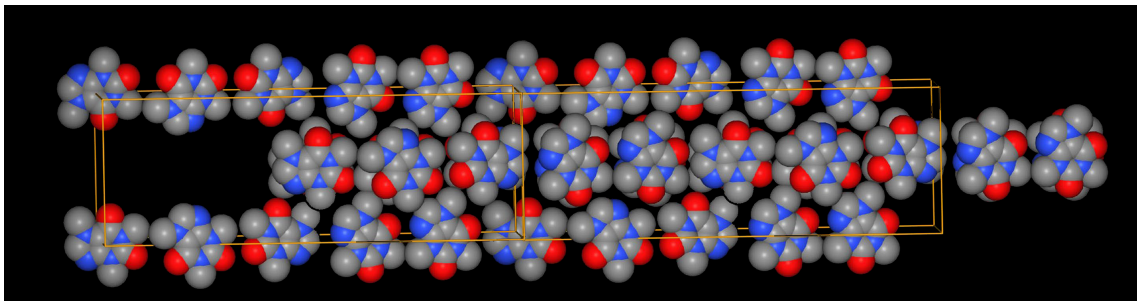


Figure 2.2: Lattice and unit cell representation with caffeine molecules ($C_8H_{10}N_4O_2$). Atoms except hydrogen are shown in a space-filling model and colored as follows: carbon (grey), nitrogen (blue), and oxygen (red). The unit cell is outlined in orange. Unit cell parameters are approximately $a = 43.04\,\text{Å},\ b = 15.07\,\text{Å},\ c = 6.95\,\text{Å},\ \beta = 99°$ [12].

Mosaicity is a metric that refers to the degree of misalignment between the crystal's plane orientations. Lower mosaicity produces clearer diffraction patterns characterized by sharp, well-defined Bragg spots with high signal-to-noise ratios, minimal background scattering, and reduced spot overlap or streaking. This facilitates more accurate spot integration, enhances the precision of intensity measurements, and enables higher-resolution structural determination. A good crystal with low mosaicity can produce high-quality data, but proteins are usually challenging to crystallize perfectly. Especially for macromolecules that are insoluble in many solvents, growing a large crystal with low mosaicity is difficult. Protein crystals are typically grown through a slow precipitation process in a solvent. In a saturated protein solution, specific conditions cause the protein molecules to precipitate. For example, a process called salting out involves adding inorganic ionic compounds (salts) to the protein solution to induce controlled nucleation and promote crystal formation. Initially, precipitation may produce many small crystals or polycrystalline aggregates. However, by carefully optimizing conditions such as salt concentration, temperature, and evaporation rate, it is possible to favor the growth of a single, well-ordered crystal suitable for protein crystallography experiments. Some organic solvents are used during crystallization because

they interact with hydrophobic portions of proteins, aiding in precipitation. Additionally, super-cooling the protein solution is a common approach to crystallization.

However, protein crystallization is a complex problem. The formation of a high-quality and low-mosaicity crystal versus an amorphous solid depends on many factors, such as temperature, air pressure, and intermolecular forces. This report only scratches the surface of protein crystallization, a complex issue requiring numerous careful experiments.

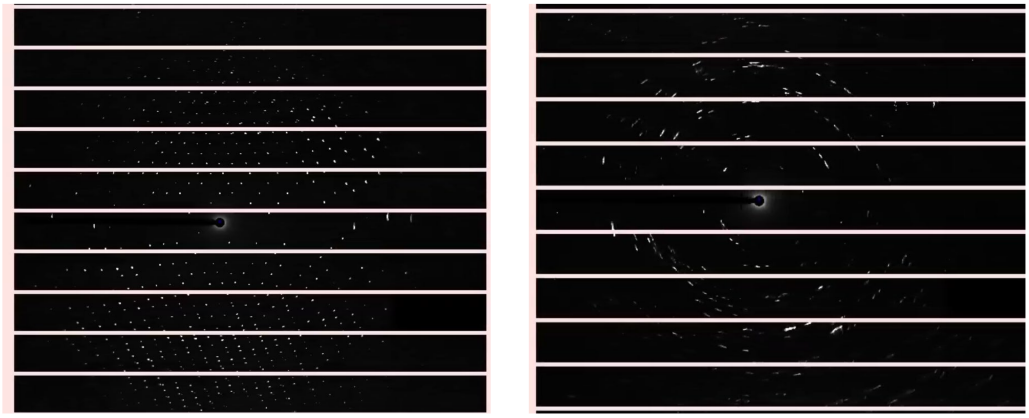


Figure 2.3: Examples of diffraction patterns where high mosaicity results in streaked, broadened, and less well-defined diffraction spots.

2.1.2 Diffraction theory and data acquisition

When the X-ray beam is focused on the crystal sample, the X-ray photons are diffracted by the dense electron density surrounding the atoms of the molecules. A detector placed behind the crystal sample detects these diffracted photons. Figure 2.3 shows an example of a diffraction pattern. The center and a line toward the top left are white due to the presence of a metal bar (beam stop) that prevents direct irradiation of the detector. The black spots represent the diffraction patterns caused by the diffracted X-ray photons, commonly referred to as reflections. The deeper the color of the spot, the more photons have contributed to it, indicating that the reflection has a higher intensity. The reflections collected are in

a certain order, primarily due to the orderly arrangement of the molecules within the crystal.

Bragg's Law

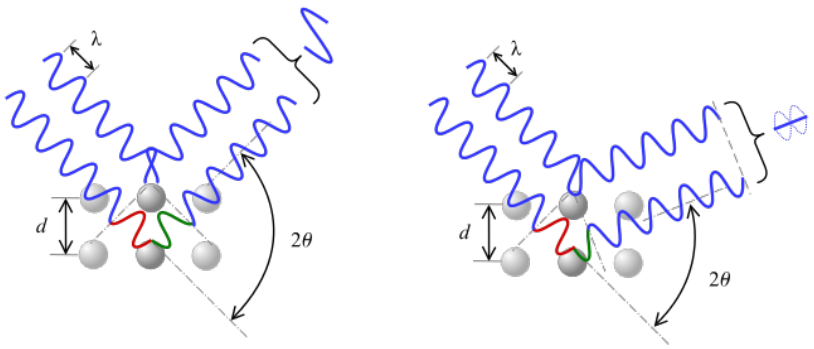


Figure 2.4: Graph representation of Bragg's law with constructive interference (left) and destructive interference (right) [13]

In the previous section on crystal structure, Figure 2.2 illustrates the regular arrangement of molecules, which can be viewed as multiple parallel planes in different orientations. This concept is further explained in Figure 2.4. The regular molecules in each small unit cell from Figure 2.2 are considered objects in parallel planes in Figure 2.4. When an X-ray wave encounters an atom in a unit cell, it is reflected at an angle equal to the incident angle due to the law of reflection. Simultaneously, another atom in a parallel plane beneath the first will also reflect the wave. Figure 2.4 demonstrates that two parallel incident X-ray photons will result in parallel reflected photons, regardless of the incident angle, due to reflections from parallel planes.

X-rays are a type of electromagnetic wave, possessing wave-like properties, including the ability to interfere. As shown on the left side of Figure 2.4, the reflected wave from the upper repeating unit is in phase with that from the lower repeating unit, resulting in constructive interference with increased intensity. The spheres illustrated in Figure 2.4 represent repeating motifs in a crystal, which could correspond to single atoms, small molecules such as salts, or larger biological molecules like proteins. Conversely, the right

side of Figure 2.4 shows waves out of phase, leading to destructive interference with zero intensity. This phenomenon was demonstrated by Lawrence Bragg and his father, William Henry Bragg [3]. The formula for Bragg's law is given below:

$$2d_{hkl}\ \sin(\theta) = n\lambda \quad n \in \{1, 2, 3, 4..., \infty\} \tag{2.1}$$

where d_{hkl} is the distance between corresponding hkl successive crystallographic planes (hkl is called Miller indices that define the orientation of these planes within the crystal lattice), θ is the incident angle of the X-ray wave, λ represents the wavelength of the incident X-ray, and n is the diffraction order ($n = 1$ is first-order, $n = 2$ is second order and so on). This equation specifies the condition on θ required for the optimal constructive interference, resulting in a bright reflection on the detector.

In theory, for a perfect crystal, Bragg diffraction occurs only when the incident angle precisely satisfies the Bragg condition. If the angle deviates even slightly, phase mismatch accumulates across the many atomic planes, leading to destructive interference and practically no diffraction signal. In such a case, the diffraction condition is extremely sharp, and the reflected intensity would be observed over an infinitesimally narrow angular range. This ideal behavior is directly related to the Lorentz factor, which corrects for the geometrical probability that a reflection is captured during a scan and will be mentioned later. In practice, however, real crystals exhibit mosaicity, imperfections, and finite size effects. These imperfections broaden the rocking curve, allowing diffraction to occur over a small but finite range of angles around the Bragg condition. As a result, the diffraction signal is observed even when the incident angle is slightly off, although with reduced intensity. The Lorentz factor remains important in practical measurements to correct for this angular spread and ensure accurate determination of reflection intensities.

Miller Indices

The previous section demonstrated that the angles at which diffracted beams emerge from a crystal can be computed by treating diffraction as reflections from groups of equivalent,

parallel planes of atoms in the crystal. This is why the spots on the detector are called reflections. For a comprehensive understanding and analysis of the crystal structure, the dimensions of a unit cell are determined by six parameters: the lengths of three independent sides a, b, and c, and three independent angles α, β, and γ. These parameters can be calculated during the indexing process, which assigns Miller indices to each reflection (this will be described in Section 2.1.2), using software packages such as XDS [14], MOSFLM [15] and DIALS [16].

The dimensions of the unit cell alone are not sufficient to describe the details of the molecule in each unit cell. As discussed in the previous section, the crystal structure can be described in terms of sets of parallel planes. These planes are an abstract mathematical construct that represents the periodic arrangement of atoms or molecules within the lattice in the crystal. They do not correspond to actual physical layers but are useful for interpreting diffraction patterns and crystal symmetry. These planes can be analyzed and identified based on the specific positions and orientations of the atoms within the unit cells. A notation system called Miller indices is introduced to express these sets of planes in each unit cell. Miller indices *h, k, l* are three integers that label the sets of parallel planes in real space and are also parameters of the basis vectors of the reciprocal space, which is useful for modeling situations when Bragg's condition is satisfied. The indices *h, k, l* correspond to the number of planes in each unit cell in the directions of *x, y, z* in Cartesian coordinates, respectively.

Figure 2.5 illustrates key features of Miller indices (hkl) through three examples. Miller indices describe sets of parallel planes within a unit cell by indicating how the planes intersect the crystallographic axes. In the first example, (111), the planes intersect each axis once, forming a symmetric diagonal across the unit cell. In the second example, (102), the planes intersect the a_1 and a_3 axes once and twice, respectively, but do not intersect the a_2 axis because the middle index $k = 0$. This means that the planes are parallel to the a_2 (y-) axis. The third example, $(\bar{1}02)$, shows a negative index: the bar over the 1 indicates that the plane intersects the negative side of the a_1 axis, reversing the orientation compared to (102). These examples highlight three important features of Miller indices:

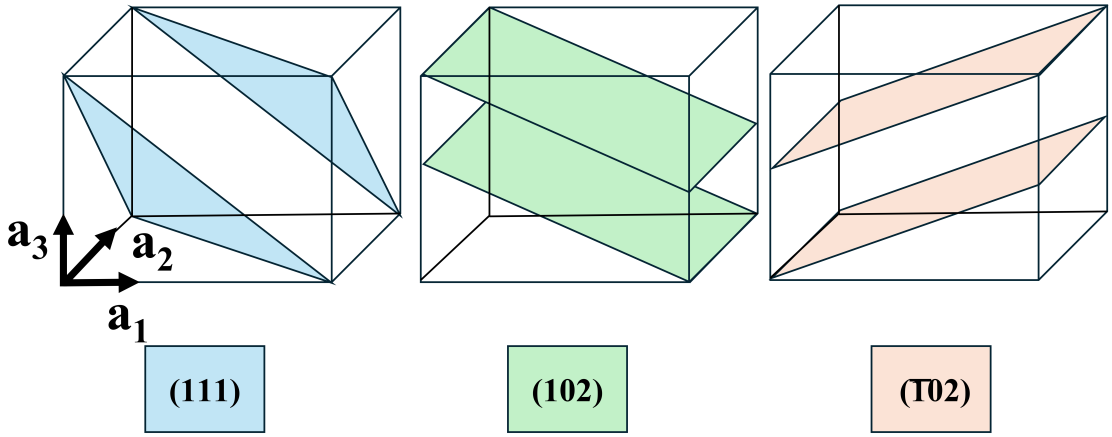


Figure 2.5: More examples with 0 value and negative value of the Miller indices [17]

(1) nonzero indices determine the number of intersections with the corresponding axis, (2) a zero index indicates that the planes are parallel to that axis, and (3) a negative index signifies intersection in the negative coordinate direction.

Reciprocal Lattice and data indexing

Bragg's law states that when a certain angle satisfies the condition given by Equation 2.1, there will be a strong reflection resulting from the interaction between the X-ray beam and the electrons in the molecules. The parameter d_{hkl} represents the spacing between the sets of planes with Miller indices (*hkl*). However, the critical information in Bragg's condition is angular and inconvenient to analyze. Therefore, a reciprocal lattice is introduced to transform the angular relationship into a regular lattice arrangement, similar to a crystal. The construction of the reciprocal lattice is illustrated as a 2D plane in Figure 2.6. The black lines represent the natural lattice of the crystal, with the origin of the direct natural lattice marked as O (in blue) and unit cell dimensions *a, b*. The reciprocal lattice and the natural lattice share the same origin. The indices of the reciprocal lattice are calculated

and formatted by a blue line perpendicular to a plane with the set of Miller indices passing through the origin. In this 2D example, the plane of Miller indices is shown as the red lines in Figure 2.6. The other end of the line terminates at a length of $\frac{1}{d_{hkl}}$ from the origin. Finally, the reciprocal lattice points corresponding to different Miller indices (*hkl*) are constructed. The inverse relationship ($\frac{1}{d_{hkl}}$) between real space and reciprocal space indicates that a smaller unit cell in the natural lattice will result in a larger reciprocal lattice and vice versa. Additionally, for non-orthogonal unit cells, the reciprocal axes will not align with the axes of the natural lattice, aiding in the determination of unit cell dimensions.

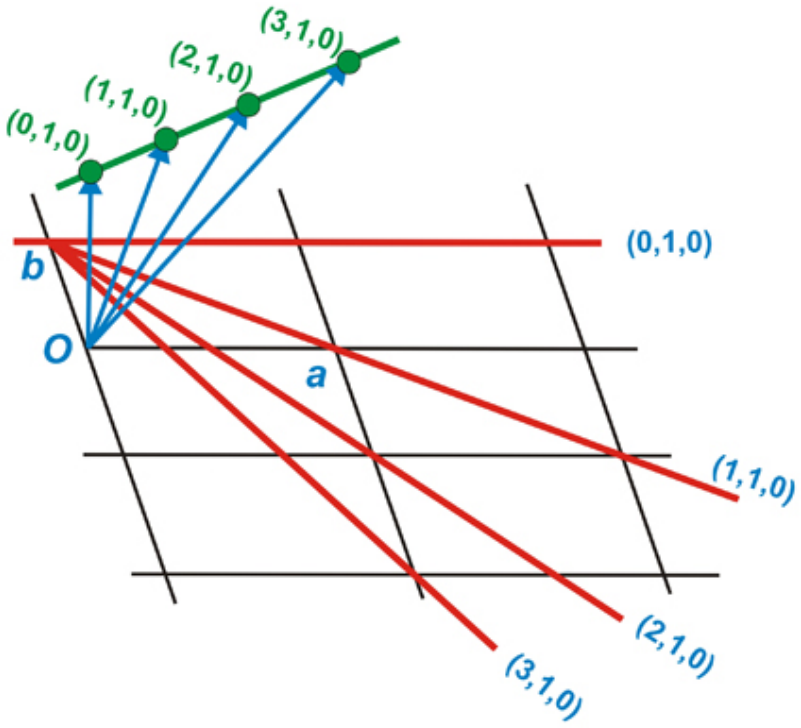


Figure 2.6: Direct real lattice of the crystal (black) and the reciprocal lattice (green) in the 2D plane. *a,b* (blue) is the unit cell edge lengths, and the values in brackets (blue) are the Miller indices [18]

By introducing the reciprocal lattice with points (*hkl*) at a distance of $\frac{1}{d_{hkl}}$ from the origin, Bragg's condition is satisfied when the diffracted beam touches these points. Figure 2.7 illustrates how the reciprocal lattice aligns with respect to the X-ray beam to satisfy Bragg's law and predicts the diffracted beam in the 2D case. Specifically, the incident beam wavevector k_0 propagates in a specific direction and terminates at the origin of both

the natural and reciprocal lattices. This origin can be chosen arbitrarily, allowing for the incident beam wavevector k_0 to be defined accordingly. The diffracted beam wavevector k_1 forms an angle 2θ with k_0, but both have identical magnitudes of $\frac{1}{\lambda}$, where λ is the wavelength. If the diffracted beam wavevector k_1 also contacts a point on the reciprocal lattice, as shown in Figure 2.7, then a position vector d^* for k_1 with a length of $\frac{1}{d_{hkl}}$ can be constructed. By solving the resulting geometric triangle, the following equation is derived:

$$\sin(\theta) = \frac{\frac{d^*}{2}}{|k_0|}$$

$$\sin(\theta) = \frac{\frac{1}{2d_{hkl}}}{\frac{1}{\lambda}}$$

$$2d_{hkl} \sin(\theta) = \lambda \tag{2.2}$$

This satisfies Bragg's law with $n = 1$. By constructing a circle with a radius of $\frac{1}{\lambda}$ centered at the start of the wavevector k_0, Bragg's condition is fulfilled when the edge of the circle touches the points of the reciprocal lattice. These points directly represent the real (*h, k, l*) planes in the real lattice, and the angle θ at this moment is the diffracted angle. This relationship is crucial for predicting the locations of reflections on the detector and indexing the reflections with Miller indices (*h, k, l*).

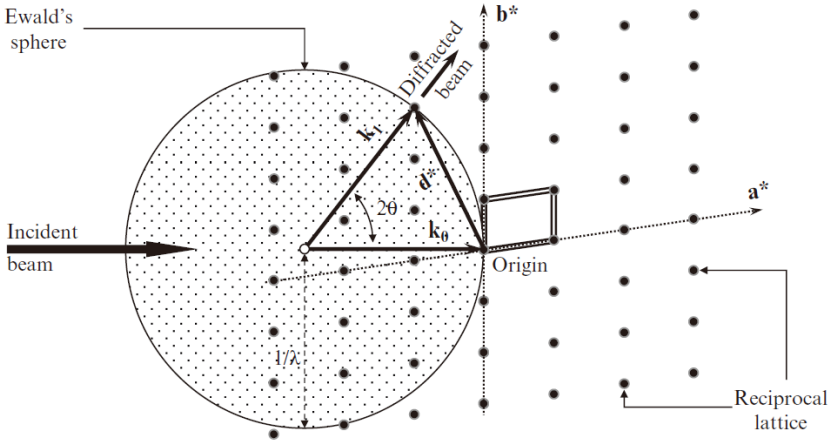


Figure 2.7: Illustration of Ewald sphere in 2D with a radius of $\frac{1}{\lambda}$ in the reciprocal lattice space [19]

In three dimensions, the circle becomes a sphere, known as the Ewald sphere. The points of the reciprocal lattice that satisfy the diffraction condition correspond to the (h, k, l) planes. Therefore, by rotating the crystal around 360°, all the (*h, k, l*) planes and their corresponding intensities can be obtained, facilitating data analysis to reconstruct the molecule's structure in the unit cell. The process to match the measured intensities with the Miller Indices is called **data indexing** and it identifies the geometric relationship between the crystal lattice and the observed diffraction reflections.

2.1.3 Data correction

In crystallography, the spatial distribution of electrons within a crystal, known as the electron density $\rho(\mathbf{r})$, is a periodic function due to the regular arrangement of unit cells. Such periodic functions can be expressed using a Fourier series, which decomposes the function into a sum of sinusoidal (or complex exponential) components.

In one dimension, a complex-valued function $s(x)$ with period P, integrable over $[0, P]$, can be written as a Fourier series of the form:

$$s(x) = \sum_{n=-\infty}^{\infty} c_n \, e^{i2\pi nx/P} \tag{2.3}$$

where the Fourier coefficients c_n are given by:

$$c_n = \frac{1}{P} \int_0^P s(x) \, e^{-i2\pi nx/P} \, dx \tag{2.4}$$

These coefficients c_n capture the contribution of each frequency component n to the overall function. Extending this concept to three dimensions, the electron density $\rho(\mathbf{r})$, with $\mathbf{r} = (x, y, z)$, is periodic in all three spatial directions. It can thus be expanded as a three-dimensional Fourier series over reciprocal lattice vectors $\mathbf{h} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$:

$$\rho(\mathbf{r}) = \sum_{\mathbf{h}} F_{\mathbf{h}} \, e^{2\pi i \, \mathbf{h}\cdot\mathbf{r}} \tag{2.5}$$

Here, the coefficients $F_{\mathbf{h}}$ are the three-dimensional analogues of the 1D Fourier coefficients c_n, and are referred to as the **structure factors**. In three dimensions, the concept of

periodicity extends from an interval of length P to a unit cell of volume V. Therefore, $F_{\mathbf{h}}$ is calculated using:

$$F_{\mathbf{h}} = \frac{1}{V} \int_{\text{unit cell}} \rho(\mathbf{r})\, e^{-2\pi i\, \mathbf{h}\cdot\mathbf{r}}\, d\mathbf{r} \tag{2.6}$$

where V is the volume of the unit cell. In crystallographic notation, **h** is indexed by Miller indices (h, k, l), so the structure factor is commonly written as F_{hkl}.

When X-rays are incident on a crystal, they are scattered by the electrons within the unit cell. The scattered waves from different electrons interfere constructively or destructively, depending on their relative phases. The total scattered amplitude in a given direction is described by the structure factor F_{hkl}, which encodes both the magnitude and phase of the wave scattered by the entire unit cell. The magnitude $|F_{hkl}|$ determines the intensity of the corresponding diffraction peak, while the phase $\arg(F_{hkl})$ governs the spatial features of the reconstructed electron density map.

However, diffraction detectors can only record the energy of the scattered waves, not their phase. As a result, only the intensities I_{hkl}, which are proportional to the square of the magnitudes of the structure factors, can be measured as $I_{hkl} \propto |F_{hkl}|^2$. In practice, the relationship between measured intensity and the ideal structure factor is affected by several experimental factors. Diffraction theory assumes ideal conditions, such as perfect crystal order, a uniform X-ray source, and precise alignment with the Ewald sphere, which are rarely achieved. Therefore, to accurately determine the structure factors from measured intensities, a series of corrections must be applied. These include the polarization factor P, the Lorentz factor L, absorption corrections A, and other instrumental or geometrical corrections denoted collectively as O. Incorporating these, the corrected intensity can be expressed as $I_{hkl} = PLAO \cdot |F_{hkl}|^2$. These corrections are essential to ensure that the final reconstructed electron density reliably reflects the contents of the unit cell.

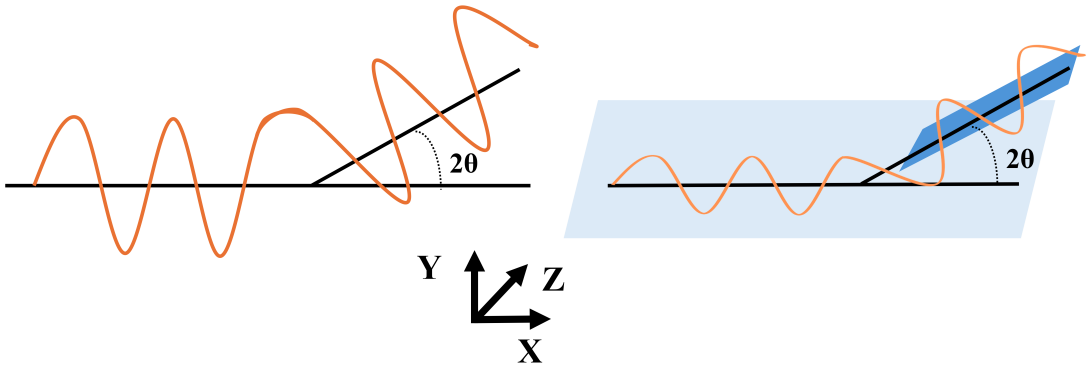


Figure 2.8: Polarization of two extreme cases. The left represents the maximum polarization error, with oscillation direction along with the Y-axis. The right oscillates along with the Z-axis, which has no intensity change of the wave.

Polarisation factor

Since X-rays are a type of electromagnetic wave, the effect of polarization must be taken into account in crystallographic experiments. Polarization errors arise because the scattering process depends on the orientation of the X-ray's electric field relative to the scattering plane. Specifically, the intensity of the scattered X-rays varies with the polarization direction before and after scattering.

Physically, X-ray scattering occurs because the incident electric field drives oscillations of electrons, which then re-radiate scattered waves. The strength of the scattered wave depends on the component of the electric field that is perpendicular to the scattering direction. If the electric field lies within the scattering plane, the scattered amplitude is reduced, and if it is perpendicular, scattering is maximized. The amplitude reduction follows a $\cos(2\theta)$ dependence, where 2θ is the scattering angle. Since the detected intensity is proportional to the square of the amplitude, the polarization correction factor becomes $\cos^2(2\theta)$.

Two extreme cases can be considered to correct for polarization effects. These cases depend on whether the polarization vector remains in the same plane before and after scattering. As illustrated on the left side of Figure 2.8, in the first case, the polarization direction changes with scattering and forms an angle of 2θ relative to its original direction.

The scattered electric field can be projected onto the horizontal axis, and the corresponding intensity reduction factor is $\cos^2(2\theta)$. In contrast, as shown on the right side of Figure 2.8, if the polarization vector remains unchanged by the scattering process, there is no loss in intensity, and the reduction factor is 1.

In principle, both extreme cases can be realized experimentally at synchrotron facilities such as the Diamond Light Source (UK) by carefully controlling the incident beam polarization and the diffractometer geometry. However, in typical laboratory setups or when using unpolarized X-ray beams, the observed intensity is an average over all possible polarization states. Therefore, the effective polarization correction factor becomes:

$$P(2\theta) = \frac{1 + \cos^2(2\theta)}{2} \tag{2.7}$$

where $P(2\theta)$ is the polarization correction factor as a function of the scattering angle 2θ. Additionally, input beam divergence must be considered when correcting for polarization effects. Beam divergence means that the incident X-rays are not perfectly parallel but spread over a small range of angles. This causes a distribution of effective scattering angles for each reflection, slightly smearing the measured intensities. As a result, the polarization correction based on a single scattering angle becomes an approximation. For small divergences, the corrected polarization factor can be approximated by expanding $P(2\theta)$ in a Taylor series, leading to:

$$\langle P(2\theta) \rangle \approx \frac{1 + \cos^2(2\theta)}{2} + \sigma^2 \left(\sin^2(2\theta) - \cos^2(2\theta) \right) \tag{2.8}$$

where σ is the standard deviation of the incident beam divergence (in radians), and $\langle P(2\theta) \rangle$ is the averaged polarization correction factor. In this expression, the second term accounts for the influence of divergence. In practice, for highly collimated synchrotron beams where σ is extremely small, the correction is negligible. However, for laboratory sources with significant divergence or for very high-angle reflections, applying this correction improves the accuracy of intensity measurements.

Lorentz Factor

In Bragg's law, reflections are observed only when the Bragg condition is satisfied, given by $2d_{hkl}\sin(\theta) = n\lambda$, where n is an integer. This condition corresponds to constructive interference of X-rays scattered by the crystal lattice, resulting in strong reflections. Geometrically, in reciprocal space, reflections occur when a reciprocal lattice point lies on the surface of the Ewald sphere.

However, in practice, reflections are observed not only at the exact Bragg condition but also within a small angular range around it, due to the finite width of diffraction peaks and experimental resolution. As the crystal rotates during data collection, reciprocal lattice points move relative to the Ewald sphere. The intensity measured for a given reflection depends on how long the corresponding reciprocal lattice point stays near the surface of the Ewald sphere during this rotation.

This effect is quantified by the Lorentz factor. The Lorentz factor accounts for the fact that reflections corresponding to different scattering angles θ traverse the Ewald sphere at different speeds. Specifically, reflections at lower angles θ spend more time satisfying the Bragg condition because, during crystal rotation, the corresponding reciprocal lattice points move more slowly across the Ewald sphere. At small angles, the points are nearly parallel to the rotation axis, resulting in slower traversal and a longer duration of diffraction. Conversely, at higher angles, the reciprocal points cross the Ewald sphere more quickly, leading to shorter diffraction times. As a result, low-angle reflections naturally accumulate more intensity than high-angle reflections, unless corrected. The Lorentz factor compensates for this difference by adjusting the observed intensities based on the scattering angle θ. Mathematically, the Lorentz factor $L(\theta)$ is given by:

$$L(\theta) = \frac{\text{constant}}{\sin(\theta)\sin 2\theta} \tag{2.9}$$

where θ is the Bragg angle. The factor $\sin(\theta)\sin 2\theta$ arises from the geometry of the reciprocal lattice point movement relative to the Ewald sphere. In many practical cases, the constant prefactor is taken as 1 for normalization.

In summary, while Bragg's law determines the condition for observing a reflection, the Lorentz factor corrects for the varying duration that reflections are observed depending on their scattering angle. Importantly, the Lorentz correction arises purely from the measurement geometry, specifically the relative motion of reciprocal lattice points during crystal rotation, and does not reflect any intrinsic property of the sample itself. Applying the Lorentz correction is essential for accurate intensity measurements in X-ray crystallography, ensuring that observed intensities properly reflect the underlying structure factors.

X-ray absorption

X-rays are absorbed as they pass through materials according to the Beer-Lambert law:

$$I = I_0 e^{-\mu L} \tag{2.10}$$

where I_0 is the incident intensity, I is the attenuated intensity after the X-ray travels a length L through the material, and μ is the absorption coefficient of the material. The value of μ varies with the atomic composition of the material and the X-ray wavelength. Equation 2.10 demonstrates the linear relationship between the incident intensity I_0 and the transmitted intensity I of X-rays passing through a sample. This relationship is typically expressed in terms of the linear absorption coefficient, which assumes a simple, one-dimensional (1D) scenario where X-rays pass through a material of uniform thickness along a straight path. However, in reality, the process of X-ray transmission through a crystal sample is inherently three-dimensional (3D).

The entire crystal is fully illuminated by the X-ray beam, such that the whole volume contributes to diffraction and absorption effects. This requires integrating the X-ray attenuation over the entire crystal volume. If only part of the crystal were illuminated, or if the beam size were much smaller than the crystal dimensions, a more localized or partial volume correction would be necessary. This complexity must be accounted for in a more accurate representation of the intensity relationship. To modify the equation from a 1D to a 3D process, the simple linear relationship needs to be extended to consider the entire

volume of the crystal sample. This involves integrating the X-ray absorption effect over the entire X-ray path through the crystal, given by:

$$T = \frac{1}{V}\int_z\int_y\int_x e^{-\mu L}\, dx\, dy\, dz \tag{2.11}$$

where T is the transmission ratio $\frac{I}{I_0}$ and V is the volume of the crystal bathed in the X-ray beam and it represents the mean attenuation over every possible diffracted X-ray path.

Other experimental factors

In addition to the aforementioned factors, various experimental conditions can significantly influence the accuracy of the structure factor. One such factor is the variation in the intensities of the incident X-ray as a function of rotation. This variation can arise from changes in the illuminated volume or due to rotation-dependent variations in the X-ray path length through the crystal, which in turn alters the average beam absorption. Another important factor is the imperfect mosaicity of the crystal or the presence of bulk disorders within the crystal, such as global radiation damage [20]. Potential changes within the crystal can lead to variations in intensity as a function of resolution (d-spacing). In the context of data processing software such as AIMLESS [21] and DIALS [22], these variations are referred to as the scale term and decay term, respectively. These terms account for physical phenomena occurring during the experiments and are crucial for accurate data interpretation.

2.1.4 Model Building

The electron density map is mathematically derived as the inverse Fourier transform of the structure factor F_{hkl}. The structure factor F_{hkl} represents the scattered wavefront and can be expressed in complex form, separating its magnitude and phase angle as follows:

$$\rho(x,y,z) = \sum_h \sum_k \sum_l |F_{hkl}|\, e^{-2\pi i(hx+ky+lz-\phi'_{hkl})} \tag{2.12}$$

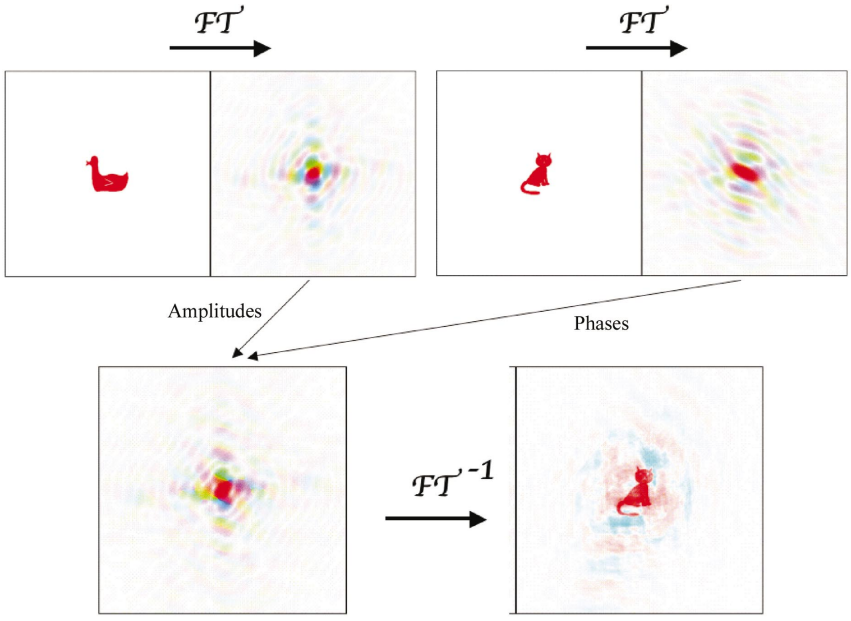


Figure 2.9: An illustration of the importance of phase information. The inverse Fourier transform (IFT) of the combination between the magnitude of the Fourier transform (FT) of the duck image, and the phase of the FT of the cat figure still produces a blurred cat image. [23]

where ϕ'_{hkl} is the phase angle. However, diffraction detectors can only measure the magnitude of the scattered X-rays (the structure factor) by counting the number of photons, which relates to the intensity by $I_{hkl} \propto |F_{hkl}|^2$. To perform the inverse Fourier transform, both the magnitude and the phase of F_{hkl} are required. In X-ray crystallography data analysis, the inverse Fourier transform is discrete, and each reflection corresponds to a specific phase. Unfortunately, directly detecting the phase information in crystallography is currently unfeasible. This challenge is known as the phase problem of crystallography. Phase information is vital for reconstructing the electron density map from the structure factor because the phases indicate the positions of the (*h, k, l*) planes. Figure 2.9 from Kevin Cowtan [23] provides a good visualization of this concept. By combining phases with magnitudes, the inverse Fourier transform can be used to produce the electron density map of the unit cell.

Previously, the Direct Method [24] and the Patterson map method [25] were popular for solving small molecules with a limited number of reflections. The Direct Method is based on a triplet relation between the phases of three reflections ($\alpha_h + \alpha_{h'} + \alpha_{h-h'} = 0$) [24].

The Patterson map method produces a map with peaks at inter-atomic vectors rather than at absolute atomic positions, and it does not require phase information [26].

However, for larger and more complex molecules, these methods often do not provide satisfactory results. Practically, although the phase cannot be directly measured, it can be circumvented by using phases from a homologous structure that was previously solved, a technique known as Molecular Replacement [27]. For macromolecules without a sufficiently similar known structure, experimental preparation is required to determine the phases, necessitating the calculation of the phase of each structure factor from the diffraction intensities.

Another approach is Isomorphous Replacement, which involves adding heavy atoms to the molecule to create isomorphous heavy-atom derivatives (same unit cell and orientation of the protein in the cell but with additional heavy atoms). The phase of the heavy atoms can be calculated by comparing the intensities of the original crystal with those of the crystal containing heavy atoms. Thus, the phases of the remaining atoms can be deduced. Similarly, anomalous scattering of heavy atoms produces measurable differences (anomalous differences) between the intensities of Friedel pairs, which are defined as I_{hkl} and I_{-h-k-l}.[28, 29, 30, 31, 32]. SAD involves collecting data at a single wavelength close to the absorption edge of the heavy atom, which simplifies the experiment and reduces the amount of data required. Currently, native SAD has become one of the most popular phase determination methods if molecular replacement is not feasible [32].

Estimating the electron density map from the phases of heavy atoms is often crude, necessitating iterative phase refinement to improve the interpretability of the electron density map. This iterative process involves reducing the discrepancies between the experimental data and the data predicted by the current model. However, an incorrectly initialized map can lead to longer computational times and poorer final electron density map resolutions, especially in anomalous diffraction phasing.

Recently, AlphaFold2 [8] has demonstrated a significant breakthrough in protein structure prediction from amino acid sequences. This success suggests that neural networks can

potentially solve the transformation from initial prior knowledge to the electron density map. Researchers have explored using AlphaFold2 for molecular replacement, showing that AlphaFold2 can serve as a distillation model that integrates various phasing methods into a single black-box approach [33].

2.1.5 Judging and refining data quality

After applying the data correction methods mentioned in Section 2.1.3, phase determination can commence. By combining the magnitude of the structure factors with the corresponding phase information, the electron density map can be generated from Equation 2.12. Once the electron density map is obtained, the next step involves interpreting this map to identify the positions of atoms within the crystal structure. This is typically done by fitting the electron density peaks to atomic models, where the peaks correspond to the locations of individual atoms or groups of atoms. The atomic coordinates are then refined through iterative cycles of model building and refinement. This process leads to the determination of the precise arrangement of atoms in the crystal, which can then be validated against known structural data.

A perfect electron density map requires accurate magnitude and phase information for each reflection. However, perfect magnitude measurements for reflections are unattainable due to the various factors discussed in section 2.1.3, particularly the absorption effect in macromolecular X-ray crystallography. According to Bragg diffraction theory, the intensity of paired diffraction peaks corresponding to the same set of Miller indices should be consistent and exhibit the same magnitude. This is because, for a given set of Miller indices (hkl), the crystal planes that generate these reflections are symmetrically equivalent, meaning they are related by the symmetry crystal's planes. As a result, the scattering of X-rays by these planes should produce diffraction peaks of equal intensity when measured under identical conditions. Any significant discrepancies in the intensities of such paired reflections could indicate experimental errors or intrinsic factors within the crystal structure itself. For example, anisotropic displacement refers to atoms vibrating more in certain

directions than others, which can affect the scattering intensity.

This fact has been used to define various measures of the internal consistency of the corrected intensities, such as the R_{merge} value [34], the R_{meas} value [35] and the R_{pim} value [36], given by:

$$R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n}|I_i(hkl) - \overline{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)} \tag{2.13}$$

$$R_{meas} = \frac{\sum_{hkl}\sqrt{\frac{n}{n-1}}\sum_{i=1}^{n}|I_i(hkl) - \overline{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)} \tag{2.14}$$

$$R_{pim} = \frac{\sum_{hkl}\sqrt{\frac{1}{n-1}}\sum_{i=1}^{n}|I_i(hkl) - \overline{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)} \tag{2.15}$$

Although the R_{meas} value and the R_{merge} value only differ by a factor of $\sqrt{\frac{n}{n-1}}$, this factor makes R_{merge} independent of the multiplicity and redundancy. Diederichs & Karplus found that the R_{merge} quality indicator can still have a high value even if the corrected intensities are internally consistent because R_{merge} has an implicit dependence on the redundancy of the data [35]. In contrast, R_{pim} shows the precision of the averaged measurement and the merged data. If the R_{meas}, R_{merge}, and R_{pim} values are all very low, it indicates that the data correction procedures have been effective. Low values for these *R* factors suggest that the corrected intensities are consistent, the redundancy does not inflate the R_{merge}, and the precision of the averaged measurements is high, reflecting the overall quality and reliability of the data.

Also, accurate phases are rarely obtained initially, calling for iterative refinement. Iterative refinement aims to minimise the difference between measured integrated intensities and those predicted by the current model built by the integrated intensities and the current corresponding phases until a specified criterion is met (*R* factors see below). Least-Squares minimisation is the most popular data refinement technique used to make the model better fit the experimental data and improve the *R*-values (R_{work} value and R_{free} value [37]).

The R_{work} value and R_{free} value are both calculated using the same formula:

$$R = \frac{\sum_{hkl} |F_{\text{obs}}(hkl) - F_{\text{calc}}(hkl)|}{\sum_{hkl} F_{\text{obs}}(hkl)} \tag{2.16}$$

where the summations are taken over different sets of reflections. For R_{work}, the reflections used in model refinement are included, while for R_{free}, a separate set of reflections withheld from refinement is used to independently assess model quality and detect overfitting. Since the model is directly optimized to fit this data, some level of overfitting is inevitable, potentially leading to an overly optimistic measure of model accuracy. The primary difference between these two values lies in the way the data is used during the calculation. When calculating the R_{free} value, a small fraction of the total data, typically around 5-10%, is held back and not used in the model refinement process. This reserved data set is used exclusively to test the model's predictive accuracy to mitigate overfitting, providing a more objective measure of how well the model generalizes beyond the data it was directly fitted to. Typical values for well-refined macromolecular structures are around 0.20 for R_{work} and 0.25–0.30 for R_{free} [38]. An R_{free} significantly higher than 0.30 often indicates potential issues such as overfitting, model bias, or problems with the crystal or data quality. Generally, the lower the R-values, the more accurate and reliable the determined structure. Larger R_{free} gaps or high absolute values are correlated with greater errors in atomic positions, B-factors, and overall model geometry. Therefore, maintaining low R-values is essential to minimize errors and ensure the physical realism of the final structure.

2.2 Absorption correction in macromolecular X-ray crystallography

Absorption correction is the primary focus of this project due to its significant impact on long-wavelength macromolecular crystallography.

The absorption effect increases approximately with the cube of the X-ray wavelength, scaling as λ^3. For example, at a wavelength of 4 Å the absorption coefficients would be $(4/2)^3 = 8$ times stronger compared to a 2 Å experiment. Considering the exponential attenuation of the transmitted intensity, the structure factor would be reduced by a factor

of $e^{-8} \approx 0.0003$, illustrating the severe impact of absorption at longer wavelengths. The pronounced absorption effect in this context can compromise data quality, making accurate scaling and correction vital. These improvements are crucial for enhancing the success rates of phasing methods such as molecular replacement, anomalous scattering, and isomorphous replacement, which are essential for generating low-resolution electron density maps. Given the complexity of evaluating the integral in Equation 2.11, various methods have been developed to address this challenge.

Compared to other experimental corrections such as beam divergence, polarization effects, or mosaicity, absorption correction becomes particularly critical at long X-ray wavelengths due to the increased absorption cross-section of biological samples. In this regime, uncorrected absorption effects can lead to systematic errors in intensity measurements of up to 10–30%, significantly larger than typical polarization or divergence corrections, which usually account for only a few percent. Therefore, absorption correction represents one of the dominant sources of systematic error in long-wavelength macromolecular crystallography and must be prioritized during data processing to ensure accurate scaling and reliable phasing.

2.2.1 Analytical and numerical absorption correction

The analytical absorption correction employs a fundamental approach to simulate the actual diffracted and incident paths for calculating the transmission factor. A 3D case is presented in Equation 2.11. Since the path length L varies with the size and shape of the crystal and the diffracting plane [9], the overall intensity of the diffracted beam is determined by averaging the integral of the transmission factors for each individual diffracted path. Solving the integral in Equation 2.11 is extremely challenging and often intractable for complicated crystal shapes. Thus, both analytical and numerical solutions have been introduced to address this problem.

Initially, to solve the integral, an analytical approach was proposed that involved dividing the crystal into multiple tetrahedrons using Howells' polyhedra method [39], and summing

the integrals over all tetrahedrons [40]. This transforms Equation 2.11 into:

$$T = \sum_{i=1}^{N} \frac{1}{V_i} \iiint_{V_i} e^{-\mu(L_1+L_2)}\, dV_i \tag{2.17}$$

where L_1 and L_2 are the incident and diffracted X-ray path lengths. Each integral calculates the transmission factors over a tetrahedron, which is analytically solvable. The equation describing a tetrahedron with vertices at $(0,0,0)$,$(u,0,0)$,$(0,v,0)$,$(0,0,w)$, and the equation of the path are given by [40]:

$$\frac{x}{u} + \frac{y}{v} + \frac{z}{w} = 1 \tag{2.18}$$

$$L = L_1 + L_2 = px + qy + rz + s \tag{2.19}$$

where u, v, w, p, q, r, and s are constants, and $px + qy + rz + s$ represents the equation of the diffracted surface. Consequently, a solvable integral over a single tetrahedron can be computed as:

$$T_i = \iiint_{xyz} e^{-\mu(px+qy+rz+s)} dxdydz \tag{2.20}$$

$$T_i = e^{-\mu s} \int_0^{x=u} \int_0^{y=1-\frac{x}{u}} \int_0^{z=1-\frac{x}{u}-\frac{y}{v}} e^{-\mu px} e^{-\mu qy} e^{-\mu rz} dxdydz \tag{2.21}$$

Summing all individual transmission factors T_i yields the overall transmission rate of the crystal. However, indexing the faces and vertices of all polyhedra is computationally expensive. To address this, several approaches have been proposed to accelerate the process. Alcock *et al.* [41] reduced the computational time by identifying parallel faces with identical absorption properties. Clark (1993) [42] and Clark *et al.* (1995) [43] transformed volumetric integration into edge integration using Gauss's theorem and Stokes's theorem. Alternatively, numerical methods have been popular for solving the integral, especially in cylindrical and spherical cases [10]. For example, the ABSORB software [44] calculates absorption correction for crystals in capillaries with trapped mother liquor. These numerical integrations are often based on Gaussian Quadrature approximation, as shown in:

$$T = \frac{1}{V}\sum_{j=0}^{n}\sum_{x_j,y_j,z_j} w_j e^{-\mu L(x_j,y_j,z_j)} \tag{2.22}$$

where w_j are the weights of the absorption contributions from the capillary and trapped liquid [44]. However, these methods still rely on indexing the faces or edges of the crystal, which can be challenging due to the variability in crystal shape and size. Moreover, traditional analytical and numerical methods typically consider only the crystal's absorption. Other components in the experimental setup, illustrated in Figure 1.4, also contribute significantly to the X-ray absorption effect, further complicating data scaling and correction.

2.2.2 Empirical absorption correction

The effectiveness of analytical and numerical absorption corrections is highly dependent on the shape and dimensions of the crystal, posing challenges when the crystal has a complex shape. This complexity is further exacerbated when other materials, such as mother liquor or the mounting loop, are attached to the crystal. To address these issues, North *et al.* (1968) [45] and Furnas (1957) [46] proposed semi-empirical methods involving multiple azimuthal scans to calculate a curve of relative transmission ratio T against the azimuthal angle ϕ for the corresponding reciprocal lattice level. By assuming that the absorption of any reflection is the same as the mean direction of the incident and reflected beams ($T(hkl) = \frac{T(\phi_{\text{incident}})+T(\phi_{\text{ref}})}{2}$), an absorption correction that considers the mother liquor can be calculated. However, this assumption becomes invalid and the semi-empirical method is imprecise when applied to asymmetrical crystals. Also, they require multi-axis goniometers and the additional data needed for the azimuthal scans can contribute significantly to radiation damage on modern synchrotron light sources.

Katayama *et al.* (1972) [47], Katayama (1986) [48], and Blessing (1995) [49] introduced an alternative approach to address this problem. Instead of using azimuthal measurements, they employed an algebraic method to empirically perform absorption correction by determining the relative absorption ratio, known as the absorption surface. They introduced a weighting factor w_{hi} and an anisotropy factor A_{hi} to perform a least-squares minimization,

approximating the observed intensity to the exact intensity. The least-squares formulation is given by:

$$\Phi = \sum_{h} \sum_{i=1}^{n} w_{hi}(A_{hi}I_{hi} - \langle A_{hi}I_{hi}\rangle_h)^2 \tag{2.23}$$

where I_{hi} is the ith observation of the symmetry-unique reflection h, A_{hi} is the absorption function, and w_{hi} is the weighting factor, typically set proportional to the inverse variance of the intensities ($w_{hi} = \frac{1}{\sigma^2(I_{hi})}$) for optimal least-squares estimates [49]. Since the exact intensity of each reflection is unknown, the term $\langle A_{hi}I_{hi}\rangle_h$ represents the average intensity, modelling the true intensity. The anisotropy factors A_{hi} vary across methods. Katayama *et al.* (1972) [47] initially used Fourier series for A_{hi}, but it was later found to lack rotational invariance [48]. Surface harmonics with Legendre functions were subsequently applied to address this issue. Blessing (1995) [49] later introduced spherical harmonics to model the spherical harmonic surface, overcoming the problem of over-correction due to anisotropic extinction. Anisotropic extinction occurs when the absorption of X-rays by a crystal varies in different directions, leading to uneven intensity reductions in the diffraction data. By using spherical harmonics, Blessing was able to more accurately model these directional dependencies, reducing the risk of over-correcting the data.

With the introduction of large area detectors, these numerical methods to obtain an empirical correction for absorption have become popular and spherical harmonics are now the basis for absorption correction in most data reduction software packages for macromolecular crystallography, such as AIMLESS [21], hkl3000 [50], SADABS [51], and DIALS [16, 22], while XDS uses alternative numerical methods without spherical harmonics [14].

Empirical methods aim to model global changes, such as the overall decrease in intensities, rather than the gradual intensity changes caused by different diffracting parts of the sample so data multiplicity needs to be high to ensure the success. The absorption surface is calculated based on the completeness of symmetry-related paired reflections. For instance, irregular crystal shapes can lead to significant attenuation in one of the measured symmetry-related intensities, while the other may have low absolute absorption. Consequently, the

least-squares minimization may inaccurately correct the former and suppress the latter, leading to an incorrect absorption surface. Therefore, empirical methods may fail in long-wavelength X-ray crystallography, due to higher absorption. Additionally, for radiation-sensitive crystals in low-symmetry space groups, it can be difficult to obtain sufficient data multiplicity to ensure the success of empirical methods.

This is because prolonged exposure to X-rays causes specific radiation damage events within the crystal, including bond breakage and structural disorder, which gradually reduce the crystalline order. As a result, the quality of the diffraction spots deteriorates.

2.2.3 Analytical absorption correction by 3D models

As the analytical absorption correction does not rely on refining parameters to minimize differences between structure factor amplitudes of symmetry-related reflections, its effectiveness is independent of data multiplicity. To analytically calculate absorption correction factors for a sample with an irregular shape, detailed characterisation of its shape and orientation is required. Previous studies utilized optical microscopy to reconstruct a three-dimensional model of the sample, including the crystal, sample mounting loop, and mother liquor, demonstrating that absorption correction was feasible and beneficial at lower levels of data multiplicity [52, 53].

An alternative method to obtain a 3D model of the sample is X-ray tomography, which has been used to characterize or visualize crystals [54, 55]. Brockhauser *et al.* [56] suggested using tomographic reconstructions and segmentations as a basis for absorption correction. This approach enables the calculation of X-ray path lengths through the different materials in the sample (crystal, sample mount, and mother liquor), as illustrated in Figure 1.4.

A typical protein crystallography data set contains hundreds of thousands of reflections. There are typically millions of crystal voxels in a 3D model, and each path length calculation can involve determining thousands of voxels along the incident and diffracted X-ray paths. Consequently, calculating all absorption correction factors for samples in protein crystallography is computationally expensive. Leal *et al.* [52] presented a Gaussian random

sampling to decrease the number of crystal voxels in the calculation. However, there is no evidence to show that the sampled crystal voxels are representative enough to obtain an accurate absorption factor.

2.3 Segmented X-ray tomography reconstruction

2.3.1 X-ray tomography reconstruction on synchrotron experiments

X-ray tomography is a non-invasive imaging technique used to obtain cross-sectional images of an object. The general process involves directing X-rays through the object from multiple angles and capturing the resulting attenuation data on detectors positioned around the object. Or it can be achieved by rotating the object with a fixed detector. These attenuation measurements are then used to reconstruct the internal structure of the object. The reconstruction process translates the acquired projection data into a 3D volume of the object that represents the internal features of the scanned object. This is achieved through mathematical algorithms that convert the line integrals of the object's density function into spatial information.

Radon Transform

The foundation of X-ray tomography reconstruction lies in the Radon transform. The Radon transform $R[f]$ of a function $f(x, y)$ represents the integral of f along a line defined by angle θ and distance s from the origin:

$$R[f](s,\theta) = \int_{-\infty}^{\infty} f(s\cos(\theta) - t\sin(\theta), s\sin(\theta) + t\cos\theta)\, dt. \tag{2.24}$$

Therefore, the projection image of the object can be mathematically expressed as the Radon transform of the object at an angle θ. Given a set of projection images from various angles, the Radon transform provides a comprehensive description of the object, which is defined by $f(x, y)$. The key to reconstructing the original object from its Radon transform is the inverse Radon transform, which can be expressed as follows.

$$f(x,y) = \int_0^{\pi} R[f](\theta, x\cos(\theta) + y\sin(\theta))\, d\theta \tag{2.25}$$

This integral formula demonstrates that from knowledge of the Radon transform $R[f]$ for all angles θ ranging from 0 to π, one can reconstruct the original function $f(x,y)$. That can mathematically prove that if a set of projection images with angles from 0° to 180° are obtained, the model of the object can be reconstructed.

Fourier Slice Theorem

The 2D Fourier transform of $f(x,y)$ is defined as:

$$\mathcal{F}\{f(x,y)\}(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i2\pi(ux+vy)}\, dx\, dy. \tag{2.26}$$

The Fourier slice theorem, also known as the central slice theorem, establishes a crucial connection between the Radon transform and the Fourier transform. The theorem states that the 1D Fourier transform of the Radon transform of $f(x,y)$ at a given angle θ is equal to a central slice of the 2D Fourier transform of $f(x,y)$ along a line at the same angle θ. Mathematically, if we denote the 1D Fourier transform of $R[f](s,\theta)$ with respect to s as $\mathcal{F}_1\{R[f](s,\theta)\}(\omega)$, the theorem can be expressed as:

$$\mathcal{F}_1\{R[f](s,\theta)\}(\omega) = \mathcal{F}\{f(x,y)\}(\omega\cos\theta, \omega\sin(\theta)). \tag{2.27}$$

In other words, the Fourier transform of the projection at angle θ is a radial line (slice) in the 2D Fourier transform of the original function $f(x,y)$.

The Fourier slice theorem provides a powerful approach for reconstructing an image from its projections. By taking the Fourier transform of each projection (Radon transform) and placing it at the corresponding angle in the 2D Fourier space, one can build the 2D Fourier transform of the original image. Once all the slices are assembled, the inverse 2D Fourier transform is applied to reconstruct the original image $f(x,y)$:

$$f(x,y) = \mathcal{F}^{-1}\left\{\mathcal{F}\{f(x,y)\}(u,v)\right\}. \tag{2.28}$$

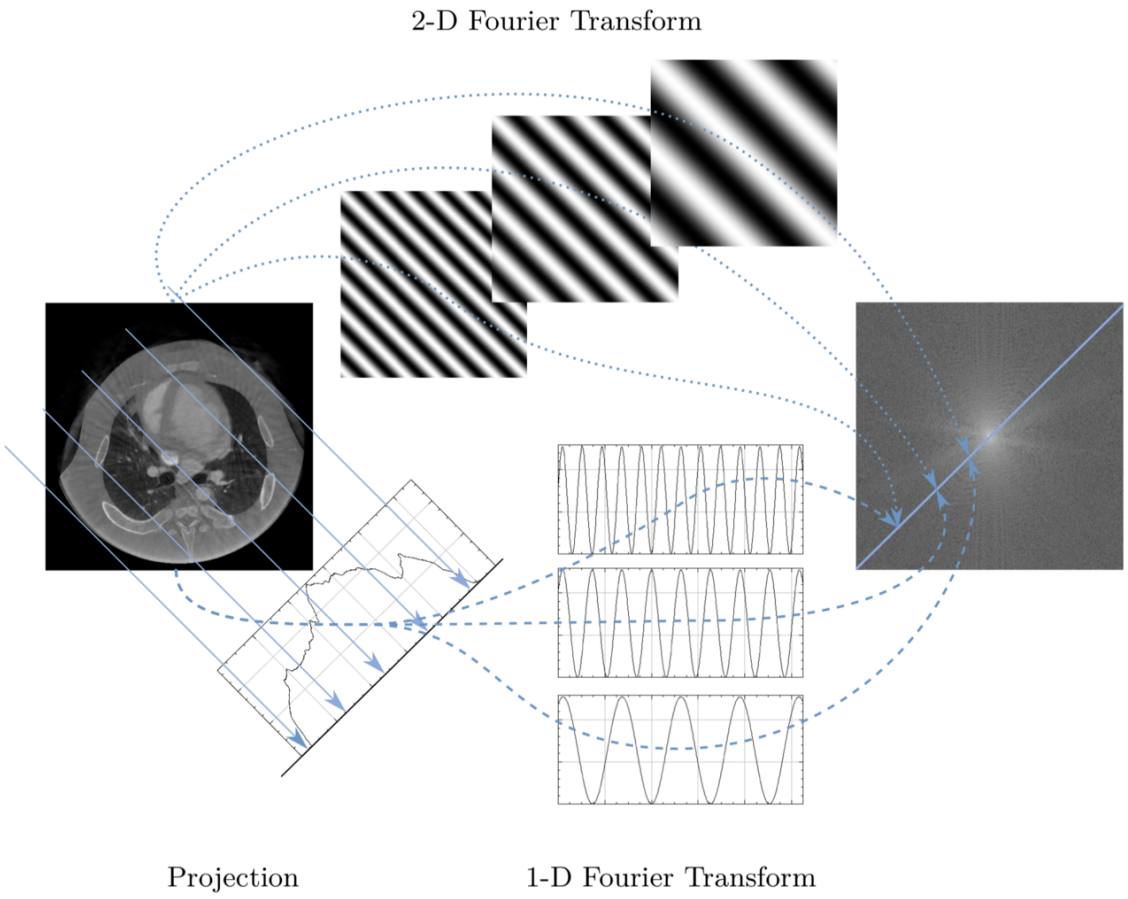


Figure 2.10: Illustration of Fourier slice theorem. The upper path indicates the 2D Fourier transform while the lower path shows the 1D Fourier transform of the projection [57]

This method forms the basis for the Filtered Back Projection (FBP) algorithm, where the filtering step corresponds to compensating for the radial nature of the Fourier space before applying the inverse transform.

Filtered Back Projection (FBP)

Filtered Back Projection (FBP) is a widely used algorithm for reconstructing images from projection data. The FBP method combines two main steps: filtering and back projection. The filtering step applies a high-pass filter to the projection data to counteract the blurring effect of the Radon transform. Mathematically,the filtered projection $\hat{P}(s, \theta)$ in Equation

2.28 is obtained by:

$$\hat{P}(s,\theta) = \mathcal{F}^{-1}\left\{|\omega|\,\mathcal{F}\{P(s,\theta)\}\right\},$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse, respectively, and $|\omega|$ is the frequency response of the filter. In the back projection step, the filtered projections are smeared back along the paths they were originally taken to reconstruct the image:

$$f(x,y) = \int_0^{\pi} \hat{P}(x\cos(\theta) + y\sin(\theta), \theta)\, d\theta.$$

FBP is computationally efficient and provides good image quality, making it a popular choice in X-ray tomography reconstruction and it can be implemented by 2D or 3D approaches. While the principles of FBP remain the same in both 2D and 3D cases, the implementation differs due to the dimensionality of the data. In 2D FBP, the reconstruction is performed slice by slice, where each slice corresponds to a 2D cross-section of the object. This method processes the projection data obtained from multiple angles around a single plane. In contrast, 3D FBP extends this approach to three dimensions, where the projection data is acquired over a range of angles covering an entire volume. The 3D FBP algorithm reconstructs the entire 3D volume by applying the filtering and back projection steps across all three dimensions, which requires more computational resources. However, this allows for the reconstruction of whole volumetric data, providing a more comprehensive representation of the scanned object.

Iterative Methods

Iterative reconstruction methods offer an alternative to traditional techniques like Filtered Back Projection (FBP), often improving image quality and reducing artefacts. These methods work by iteratively refining an initial estimate of the image (or volume in 3D) to minimize the difference between the measured projections and the projections calculated from the current estimate. One widely used iterative method is the Algebraic Reconstruction Technique (ART). ART updates the image estimate $f^{(k)}$ at each iteration k using the following update strategy:

$$f^{(k+1)} = f^{(k)} + \lambda \frac{P_i - R[f^{(k)}]_i}{\|\mathbf{R}_i\|^2} \mathbf{R}_i \tag{2.29}$$

where P_i is the i-th measured projection, $R[f^{(k)}]_i$ is the i-th computed projection from the current image estimate, $\mathbf{R}_i$ is the system matrix row corresponding to the i-th projection, and λ is a relaxation parameter. ART updates the image by considering one projection at a time, which can lead to faster convergence but may also introduce noise or artifacts if not carefully regularized. Iterative methods like ART can incorporate various constraints and regularization terms, making them flexible and robust for different imaging scenarios, such as limited-angle tomography and low-dose imaging.

Another popular iterative method is the Simultaneous Iterative Reconstruction Technique (SIRT). Unlike ART, which updates the image estimate sequentially for each projection, SIRT updates the image by considering all projections simultaneously. The update rule for SIRT is given by:

$$f^{(k+1)} = f^{(k)} + \lambda \mathbf{R}^T (\mathbf{P} - \mathbf{R} f^{(k)}), \tag{2.30}$$

where $\mathbf{P}$ is the vector of all measured projections, $\mathbf{R}$ is the system matrix representing the projection process, $\mathbf{R}^T$ is the transpose of $\mathbf{R}$, and λ is a relaxation parameter. SIRT applies a global correction to the image estimate based on all projections, leading to a more consistent and stable reconstruction process. Although SIRT tends to converge more slowly than ART, it often produces more accurate reconstructions, particularly in the presence of noise or inconsistencies in the data.

Iterative methods are generally more computationally intensive than FBP but can provide superior image quality and, in many cases, higher reconstruction accuracy, particularly under noisy or limited-data conditions, because they incorporate regularization strategies during the optimization process. Both ART and SIRT can be applied in 2D and 3D, where they iteratively refine the image (or volume in 3D) by minimizing the difference between the measured projections and the projections of the current image estimate. However, these methods involve several hyperparameters, such as the number of iterations and regular-

ization parameters, which must be carefully tuned to achieve good reconstruction quality. In 3D implementations, the increased data volume further amplifies these computational demands, making the careful selection and tuning of hyperparameters even more critical.

2.3.2 Segmentation

The tomographic reconstruction alone is insufficient for the calculation of an absorption correction since L in Equation 2.11 represents the path length of the specific material, necessitating segmentation of the reconstruction. The segmentation requires predicting the semantic material for each pixel in an image based on a predefined label set. Popular segmentation methods include point, line, and edge detection-based methods, threshold-based methods, region-based methods, and morphology-based methods. Emerging deep neural network (DNN)-based methods are also gaining attention.

Using a U-Net architecture [58], it was found that this approach outperformed traditional Otsu Thresholding [59] in terms of pixel accuracy and pore estimation [60]. However, this work focused solely on binary classification, distinguishing between the background and the content. Conversely, SAVU, widely used at Diamond Light Source [61], includes a package with thresholding and morphology-based segmentation [62]. The thresholding method employs geodesic distance thresholding (GeoDistance) with Gaussian Mixture Models (GMM) for initial material separation. The morphology-based method also starts with GMM but subsequently applies a novel region-growing technique (RegionGrow) for precise boundary detection. Although this can achieve results comparable to manual segmentation, thereby saving human effort, it requires high-quality iteratively reconstructed data. Also, at the end of the processing, it requires the user to specify each material in the segmentation results, hence this method is only semi-automatic segmentation.

FBP can introduce more ring and streak artefacts compared to iterative methods, even after applying various removal algorithms [63]. These artefacts make pure thresholding or edge detection algorithms less effective, necessitating more advanced algorithms.

U-Net

U-Net is a deep neural network (DNN) architecture specifically designed for image segmentation tasks [58]. It is renowned for its symmetric design and the strategic use of skip connections, which are essential for achieving precise localization, as illustrated in Figure 2.11. The network architecture features two primary paths: a contracting path and an expanding path. The contracting path, or encoder, captures context by progressively reducing the spatial dimensions of the input image and extracting high-level features. In this path, convolutional neural network (CNN) blocks are utilized to extract features, and max pooling layers are used to perform downsampling. Conversely, the expanding path, or decoder, increases the spatial dimensions to restore the original image size while maintaining detailed information. Transpose convolutions are used for upsampling the feature maps during the decoding process.

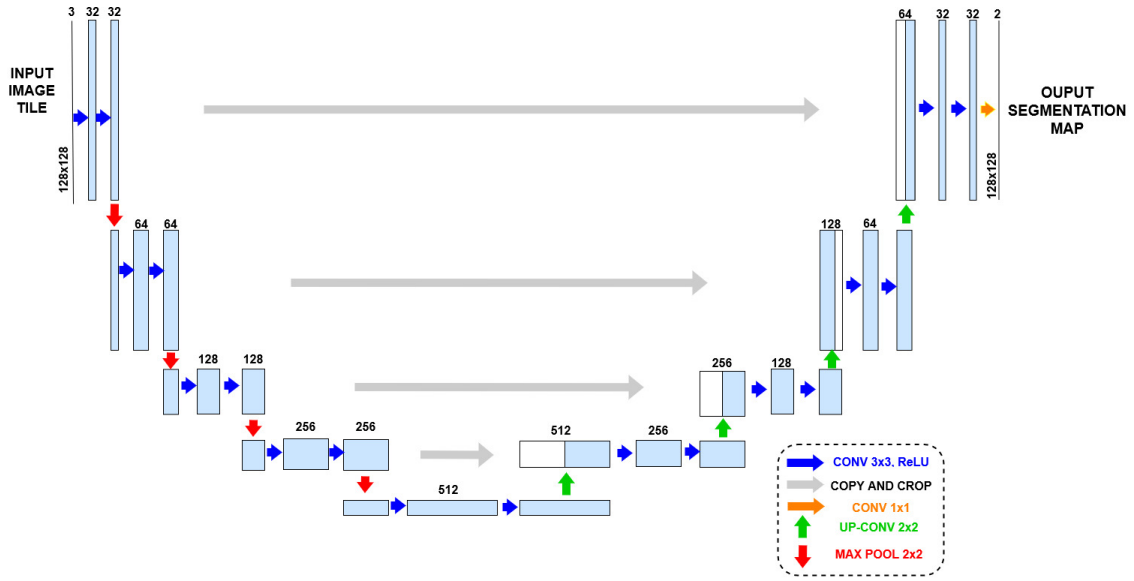


Figure 2.11: An example U-net architecture [64]. Each blue rectangle shows a multichannel feature map, with the number of channels indicated above each box. The white boxes depict the copied feature maps (skip connections).

This architecture is enhanced by skip connections that link corresponding layers in the encoder and decoder paths. These connections preserve fine-grained details by allowing features extracted during downsampling to be directly reused during upsampling, ensuring

that localization is not compromised. Moreover, U-Net is fully convolutional, allowing it to process input images of arbitrary sizes efficiently while keeping the number of parameters relatively low. This design contributes to its efficiency and flexibility. U-Net also emphasizes the use of data augmentation techniques, such as elastic deformations, to artificially expand the training dataset. This approach enhances the model's robustness and generalization capabilities, making it particularly effective in scenarios with limited training data. Overall, U-Net's innovative design and emphasis on data augmentation make it a powerful tool for image segmentation tasks in various applications.

U-Net is particularly advantageous for grey-scale and medical image segmentation due to several intrinsic features. Medical images, including X-rays, MRI, and CT scans, are predominantly grey-scale. U-Net's architecture is specifically designed to handle the subtle variations and high-resolution details typical in grey-scale images, making it an optimal choice for these applications. Furthermore, medical image segmentation demands high precision to accurately delineate structures and abnormalities. The symmetric architecture and skip connections in U-Net maintain high spatial resolution, which is crucial for medical diagnostics. Additionally, obtaining annotated medical images is challenging, resulting in limited dataset sizes. U-Net's design, coupled with effective data augmentation techniques, allows it to perform well even with small datasets, addressing the common issue of data scarcity in medical applications.

Since the introduction of U-Net, CNN-based networks have achieved state-of-the-art results across various 2D and 3D medical image segmentation tasks [65, 66, 67]. For volume-wise segmentation, tri-planar architectures are sometimes employed, combining three-view slices for each voxel, a method known as 2.5D [68, 69]. In contrast, 3D approaches directly utilize the full volumetric image, represented by a sequence of 2D slices or modalities. Multi-scan, multi-path models have been developed to capture downsampled features of the image [70]. To exploit 3D context and address computational resource limitations, hierarchical frameworks have been investigated. For instance, Roth et al. [71] proposed a multi-scale framework to obtain varying resolution information in pancreas segmentation.

These methods represent pioneering studies in 3D medical image segmentation at multiple levels, addressing issues related to spatial context and low-resolution conditions.

X-ray tomography in crystallography involves the detailed 3D imaging of protein crystal samples and requires precise segmentation for an accurate absorption correction to be calculated. U-Net's application in this domain is highly beneficial due to several factors. The need for precise delineation of materials in X-ray tomography parallels the requirements in medical imaging, and U-Net's capability to provide high-resolution segmentation makes it suitable for this application. Crystals and their mounting loops often exhibit sharp edges. The skip connections in U-Net help preserve fine details during segmentation, ensuring accurate representations of these small structures. Although the original U-Net was designed for 2D images, its architecture has been extended to 3D (known as 3D U-Net) to handle volumetric data, such as that produced by X-ray tomography. This extension allows the model to analyze 3D structures more effectively, which is critical for accurate segmentation. Additionally, the limited number of datasets in crystallography makes U-Net an appropriate model candidate due to its robustness in small dataset scenarios.

Transformer

The 3D segmentation of a tomography reconstruction necessitates the extraction of high-level features while preserving the initial 3D spatial resolution. Convolutional Neural Networks (CNNs) are adept at capturing local details and low-level information, but they often fail to capture global context and long-range spatial dependencies. Despite their success in various applications mentioned above, this limitation can significantly impact segmentation performance for more complex tasks. The transformer architecture [72] has addressed this gap by effectively capturing global information and long-range dependencies.

The transformer architecture, initially developed for natural language processing tasks, has revolutionized the handling of sequential data using self-attention mechanisms [72]. The transformer encodes the influence of different parts of the input data, allowing for

more flexible and parallel processing compared to traditional recurrent neural networks (RNNs). While RNNs process input sequences one step at a time and struggle with long-range dependencies due to their sequential nature, transformers compute attention scores between all pairs of input tokens simultaneously. Here, a "token" is a basic unit of data that the model processes, such as words or subwords. This method allows transformers to effectively model relationships and dependencies across the entire input sequence. By understanding the contextual relevance of each token in relation to others, transformers can capture complex patterns and structures in the data that traditional models often miss. This capability to grasp long-range dependencies and global context has greatly advanced natural language processing and computer vision, offering a robust framework for tasks requiring a comprehensive understanding of sequential and spatial information.

The transformer encoder block is a fundamental component of the transformer architecture, and it is usually shortened to "transformer". It processes input sequences by leveraging self-attention mechanisms and feed-forward neural networks. The overall structure of a transformer encoder block consists of two main sub-layers: Multi-Head Self-Attention (MSA) and a Feed-Forward Network (FFN). Each of these sub-layers is followed by residual connections [73] and layer normalization [74] to enhance learning stability and performance, as illustrated in Figure 2.12.

The transformer encoder block processes an input sequence $X \in \mathbb{R}^{N \times d_{\text{model}}}$, where N is the sequence length and d_{model} is the dimensionality of the input embeddings. These input embeddings are typically obtained by projecting (e.g. a linear transformation) the original input tokens into a continuous vector space, allowing the model to work with dense representations of the data numerically. This is because computers can't recognize human language and it is necessary to transform human-readable text into machine-readable codes.

First, the Multi-Head Self-Attention (MSA) mechanism allows the model to focus on different parts of the input sequence simultaneously. It consists of several attention heads that each compute self-attention for different subspaces of the input. The attention mechanism is defined as:

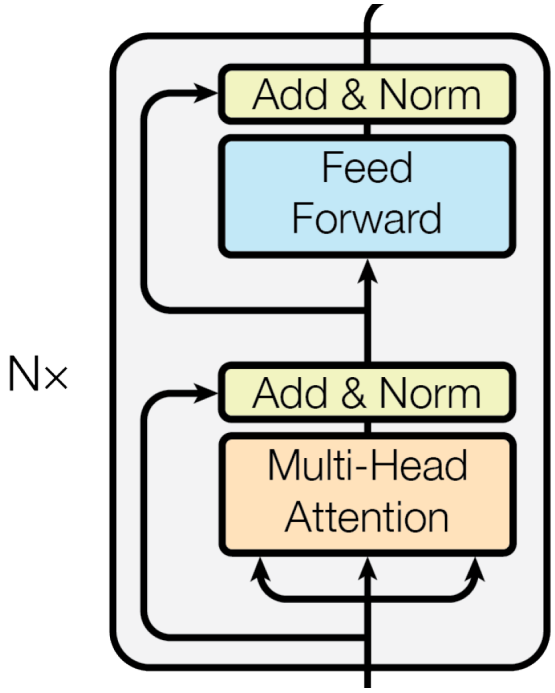


Figure 2.12: An example transformer encoder block [72]. N is the number of this encoder block and it is equal to the length of the input sequence. Add & Norm represents the residual connection [73] and layer normalization[74])

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.31}$$

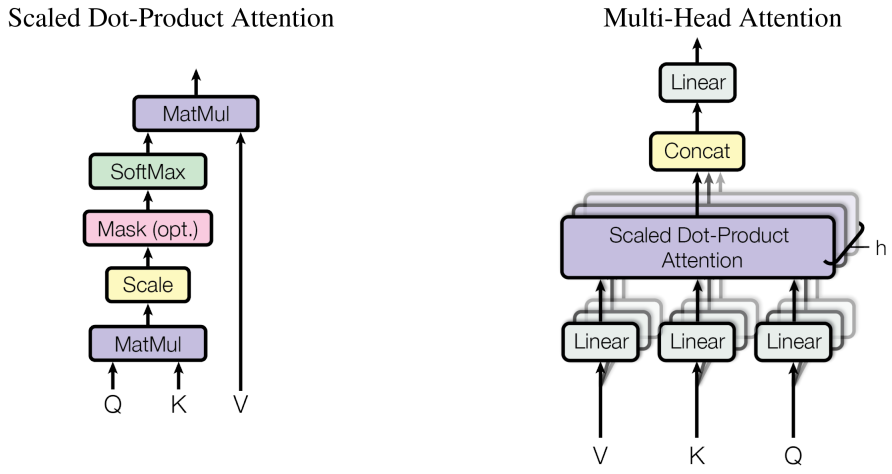


Figure 2.13: (left) Scaled Dot-Product Attention mechanism in Equation 2.31. (right) Multi-Head Attention, comprising h attention layers processing in parallel in Equation 2.32 [72].

Here, $Q = XW_i^Q$ (Query), $K = XW_i^K$ (Key), and $V = XW_i^V$ (Value), where W_i^Q, W_i^K, and W_i^V are the learned linear transformation matrices for the i-th head, and d_k is the dimensionality of the key/query vectors. The query, key, and value vectors are derived from the input embeddings and represent different linear transformations of the input sequence. For example, in a language model, the query might represent the linear transformation of the current word you want to translate (e.g. [“apple”]). The key could represent the linear transformations of all words in the sequence (e.g., “banana”, “apple”, “car”). The value vectors carry the actual information that will be used to update the representation of the query (e.g., “Banane”, “Apfel”, “Auto” corresponding to the German translation).
The outputs of the attention heads are concatenated and linearly transformed:

$$\text{MSA}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2.32}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W^O is the learned linear transformation matrix for the concatenated output.
After the MSA sub-layer, a residual connection and layer normalization [74] are applied to enhance the learning stability:

$$\text{Output}_1 = \text{LayerNorm}(X + \text{MSA}(Q, K, V)) \tag{2.33}$$

Next, the Feed-Forward Network (FFN) consists of two linear transformations with a ReLU activation [75] in between:

$$\text{FFN}(X) = \text{ReLU}(XW_1 + b_1)W_2 + b_2 \tag{2.34}$$

where W_1 and W_2 are the weight matrices, and b_1 and b_2 are the bias vectors. Finally, another residual connection [73] and layer normalization follow the FFN sub-layer:

$$\text{Output}_2 = \text{LayerNorm}(\text{Output}_1 + \text{FFN}(\text{Output}_1)) \tag{2.35}$$

The transformer encoder block effectively captures and processes the input sequence by leveraging self-attention and feed-forward networks, while maintaining stability and

performance through normalization and residual connections. For example, in vision tasks, if we treat image patches as queries, keys, and values, the self-attention mechanism can capture the relationships and dependencies between different parts of the image, leading to a more comprehensive understanding of the visual content and enabling tasks such as image segmentation, object detection, and image generation.

Vision Transformer (ViT)

Building on the foundational concept of the transformer, the Vision Transformer (ViT) applies the transformer encoder blocks to image data. ViT splits an image into a sequence of fixed-size patches, treating each patch as a token, analogous to words in a sentence, as illustrated in Figure 2.14. These tokens are then projected and processed through multiple transformer layers, enabling the model to learn global image features effectively. The key innovation of ViT is its ability to capture long-range dependencies and global context within images, which are often challenging for traditional CNNs. These long-range dependencies in the 3D spatial domain can be crucial to segment target materials in the tomography reconstruction [76] as the materials are continuous and their shapes have regular patterns. This example of the Vision Transformer (ViT) model utilizes a contracting-expanding pattern, consisting of a stack of transformers as the encoder, which is connected to a decoder via skip connections. Similar to its use in Natural Language Processing (NLP), the transformers operate on a 1D sequence of input embeddings. To handle 3D input volumes, the input $x \in \mathbb{R}^{H\times W\times D\times C}$ with resolution (H, W, D) and C input channels is divided into flattened, uniform, non-overlapping patches $x_v \in \mathbb{R}^{N\times(P^3\cdot C)}$, where (P, P, P) denotes the resolution of each patch and $N = \frac{H\times W\times D}{P^3}$ is the length of the sequence.
Subsequently, a linear layer is used to project these patches into a K-dimensional embedding space, which remains constant throughout the transformer layers. To preserve the spatial information of the extracted patches, a 1D learnable positional embedding $E_{\text{pos}} \in \mathbb{R}^{N\times K}$ is added to the projected patch embeddings, resulting in:

$$z_0 = [x_1^v E; x_2^v E; \dots; x_N^v E] + E_{\text{pos}} \tag{2.36}$$

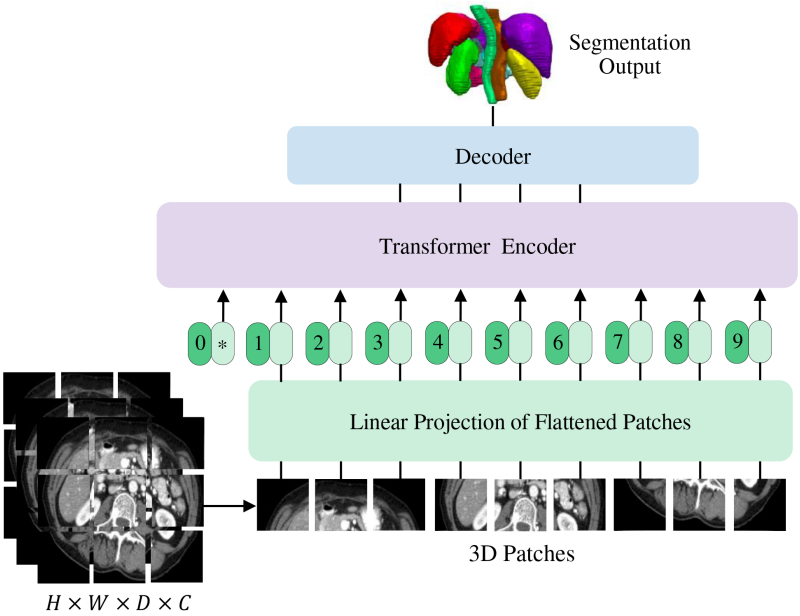


Figure 2.14: Overview vision transformer model [76] used in 3D medical image segmentation. It consists of a transformer encoder that directly utilizes 3D patches as tokens and then is connected to a CNN-based decoder via skip connections.

Then, just like in a standard transformer block, the patch embeddings can be encoded by the transformer encoder which comprises of multiple layers, each with Multi-Head Self-Attention (MSA) and Multilayer Perceptron (MLP) sublayers. The decoder processes the encoded features by progressively upsampling them to reconstruct the original resolution of the input. It utilizes skip connections to merge features from the encoder, ensuring that both local and global information is preserved. Each upsampling step is followed by convolutional layers to refine the features. The final output is produced by a 1×1×1 convolutional layer with a softmax activation function to generate voxel-wise semantic predictions.

Overall, the Vision Transformer model integrates the strengths of transformers and convolutional neural networks to process 3D volumetric data. The encoder uses transformers to capture long-range dependencies and context in the input volume, while the decoder uses CNNs and skip connections to reconstruct high-resolution output, enabling precise 3D semantic segmentation.

Chapter 3

AnACor1.0: ray-tracing analytical absorption corrections for long-wavelength crystallography

The content of this chapter is published in:

Lu, Y., Duman, R., Beilsten-Edmands, J., Winter, G., Basham, M., Evans, G., Kamps, J.J., Orville, A.M., Kwong, H.S., Beis, K., and Armour, W., 2024. Ray-tracing analytical absorption correction for X-ray crystallography based on tomographic reconstructions. *Journal of Applied Crystallography*, **57**(3), 649–658. https://doi.org/10.1107/S1600576724002243.

3.1 Introduction

In X-ray crystallography, intensities of reflections are proportional to the square of their structure factor amplitudes ($I_{\mathbf{h}} \propto |F_{\mathbf{h}}|^2$). Several factors need to be considered when calculating structure factor amplitudes from measured intensities, such as Lorentz, polarisation, sample illumination, decay and absorption corrections [77]. Away from absorption edges, sample absorption is approximately proportional to the cube of the wavelength [78]. It depends on the chemical composition, density, and the shape and size of the sample, which includes the crystal, as well as the surrounding materials like the sample mount and, mother liquor used to mount the crystals. High-quality structure determination relies on accurate structure factor amplitudes. The observed intensity depends both on the intrinsic scattering strength ($|F_{\mathbf{h}}|^2$) and the attenuation of the X-ray beam along the incident and diffracted paths (modeled by $A_{\mathbf{h}}$). Consequently, the measured intensity I_{meas}

is proportional to the product of the absorption correction factor and the squared structure factor amplitude, following $I_{\text{meas}} \propto A_h|F_h|^2$. Hence, correcting the measured intensities by calculating absorption correction factors is critical. For a crystal which is not surrounded by mother liquor or mounted in a loop, the Bragg intensities after absorption correction are given by $I_{\text{corr}} = I_{\text{meas}}/A_{\mathbf{h}}$. The absorption correction factor $A_{\mathbf{h}}$ for the reflection **h** in a crystallography experiment is given by

$$A_{\mathbf{h}} = \frac{1}{V}\int_V e^{-\mu(L_1(x,y,z)+L_2(x,y,z))}\, dV \tag{3.1}$$

where $L_1(x, y, z)$ and $L_2(x, y, z)$ (hereon referred to as L_1 and L_2) are the incident and diffracted X-ray path lengths for each crystal element dV, and μ is the absorption coefficient of the crystal [9]. Since the resulting volumetric integral calculation is intractable for irregularly shaped crystals, absorption correction for multi-faced crystals has been performed by numerical methods [79, 80]. As an alternative approach, the crystal can be partitioned into fundamental tetrahedra to calculate the integral over all the tetrahedra [39, 40, 81]. Both analytical and numerical absorption corrections require an accurate description of the shape and dimensions of the crystal. One solution from the APEX 3 software [82] is to determine and index all the crystal faces visually and perform an analytical absorption correction. However, this is difficult when the shape of the crystal is not a regular polyhedron. In addition, the presence of other materials surrounding the crystal, such as mother liquor and sample mount, adds further complication: these materials with different absorption coefficients only contribute to the absorption effect but not to the diffraction. Semi-empirical methods [83, 84] based on intensity measurements and assumptions about the incident and diffracted beams do not rely on the knowledge of the sample shape. However, they require multi-axis goniometers and the additional data needed for the azimuthal scans can contribute significantly to radiation damage on modern synchrotron light sources. Empirical methods which are independent of the sample geometry were developed either based on Fourier series of the incident and diffracted beams [85, 86] or by using spherical harmonics [49] to minimise the residual between the

intensities for symmetry-related reflections. With the introduction of large area detectors, these numerical methods to obtain an empirical correction for absorption have become popular and spherical harmonics is now the basis for absorption correction in most data reduction software packages for macromolecular crystallography, such as AIMLESS [21], hkl3000 [50], SADABS [51], and DIALS [16, 22], while XDS uses alternative numerical methods without spherical harmonics [14]. However, the efficacy of empirical methods depends on the number of symmetry-equivalent reflections, which can be difficult to achieve when data multiplicity is low, e.g. in the case of radiation-sensitive crystals in low-symmetry space groups.

As the analytical absorption correction does not depend on refining parameters to minimise differences between structure factor amplitudes of symmetry-related reflections, its success does not rely on data multiplicity. To analytically calculate absorption correction factors for a sample with irregular shape, its shape and orientation has to be characterised in detail. Previous work using optical microscopy to reconstruct a three-dimensional model of the sample containing crystal, sample mount, and mother liquor showed that absorption correction was viable and advantageous at lower levels of data multiplicity [52, 53]. An alternative approach to obtain a 3D model of the sample is X-ray tomography, which has been applied to either characterise or visualise crystals [54, 55]. The use of tomographic reconstructions and segmentations as a basis for absorption correction has previously been suggested by Brockhauser *et al.* [56]. This enables the calculation of X-ray path lengths through the different materials in the sample (crystal, sample mount and mother liquor), as illustrated in Figure 3.1.

While X-ray absorption is not normally considered an issue at standard wavelengths in MX, it is a major limiting factor in long-wavelength crystallography. Beamline I23 at Diamond Light Source, UK [6], is a unique synchrotron instrument operating in a wavelength range between 1.1 and 5.9 Å, giving access to the absorption edges of several light elements of biological significance, such as calcium, potassium, chlorine, sulfur and phosphorus. The largest anomalous signal for sulfur is expected close to its absorption

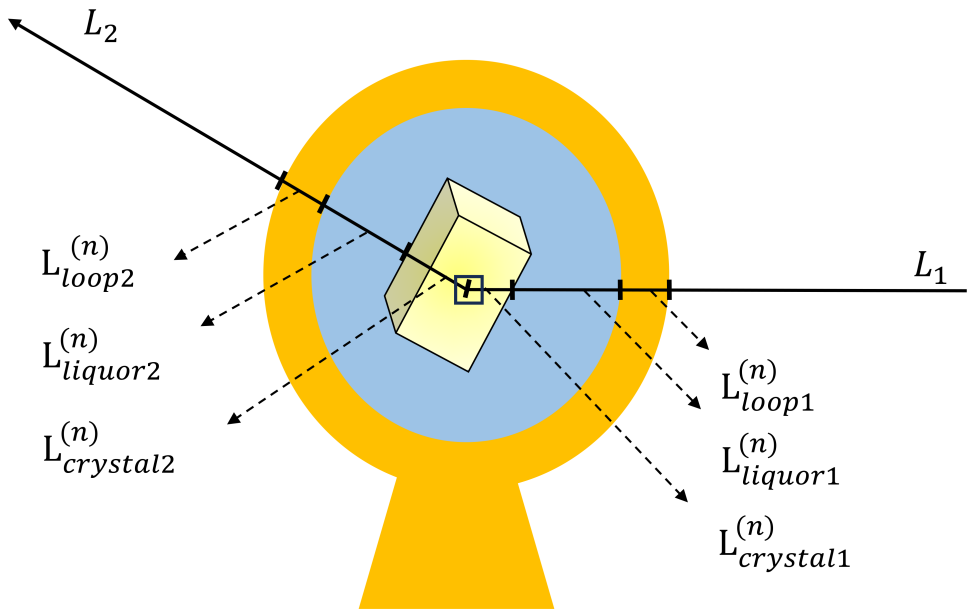


Figure 3.1: A sketch illustrating the ray-tracing method used to calculate an absorption correction factor for a crystal voxel n. $L^{(n)}_{m1}$ and $L^{(n)}_{m2}$ represent the path lengths of the incident and diffracted X-ray beams through the material m (loop, liquor and crystal)

edge (λ = 5.02 Å). However, the difficulties in correcting for increased sample absorption at very long wavelengths compromise the overall data quality, resulting in a reduction in the measured anomalous signal. Applying standard absorption correction protocols, the optimal wavelength for single-wavelength anomalous diffraction experiments based on sulfur (S-SAD) is found to be λ = 2.75 Å [5], clearly indicating the need for more sophisticated methods to exploit the full potential of long-wavelength crystallography.

In this chapter, AnACor1.0 is introduced, a computer program designed to calculate absorption correction factors for long-wavelength X-ray diffraction data. It employs a ray-tracing method to calculate the path lengths of the incident and diffracted X-rays through the sample, based on a tomographic reconstruction. The effectiveness of AnACor1.0 is demonstrated for long-wavelength datasets collected at 3.54 Å, on a crystal of the membrane protein OmpK36 GD, and at 4.13 Å, on a crystal of the heme-binding enzyme chlorite dismutase (Cld). OmpK36 GD, from here referred to as simply 'OmpK36', is a 373 amino acid outer membrane porin from *Klebsiella pneumonia* involved in nutrient

and antibiotic diffusion in gram negative bacteria [87], while Cld is a heme-*b* containing homodimeric oxidoreductase from *Cyanothece sp.* PCC7425, consisting of 181 amino acids per monomer. The choice of these two samples for this chapter was motivated by their crystallization in low-symmetry space groups, posing a challenge for the conventional absorption correction methods used in standard X-ray diffraction scaling programs.

3.2 Methodology

3.2.1 Experiment workflow and data preparation

The sample crystallization and the diffraction and tomography experiments are finished by our collaborators, Ramona Duman and Armin Wagner, at the long-wavelength MX beamline I23 at Diamond Light Source, UK. Crystals of OmpK36 were prepared and cryo-protected as previously described with no modification [87]. OmpK36 crystalized as rods in space group *C*2, with three monomers present in the asymmetric unit. Large sample-to-sample variations required extensive screening of crystals. The crystal of OmpK36 selected for this chapter had dimensions of 260 x 30 x 30 μm^3.

Cld crystals were produced using a protocol based on previously reported conditions [88] with modifications. Large Cld crystals (200 – 600 μm, diamond-shaped morphology) were obtained using sitting drop vapour diffusion (24 well plate, CrysChem 24-well, Hampton Research) at 25 °C. Cld protein (3 μL, 12 mg/mL), in 10 mm HEPES buffer pH 7.4, was mixed in a 1:1 ratio with crystallization solution (150 mm Mgso4, 100 mM MES pH 6.5 and 20 w/v% PEG 3350) and equilibrated against crystallization solution (350 μL). Crystals were obtained over several days (4-7) and allowed to mature. Cld crystals (5 × 200 – 600 μm) selected for the preparation of seed stocks were transferred into crystallization solution (100 μL; Eppendorf tube 1.5 mL) and crushed using glass seed beads (4 ×, Hampton Research) over multiple (10 ×) vortex agitation (30s) and cooling on ice (4 °C, 30s) cycles. Cld crystals (50 – 200 μm) were obtained using sitting drop vapour diffusion at 25 °C. Cld protein (3 μL, 12 mg/mL) in 10 mM HEPES buffer, pH 7.4, was mixed in a 1:1 ratio with

crystallization solution (150 mM MgSO4, 100 mM MES, pH 6.5 and 20 w/v%PEG 3350) and dilute seed stock (0.5 µL) and equilibrated against crystallization solution (350 µL). Cld crystals were matured over 24 h. Cryo-protection was performed by first transferring into glycerol solution (10 v/v% glycerol, 90 v/v% crystallization solution, 30s), then into NaCl solution (30 v/v% /150mM final NaCl, 20 v/v% glycerol and 50 v/v% crystallization solution), before plunge-freezing into liquid nitrogen. The crystal of Cld used in this chapter had dimensions of 190 x 150 x 90 μm^3. and indexed in space group *P*1, with two monomers in the asymmetric unit.

The in-vacuum sample environment comprises the cylindrical P12M detector and a multi-axis goniometer to enable collection of complete diffraction data from crystals in low-symmetry space groups even at the longest wavelengths. A tomography camera is integrated into the beamline sample environment allowing easy transition between the two experimental modes [62]. The sample preparation for in-vacuum data collection followed the standard protocol for beamline I23 [89]. For the OmpK36 crystal, 3 x 360° of data were collected at a wavelength of λ = 3.54 Å with 0.1s exposure/0.1° rotation angle and a beam transmission of 50%, with a top-hat X-ray beam adjusted to 240 x 150 μm^2.

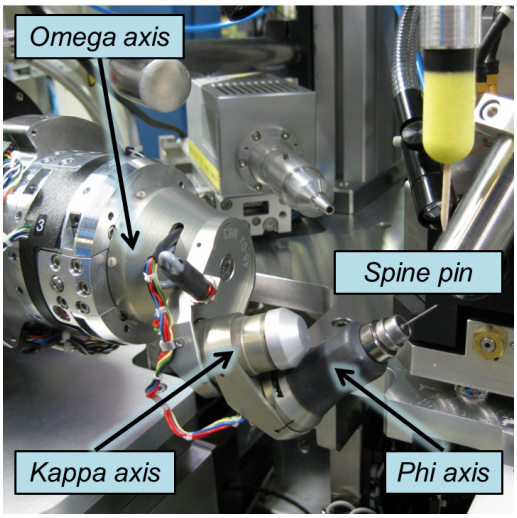


Figure 3.2: A schematic illustration of the kappa goniometer at Beamline I04 at Diamond Light Source.

In a kappa goniometer setup, the crystal is mounted on a system with three rotational axes:

omega (ω), kappa (κ), and phi (ϕ). The additional κ axis, inclined at an angle relative to the ω axis, enables reorientation of the crystal without changing the mounting. By adjusting the κ and ϕ angles, a wider range of reciprocal space can be sampled, improving completeness and data redundancy. A schematic illustration of the goniometer axes is provided in Figure 3.2. To ensure completeness of the data, two of the three datasets were collected using κ goniometry, with the κ axis rotated to -70° and the ϕ axis positioned at 0° and -120°, respectively. Each of the three datasets was measured with a photon flux of 1.36 x 10^{11} photons/s, which resulted in a total absorbed dose of 6.5 MGy/dataset, as calculated by Raddose3D [90]. Since the Cld crystal diffracted to a higher resolution than the OmpK36 crystal, it chose a low-dose data collection strategy. In total 22 x 360° were collected at a wavelength of λ = 4.13 Å with a 350 x 350 μm^2. top-hat beam, using an exposure of 0.1s/0.1°. With a beam transmission of 5%, the measured photon flux of 6.7 x 10^9 photons/s yielded an absorbed dose of 0.1 MGy/dataset. Two of the 22 datasets were collected with the κ and ϕ goniometer axes at 0°, while the rest were recorded at $\kappa = -70°$ and twenty different phi values, between -120° and 120°. The diffraction data was indexed and integrated with DIALS [16], providing a kappa/phi orientation matrix, raw intensities, incident vectors, scattering vectors and goniometer angles.

The diffraction experiment was immediately followed by tomography data collection at the same X-ray wavelength. One 180° tomography dataset was collected for each crystal, with the kappa and phi axes set at 0° and a beam size of 700 x 700 μm^2. and 100% transmission, using a propagation distance of 4.9 mm between scintillator and sample. For OmpK36 1800 projections, 30 flat-field images (without sample) and 30 dark images (without X-rays) were collected with an exposure of 0.15 s/0.1° rotation. The measured flux for this dataset was 1.5 x 10^{12} photons/s, resulting in a total absorbed dose of 4.8 MGy, as calculated by Raddose3D [90]. For the Cld crystal, 900 projections, 20 flat-field and dark images were collected with an exposure of 0.28 s/0.2° rotation, a measured flux of 4.3 x 10^{11} photons/s, yielding a total absorbed dose of 0.8 MGy.

The tomography data was processed using the SAVU pipeline [91], with a processing rou-

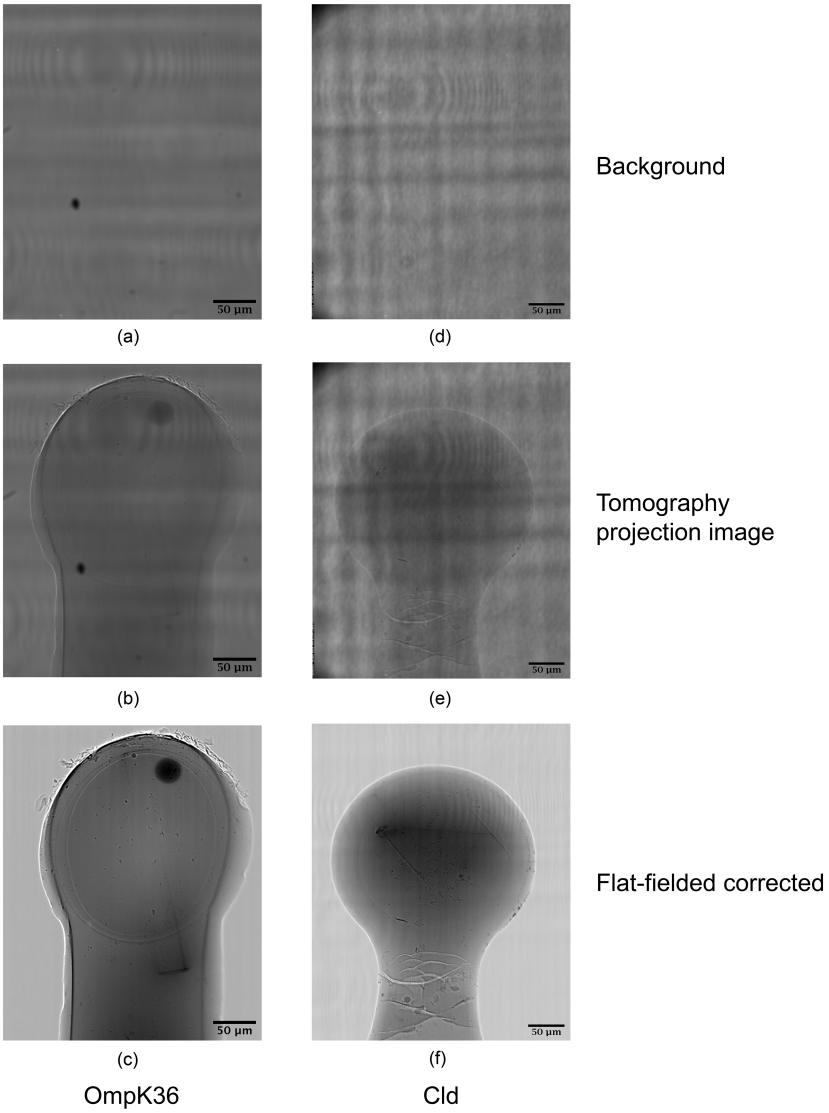


Figure 3.3: Tomography projection images for background ((a) and (d)), sample before ((b) and (e)) and after ((c) and (f)) flat-fielded correction of OmpK36 and Cld samples.

tine consisting of standard flat-field correction, followed by ring artefact removal [63] and reconstruction. For OmpK36, the reconstruction step was performed by iterative methods via the ToMoBAR module in SAVU [62], as its edge-enhancing properties gave improved results. For Cld, where the data showed better contrast, the filter-back projection (TomoPy) module [92] was used instead. No contrast transfer function correction was applied in the processing. Flat-field images, raw projections and flat-field corrected projections for both samples are shown in Figure 3.3. For ease of segmentation, reconstruction was performed on cropped data, to eliminate as much of the background as possible and reduce the size

of the images. The OmpK36 data was cropped from an initial volume of 1600 x 1200 x 1200 voxels to 1220 x 1001 x 1001 voxels, while the Cld data was reduced to 1310 x 1181 x 1181 voxels. The pixel size in the tomography images, determined from previous beamline calibrations, was 0.3 x 0.3 μm^2. Manual segmentation was performed with the visualisation software Avizo (Thermo Fisher), providing a 3D model with every voxel annotated as one of the different sample materials. Based on the sample 3D models, the absorption correction factors were calculated and exported to the scaling module in DIALS [22] to further correct the diffraction intensities. The published structures, PDBID 6RCK [87] for OmpK36, and PDBID 5MAU [88] for Cld, were used as starting models for the Dimple pipeline (`http://ccp4.github.io/dimple/`). The '- - anode' option [93] was used to calculate anomalous difference Fourier maps and anomalous peak heights and the option '- - free-r-flags' in the Refmac refinement [94] step ensured the same R-free flags for all absorption correction strategies. The Crank2 phasing pipeline [95] was used for experimental phasing by single-wavelength anomalous diffraction (SAD) with identical input parameters for the different strategies: the AFRO and PRASA modules were chosen for the F_A estimation and substructure determination steps, with the later step using 4000 trials and resolution cutoffs of 2.7 Å for Cld and 3.4 Å for OmpK36. The AFRO and PRASA modules were chosen because AFRO provides a maximum likelihood estimation of F_A values, which is more statistically robust than simple Friedel difference methods, and PRASA uses a phase retrieval approach that is often more effective than direct methods (e.g., SHELXD) for substructure determination, particularly in challenging cases.

3.2.2 Analytical absorption correction

For the calculation of the absorption correction factors, the integral (Eq. 3.1) is calculated over the crystal volume [96] as the only source of X-ray diffraction. To move from the continuous integral in Eq. 3.1 to a discrete equation, it replaces crystal elements dV by crystal voxels ΔV from the tomographic reconstruction [52]. This allows substituting the integral over the volume V with a sum over the crystal voxels. Hence, the integral in Eq.

3.1 can be rewritten discretely as

$$A_{\mathbf{h}} = \frac{1}{N}\sum_{n=1}^{N} A_{\mathbf{h}}^{(n)} \tag{3.2}$$

where N is the number of the crystal voxels in the 3D model exposed to the X-ray beam, and n represents every single crystal that is bathed in X-ray. The sample in a crystallography experiment typically contains more than one material therefore the calculation of the absorption correction factor $A_{\mathbf{h}}^{(n)}$ for a crystal voxel can be rewritten as:

$$A_{\mathbf{h}}^{(n)} = \exp\left[-\sum_{m=1}^{M} \mu_m L_m^{(n)}\right] \tag{3.3}$$

where $L_m^{(n)}$ represents the sum of the incident path length $L_{m1}^{(n)}$ and the diffracted path length $L_{m2}^{(n)}$ through the material m as shown in Figure 3.1, M shows the total number of different types of materials.

The final squared structure factor amplitudes $|F_{\mathbf{h}}|^2$ are obtained after combining their absorption correction factors with the overall scale factor, Lorentz and polarization corrections, and other standard correction and scaling techniques.

3.2.3 Standard ray-tracing method

The standard ray-tracing approach assumes X-rays incident upon a crystal voxel n and subsequently undergoes diffraction at that voxel, and consists of two algorithms: traversal and length calculation. During each ray traversal along the incident and scattered X-ray directions, the voxels' coordinates and their related material labels are calculated and subsequently recorded. After finishing the traversal, the absorption factors can be calculated from the recorded information based on the length calculation algorithm.

The construction of the model involves stacking two-dimensional segmented slices of the tomographic reconstruction, resulting in a three-dimensional array that can be referred to as a cuboid with six planes. The traversal algorithm is inspired by the Fast Voxel Traversal Algorithm [97], which is used in voxel traversal through a 3D array. A two-dimensional case is considered to better illustrate the traversal algorithm in Figure 3.4, and it is easy to

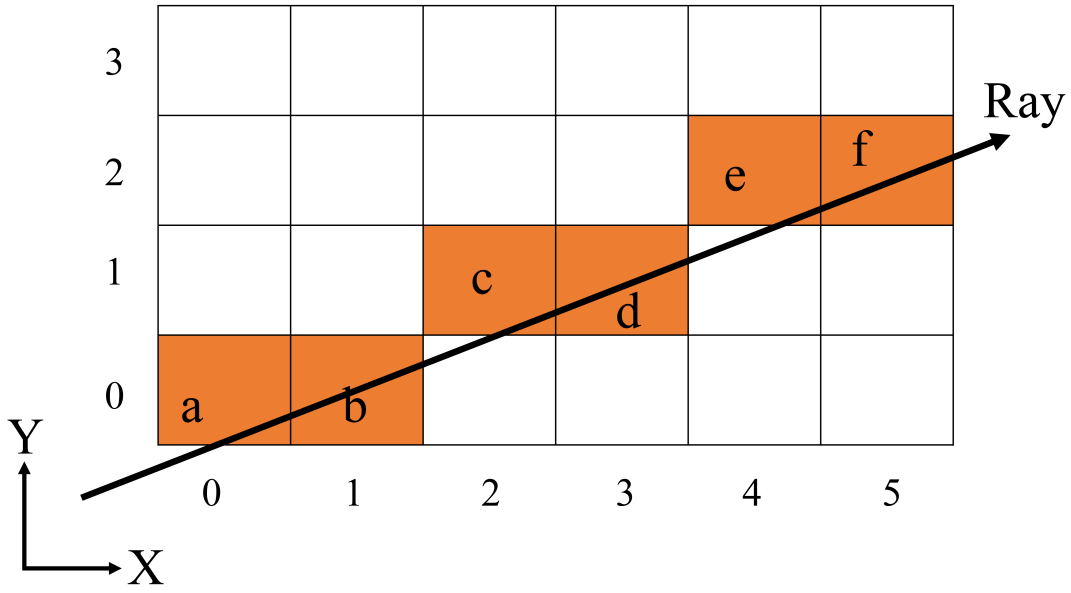


Figure 3.4: Schematic diagram of the ray-tracing traversal algorithm, illustrating a ray traversing from bottom left to top right across a 2D voxel grid. Starting at voxel a, the ray sequentially crosses voxels b, c, d, e, and f by incrementally advancing along the primary axis (here the X-axis) at fixed intervals. At each step, the nearest voxel to the ray path is selected and recorded based on rounded coordinates.

extend to three dimensions. A ray passes the pixels from bottom left to top right, which can be described by the equation $\mathbf{u} + \mathbf{v}s$, where $\mathbf{v}$ is the direction of the ray and $\mathbf{u}$ is a point on the ray.

In Figure 3.4 the ray can be divided into intervals of s along the X-axis, where each interval corresponds to one pixel. The coordinates of the traversed pixels b, c, d, e, and f can be determined by starting at pixel a ($\mathbf{u}_a = (0, 0)$) and moving sequentially and increasingly through the intervals ($s = 1, 2, 3, \ldots, n$, where n is the maximum interval, and $n = 5$ in Figure 3.4. the coordinates are rounded to the nearest pixel after each step and recorded for the later calculation. To calculate the path length, the starting and ending points are crucial, whereas the intermediate pixels along the ray are less significant. Only one pixel on the ray per interval is sufficient to calculate the path length. For this reason, the pixel below pixels c and e are not counted. Hence, in this example, the path length from a to f

is simply the Euclidean distance between the coordinates $(0, 0)$ and $(5, 2)$.

The determination of which axis to increment the intervals depends on the plane of the array intersected by the direction of the exit ray. This ensures that all calculated results remain within the array's bounds, allowing for accurate identification of each pixel traversed by the ray. For instance, if the *Y*-axis were chosen for interval incrementing in Figure 3.4, the interval sequence would be $s = 1, 2, 3$. As a result, the calculated coordinate for pixel b would coincide with point a, and this pattern would persist for the coordinates of (c, d) and (e,*f*). Consequently, this approach would yield incorrect path length calculations.

The traversal algorithm consists of two stages: initialization and incremental traversal. In the initialization stage, the exit point of the ray (such as pixel *f* in Figure 3.4) is determined. This step indicates which axis will be moved sequentially and whether the interval sequence will be ascending ($s = 1, 2, 3, \dots, n$) or descending ($s = n, \dots, 3, 2, 1$). Unlike the standard Fast Voxel Traversal Algorithm [97], our algorithm considers increments or decrements along only one direction (either x or y).

In the incremental traversal stage, in Figure 3.4, the coordinate of the new pixel is computed by the ΔY and ΔX along y and x axes from the starting pixel (X,Y) with rounding to the nearest integers. The ΔX indicates how far along the ray it must move along the x-axis component in one step s. It's straightforward to see that ΔY is determined by multiplying the direction of the ray **v**. The basic loop of Figure 3.4 is outlined in Algorithm 1:

Algorithm 1 Basic loop of traversal algorithm in 2D example in Figure 3.4

```
while NewX ≤ MaxX do
    ΔX ← s
    NewX ← Round(X + ΔX)                    ▷ Pixel values are integers
    NewY ← Round(Y + ΔY · ΔX)
    s ← s + 1
    NewPixel(NewX, NewY)
end while
```

In 3D cases, the initialization in the traversal algorithm that aims to find the exit faces of the ray, becomes more difficult. A ray casting technique is used, as outlined in Equations 3.4 - 3.6.

$$t_i = \frac{\hat{n}_i \cdot (x_i - P_0)}{\hat{n}_i \cdot d} \quad \text{for } i = 1 \text{ to } 6 \tag{3.4}$$

$$t_{\text{min}} = \min\{t_i : t_i \geq 0, i = 1, 2, \ldots, 6\} \tag{3.5}$$

$$P = P_0 + t_{\text{min}} d \tag{3.6}$$

The cuboid model consists of six faces, which can be mathematically represented as six planes that extend to infinity. These planes are defined by the vertices of the cuboid within the vector space. The calculation in Equation 3.4 determines the distance t_i between the crystal coordinate and the intersection with the plane. This is done by utilizing the unit normal vector $\hat{n}_i$ of the plane, the vertex coordinates x_i on the plane, the directional vector d of the X-ray, and the crystal point P_0. The vector d intersects with points on all six faces within the infinite vector space. The minimal value of the non-negative t_i represents the location where plane i connects with the cuboid in the positive direction of the vector d. Overall, the exit coordinate can also be determined by Equation 3.6 and this identifies the specific face of the cuboid that intersects with the X-ray, which helps finish initialization. In the path length calculation algorithm for incident path length $L_{m1}^{(n)}$, the crystal voxel n is considered as the starting point for the X-ray traversal rather than the X-ray source, so the direction of the incoming vector is reversed and it originates from the crystal voxel n. The algorithm iterates until it encounters the boundary of the 3D model and the coordinates and the corresponding labelled materials are recorded during the traversal. When the iteration stops, the total path length is calculated as the Euclidean distance between the starting point and the last voxel where the iteration stops. As depicted in Figure 3.5, the coordinates recorded exhibit a zigzag pattern along the X-ray path due to the voxelization process. To mitigate this zigzag effect, the individual path lengths ($L_m^{(n)}$) for material m are determined by first calculating the proportion of the total path length that the material m occupies during traversal, and then multiplying this proportion by the total path length. This product is then combined with the voxel size and the corresponding absorption coefficients to obtain the final exponent $\mu_m L_m^{(n)}$ in Equation 3.3. Finally, the calculation of the absorption

factor A_h for the reflection **h** involves summing $A_h^{(n)}$ for all crystal voxels, as shown in Equation 3.2.

In a tomography reconstruction, there are air/vacuum regions outside of the crystal sample, which contribute a negligible amount to the absorption effect. It could be argued that by neglecting these regions, computational time could be saved. However, as illustrated in Fig 3.5, if the traversal chooses to stop at the vacuum/air region, the absorption caused by the liquor and the loop at the end of the ray will not be counted. This is why we have chosen to stop the traversal at the model's boundary instead of any vacuum/air region, trading computational efficiency for accuracy to reduce the impact of segmentation artefacts.

3.2.4 Absorption coefficients

Absorption coefficients are determined experimentally using the intensity values in the flat-field corrected tomograms (Figure 3.3, (c) and (f)) as estimates of the ratio between the incident and transmitted intensities based on Beer-Lambert's law:

$$\mu = -\frac{1}{x} \ln\left(\frac{I}{I_0}\right) \tag{3.7}$$

where I is the intensity of the X-ray beam after passing through the material, I_0 is the initial intensity of the X-ray beam before entering the material, μ is the absorption coefficient, and x is the penetration distance through the material. The distances x through each material required for the calculation are obtained from the 3D segmentation models. The 3D models of the OmpK36 and Cld samples in different orientations are presented in Figure 3.6.

The overall process of absorption coefficient determination is outlined in Figure 3.7. The initial step, **Auto-orientation**, is critical for determining the orientation offset, as the tomography reconstruction model may not inherently align with the real-world sample. When such misalignment occurs, orientation offsets must be applied to the directions of the scattering X-rays recorded during the experiment to accurately calculate the absorption factor. This ensures that the tomography reconstruction is accurately aligned with the physical sample. Given that the sample comprises of mother liquor, a mounting loop,

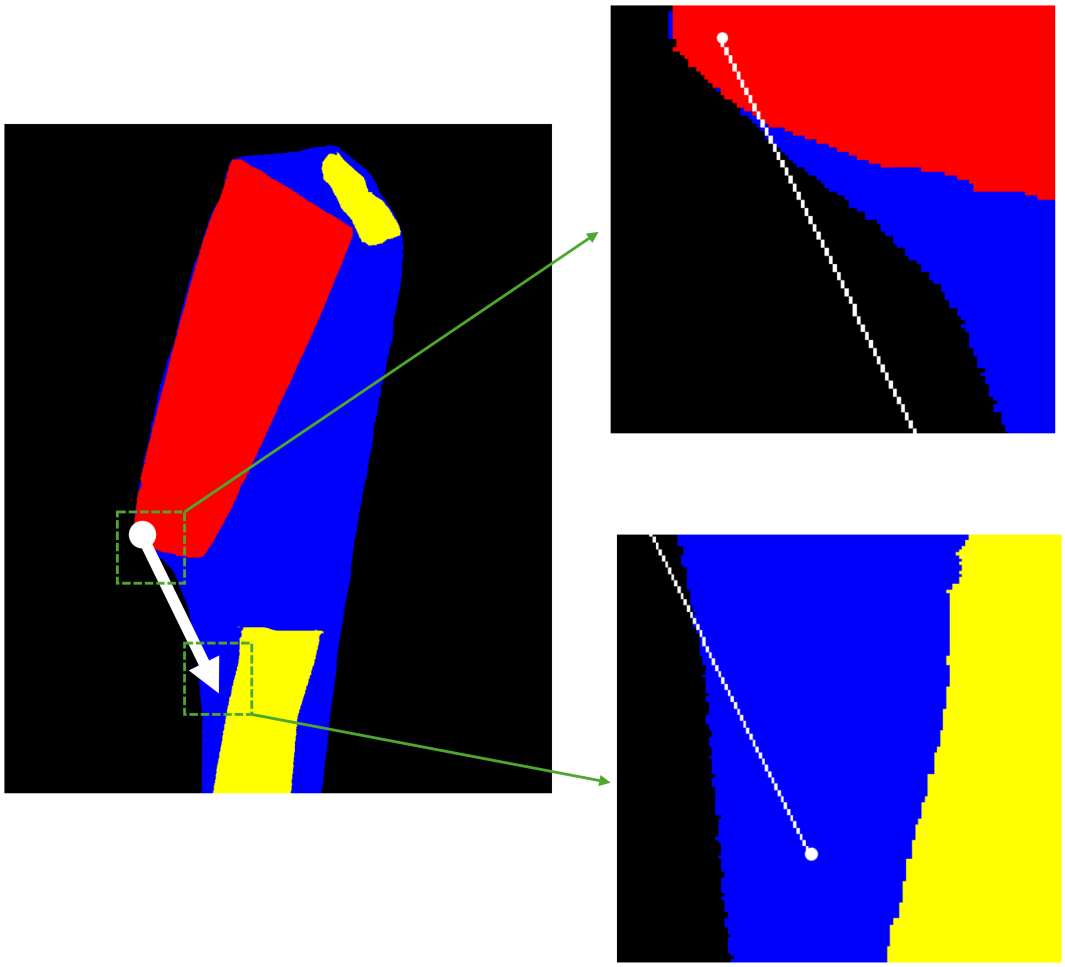


Figure 3.5: A ray-tracing path marked in white for a tomographic reconstruction slice of Thermolysin (Black: Vacuum; Red: Crystal; Yellow: Loop; Blue: Mother liquor). The diffracted path contains a large region of vacuum/air.

and a crystal, it is challenging to optimize the absorption coefficients for all materials simultaneously. Therefore, we first determine the absorption coefficients of the mother liquor, as it is more likely to have regions exclusively containing mother liquor. By applying Beer-Lambert's law, the absorption coefficients for the mother liquor can be obtained. The accuracy of the absorption coefficient determination can be influenced by both the signal-to-noise ratio (SNR) of the imaging system and the measurement strategy. Since the mother liquor typically exhibits the highest absorption among the materials present, its relatively strong attenuation provides a better signal compared to weakly absorbing

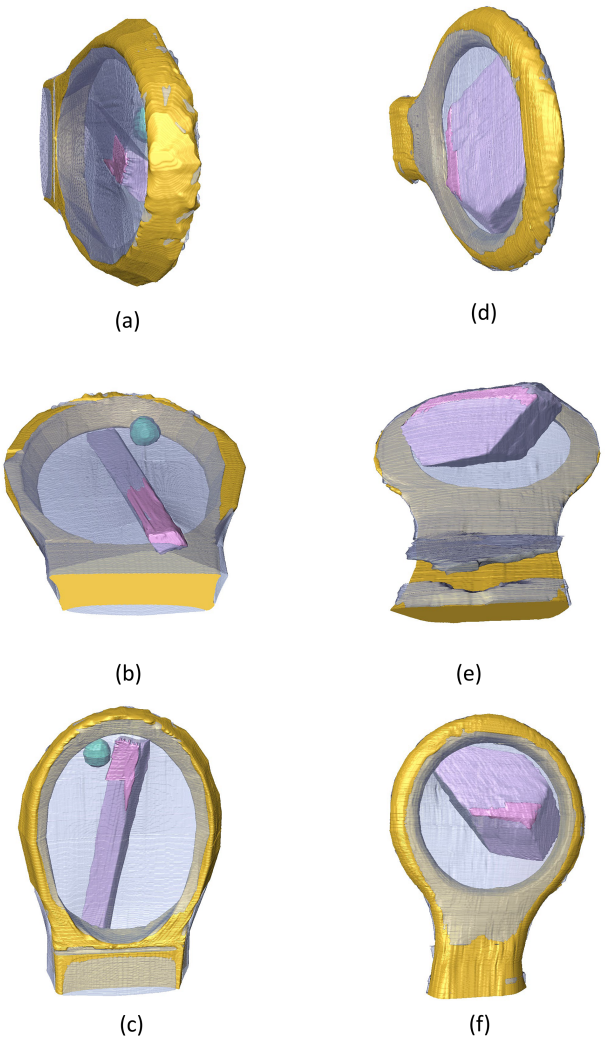


Figure 3.6: Volume renderings of segmentations of OmpK36 ((a) - (c)) and Cld (((d) - (f)). Transparent blue: mother liquor, gold: loop, pink: crystal, green: protein/detergent aggregate. The width of each view corresponds approximately to 400 µm.

components. To mitigate the effects of detector noise and imaging artifacts, absorption coefficients are estimated over multiple orientations and positions of the sample, and the median value is taken, which reduces the impact of local fluctuations and systematic errors. A comparison with theoretical absorption coefficients of the mounting loop and the crystal calculated using RADDOSE-3D shows that the percentage error is typically smaller than 10%, confirming the reliability and robustness of the estimation method.

The **Auto-viewing** step aims to find the goniometer angle at which the plane of the sample mounting loop is perpendicular to the incident X-ray beam. This orientation ensures that

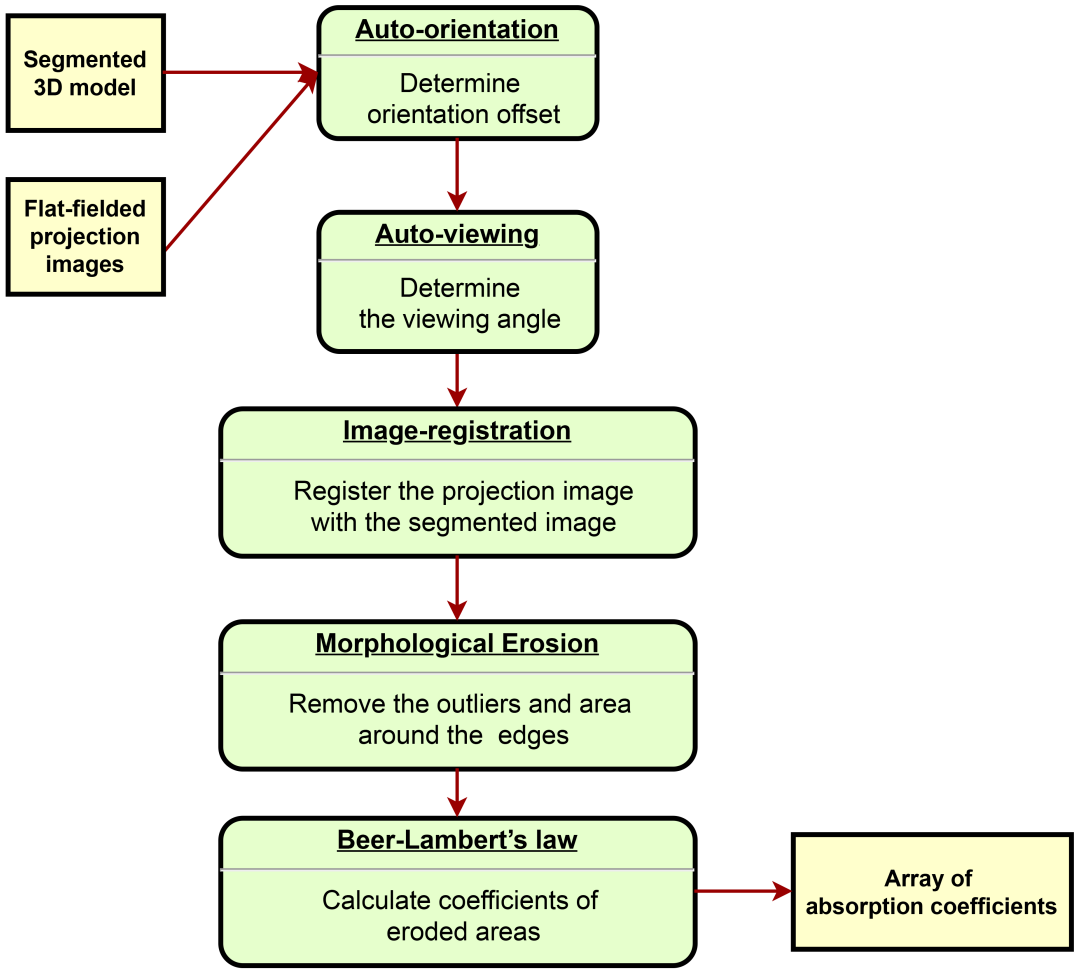


Figure 3.7: Flow chart of determining absorption coefficients. The inputs are the flat-fielded corrected projection images before tomography reconstruction and the segmented 3D model. The final output is an assembly of linear absorption coefficients based on Beer-Lambert's law

the X-rays pass through the mother liquor with minimal obstruction from the loop or crystal, facilitating a clearer measurement of the absorption of the mother liquor. As illustrated in Figure 3.6(c) and (f), maximizing the projected area of the sample can improve the identification of regions containing only mother liquor, which is essential for accurate absorption coefficient estimation. Additionally, a larger illuminated area provides more data points for the histogram analysis, thereby enhancing the precision and reliability of the absorption correction.

After aligning the orientation of the 3D model with the flat-fielded projection, the tomog-

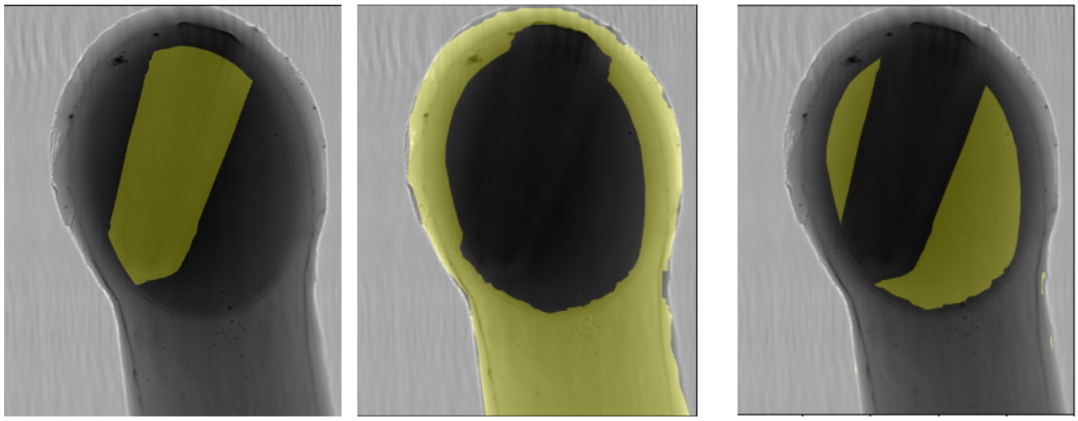


Figure 3.8: Aligned masks of different material on the tomography images of Thermolysin

raphy reconstruction process often results in a reconstructed sample that is centralized within the image, causing misalignment with the original projection data. To correct this, a **Image registration** step is necessary to align the positions of the 3D model projection with the sample region in the flat-fielded projection images. Intensity-based thresholding algorithms, including Triangle[98], Mean[99], Otsu[100], Li[101], Yen[102], Isodata[103], and Local[104], are employed to extract the sample region mask from the flat-fielded projection image. The Phase Cross-Correlation technique is then used to produce a shift vector to align this mask with the projection of the 3D model, which is illustrated in Figure 3.8.

Phase Cross-Correlation (PCC) is a technique used for determining the relative translational shift between two images. It operates in the Fourier domain and is based on the properties of the Fourier transform. Given the mask in the flat-field image $f(x, y)$ and the projection image of 3D model $g(x, y)$, which are related by a translation $(\Delta x, \Delta y)$, the relationship between their Fourier transforms $F(u, v)$ and $G(u, v)$ can be expressed as:

$$G(u, v) = F(u, v) \cdot e^{-2\pi i(u\Delta x + v\Delta y)} \tag{3.8}$$

Here, u and v are the frequency components in the Fourier domain, and Δx, Δy represent the translation in the spatial domain. The cross-power spectrum is defined as:

$$R(u,v) = \frac{F(u,v) \cdot G^*(u,v)}{|F(u,v) \cdot G^*(u,v)|} \tag{3.9}$$

where $G^*(u,v)$ is the complex conjugate of $G(u,v)$, and $|\cdot|$ denotes the magnitude. The cross-power spectrum normalizes the Fourier transforms, eliminating the amplitude information and retaining only the phase information. The inverse Fourier transform of the cross-power spectrum yields a function $r(x,y)$ that is a delta function centred at the displacement $(\Delta x, \Delta y)$:

$$r(x,y) = \mathcal{F}^{-1}\{R(u,v)\} \tag{3.10}$$

The peak of $r(x,y)$ indicates the translation vector $(\Delta x, \Delta y)$ between the two images. However, the effectiveness of this approach heavily depends on the similarity between the thresholded mask $f(x,y)$ and the projection of the 3D model $g(x,y)$. When these images are not alike, such as when the thresholding results in a mask with a significantly different area due to under- or over-thresholding, the phase cross-correlation may struggle to determine the correct translation vector. This occurs because the cross-power spectrum, which the method relies on, depends on the similarity of phase information between the images. Significant differences in content can cause the cross-correlation function to produce a broader or lower peak, leading to an inaccurate estimation of the shift vector. To mitigate this, multiple thresholding algorithms are evaluated, and the result with the smallest mean square error is selected after determining and applying the shift vector.

The crystal and mounting loop often have sharp contrast and strong scattering effects around the edges that introduce artefacts in the calculation of absorption coefficients. To reduce these artefacts, **Morphological Erosion** with a kernel size of 10×10 is applied to the regions of interest. This kernel size was empirically selected to effectively eliminate high-contrast edge artefacts after experiments with different kernel sizes, without excessively shrinking the internal region of each segmented phase. It balances removing noisy or unreliable boundary voxels and preserving enough volume for a robust estimation of absorption coefficients. After erosion, Beer-Lambert's Law is used to determine the distribution of

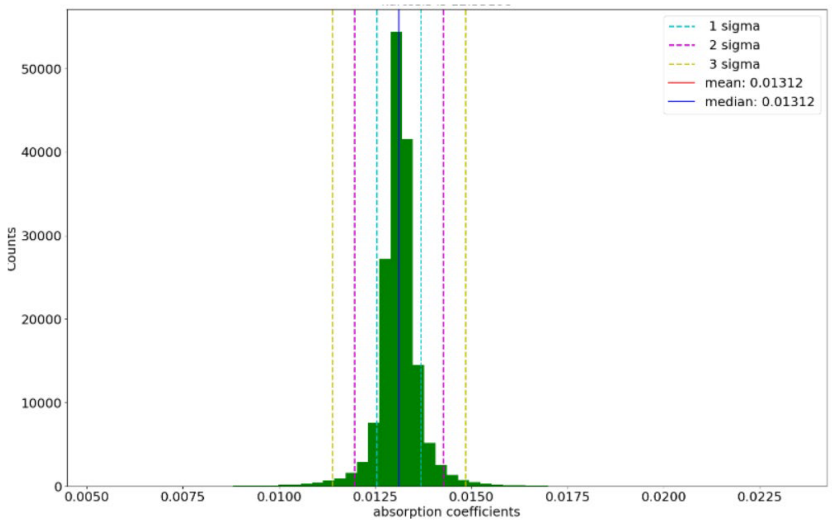


Figure 3.9: Histogram of absorption coefficients

absorption coefficients within these eroded regions. Segmentation artefacts can also affect the accuracy of the calculated absorption coefficients. To address this, only the longest 50% of linear path lengths through the materials are considered in the final calculation. The absorption coefficient is then determined as the median of the resulting distribution of these calculated coefficients, An example histogram of the estimated absorption coefficients is shown in Figure 3.9. Ideally, the histogram should resemble a normal distribution with a narrow spread, indicating consistency across measurements from different orientations and positions of the sample. If the absolute difference between the mean and the median exceeds 10%, a warning is raised, as this may indicate systematic errors or poor alignments.

Sample	Crystal	Mother Liquor	Loop	Protein/Detergent Aggregate
OmpK36	0.01053 (0.00977)	0.01208	0.00931	0.0322
Cld	0.0160 (0.0178)	0.01856	0.01724	N/A

Table 3.1: Linear absorption coefficients (μm^{-1}) of different materials in OmpK36 (λ = 3.54 Å) and Cld (λ = 4.13 Å) samples. The predicted results from Raddose are included in the bracket.

3.2.5 Implementation details

A standard ray-tracing method is applied to compute the path lengths $L_m^{(n)}$ for each crystal voxel n of the reflection **h** in Eq. 3.3. For a crystal voxel n, it assumes an incoming and a diffracted X-ray originating from the voxel. These X-rays, after applying the rotational matrix of the goniometer $\boldsymbol{R}$ to the reflection **h**, will propagate through the 3D segmented model. The coordinates of each voxel, along with its corresponding material label, are recorded. Then, the path lengths $L_m^{(n)}$ of material m can be determined by the distance between the coordinates of the boundaries of the materials. By combining the absorption coefficients of the corresponding materials, the absorption factor $A_{\mathbf{h}}^{(n)}$ for the crystal voxel n can be determined (Eq. 3.3). Finally, the total absorption factor $A_{\mathbf{h}}$ for the reflection **h** is calculated by summing $A_{\mathbf{h}}^{(n)}$ for all crystal voxels according to Eq. 3.2.

It is computationally intensive to rotate the overall 3D segmented model for each absorption factor calculation according to the rotational matrix of the goniometer $\boldsymbol{R}$. Instead, AnACor1.0 rotates the vectors of the incoming and diffracted beams to calculate the path lengths by inverting the goniometer matrix. The tomography experiments are always performed at kappa/phi orientations $\kappa = 0°$ and $\phi = 0°$. To correct data from diffraction experiments with varying kappa/phi orientations, it is essential to transform the vectors of both the incoming and diffracted beams with the kappa/phi orientation matrices $(\boldsymbol{R} \cdot \boldsymbol{R})^{-1}$ taken from the DIALS experiment model. Hence, the overall transformed vectors of these beams are in the form of $s_t = (\boldsymbol{R} \cdot \boldsymbol{R} \cdot \boldsymbol{R})^{-1} \cdot s_r$, where s_r is either the vector of the incoming or that of the diffracted beam taken from the DIALS reflection data. The resulting directional vectors s_t are used in the ray-tracing method. The X-ray beam is assumed to have a top-hat profile, meaning that its intensity is spatially uniform across the cross-section. Although the beam size is larger than the crystal sample, the absorption correction algorithm accounts for only the actual illuminated volume of the sample, ensuring that edge effects or contributions from outside the sample are not included in the correction. If the crystal is larger than the incident X-ray beam, a discriminator in the ray-tracing algorithm is used to determine whether a crystal voxel is inside the X-ray beam.

The absorption correction software AnACor1.0 is written in Python to facilitate future integration into DIALS [16]. In order to enhance computational efficiency, Numpy 1.23.2 [105] is used for data loading and pre-processing. Numba 0.56.2 [106] is used for JIT (just-in-time) compilation. A typical protein crystallography dataset contains hundreds of thousands of reflections. There are typically millions of crystal voxels in a 3D model, and each path length calculation can involve determining thousands of voxels along the incident and diffracted X-ray paths. Consequently, calculating all absorption correction factors for samples in protein crystallography is computationally expensive. To mitigate this, a systematic sampling method with a sampling interval of 2000 is applied. This value was empirically chosen to balance computational efficiency and accuracy, after experiments with different sampling intervals. For the two samples studied, smaller intervals resulted in only marginal improvements in precision while significantly increasing computational time. Further details and investigations are provided in a later section. This sampling approach relies on the sorted arrangement of the crystal voxels, which helps in identifying the subsections of the crystal where the path lengths (L_1 and L_2) are similar. By selecting every 2000^{th} voxel from this sorted list, it ensures that sampling is consistently applied across the crystal. Therefore, it can capture the essential characteristics of the sample with far fewer data points, maintaining accuracy in Eq. 3.2 calculations while reducing computational load.

Parallel computing is used by the built-in *multiprocessing* package in Python, and the calculations of all the reflections are evenly distributed to each CPU core. After applying sampling and parallel computing, on a cluster node with 48 CPU cores, the computational time for the analytical absorption correction of one dataset of OmpK36 and Cld is about 40 and 30 minutes, respectively, with the total RAM usage of around 200 GB.

3.2.6 Proof of correctness by tabulated results

To evaluate the accuracy of the ray-tracing method with and without tomographic volume sampling, the absorption factor calculations were compared with previously published

numerical solutions [10]. Three simulated shapes were considered: cubic, cylindrical, and spherical, consisting of crystal material only, presented in Fig. 3.10 (a) - (c). For consistency, a voxel size of 0.3 x 0.3x 0.3 μm^3 and the same sampling interval of 2000 were applied. Also, a smaller voxel size of 0.1 x 0.1x 0.1 μm^3 is also investigated.

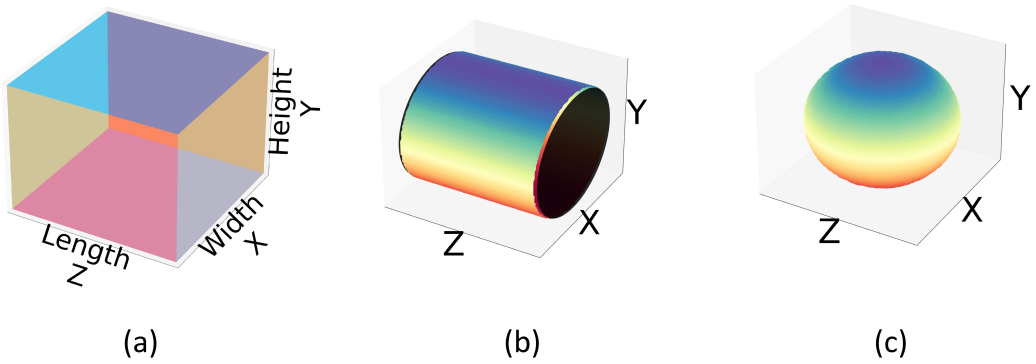


Figure 3.10: 3D visualization of simulated crystals in the shapes of a cube (a), cylinder (b), and sphere (c).

In the simulation, a single material with an absorption coefficient of 0.01 μm^{-1}, close to the absorption coefficient of the OmpK36 crystal, measured at 3.53 Å, is assumed. For the cube, the resultant integrals can be solved analytically while for the cylinder and sphere, the numerical solutions are taken from the International Tables [10]. The incident X-rays are simulated as going along the X-axis, and the diffraction angles are the angles between the X-axis and the diffracted X-rays, as described in the International Tables. Three different dimensions, similar to those of the crystals in this chapter, are used for each object.

- Cuboid:
  - (A): Length=100 µm, Width=100 µm, Height=50 µm
  - (B): Length=100 µm, Width=100 µm, Height=100 µm
  - (C): Length=100 µm, Width=100 µm, Height=150 µm
- Cylinder:
  - (A): Length=50 µm, Radius=10 µm
  - (B): Length=50 µm, Radius=50 µm

  - (C): Length=50 µm, Radius=100 µm

- Sphere:
  - (A): Radius=10 µm
  - (B): Radius=50 µm
  - (C): Radius=100 µm

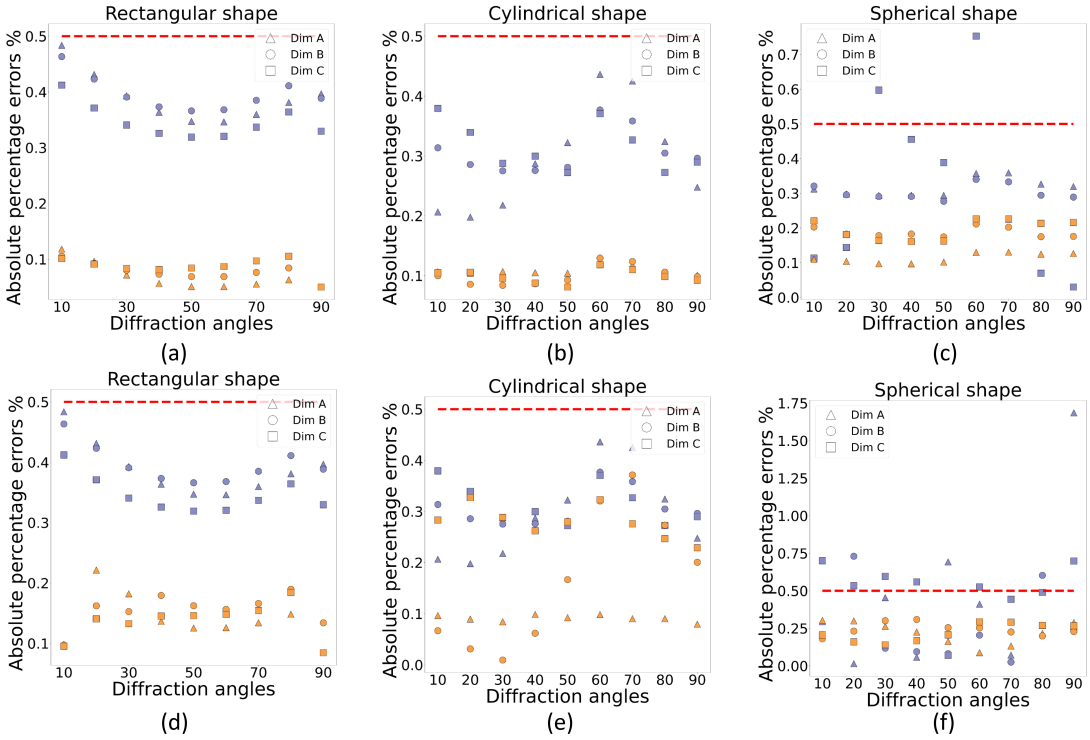


Figure 3.11: Absolute percentage differences between the absorption values from the International Tables [10] and those calculated by the ray-tracing method, shown for three shapes: rectangular, cylindrical, and spherical. The top row (a–c) shows results without tomographic sampling; the bottom row (d–f) shows results with tomographic volume sampling. Each plot compares two voxel sizes: 0.3 x 0.3 x 0.3 μm^3 (purple) and 0.1 x 0.1 x 0.1 μm^3 (orange). Dim A, B, and C represent the three principal dimensions of the shape. It shows that the absolute percentage errors between the theoretical absorption values and those computed by the ray-tracing method are mostly below 0.5%. The results indicate that larger voxel resolution (coarser sampling) leads to higher errors, while finer resolution improves accuracy more effectively than tomographic sampling alone. Percentage error is computed as $\frac{|\text{table}-\text{raytracing}|}{\text{table}} \times 100$.

Percentage errors between the values from numerical solutions and those from the ray-tracing method presented in this chapter are plotted in Fig. 3.11, with both no sampling (a-c) and sampling method employed here (d-f). It can be observed that without sampling, the general percentage errors for all three volumes are smaller than 0.5%. After applying

the sampling method, the overall percentage errors are still around 0.5%, except for some diffraction angles for the sphere. To evaluate the effect of the resolution of the 3D model, a smaller voxel size of 0.1 x 0.1 x 0.1 μm^3 is used (orange symbols in Fig. 3.11). To maintain the same number of crystal voxels that are involved in the calculation, the sampling ratio is decreased by a factor of 3^3 (sampling every 54000th coordinate). It is clear that the errors for all shapes of the non-sampling method with higher resolution are around 0.15%. With the sampling method, although the errors become slightly higher, they are still smaller than 0.5%. Also, the errors of higher resolutions are generally smaller than those of standard resolution (0.3 x 0.3 x 0.3 μm^3). As voxel sizes decrease, the percentage errors also become smaller, showing that the effects of sampling are less significant than the impacts resulting from the voxel size.

3.2.7 Absorption correction strategies

Data scaling is performed by the *dials.scale* program in DIALS [22] using the following custom scaling model:

$$g_{\mathbf{h}l} = C_{\mathbf{h}l} T_{\mathbf{h}l} S_{\mathbf{h}l} A_{\mathbf{h}l} \tag{3.11}$$

where $g_{\mathbf{h}l}$ is the overall inverse scale factor that needs to be determined for the l^{th} observation of symmetry-unique reflection **h**. The scale factors are determined by optimizing the scaling model parameters using a least-squares target function as previously described [22]. $C_{\mathbf{h}l}$, $T_{\mathbf{h}l}$, and $S_{\mathbf{h}l}$ are the scale term, the decay term and the spherical harmonics correction term of the default physical model. The absorption correction factors $A_{\mathbf{h}l}$ are precalculated by AnACor1.0 for each reflection $\mathbf{h}l$ and not optimised during the scaling process.

The scale term $C_{\mathbf{h}l}$ models intensity variations as a function of rotation while the decay term $T_{\mathbf{h}l}$ is a function of resolution and rotation. The spherical harmonics term S_{hl} is used to model anisotropic absorption effects that vary smoothly with the orientation of the crystal during data collection. Spherical harmonics provide a physically meaningful basis because they are capable of representing angular variations over the sphere, making them well-suited for describing the directional dependence of absorption. In this context, the

absorption anisotropy introduced by sample shape or mounting geometry is assumed to vary smoothly with rotation, and can be effectively approximated by a truncated series of spherical harmonic functions. The *absorption_level=high* option in *dials.scale* [107] was used for all approaches that included this term, which reduces the program's restraints on $S_{\mathbf{h}l}$ and uses six orders of spherical harmonic basis functions, to allow high and complex levels of absorption to be modelled. The *anomalous=False* option in *dials.scale* was used, as the low multiplicity of individual datasets was found to lead to unstable error model refinement for some datasets when the option *anomalous=True* was used.

To evaluate the analytical absorption correction by ray-tracing in AnACor1.0, four approaches are compared:

- No absorption correction (labelled as NO)

  $(g_{\mathbf{h}l} = C_{\mathbf{h}l}T_{\mathbf{h}l})$

- Spherical harmonics correction (default in *dials.scale*, SH)

  $(g_{\mathbf{h}l} = C_{\mathbf{h}l}T_{\mathbf{h}l}S_{\mathbf{h}l})$

- Analytical absorption correction described in this work (AC)

  $(g_{\mathbf{h}l} = C_{\mathbf{h}l}T_{\mathbf{h}l}A_{\mathbf{h}l})$

- Analytical absorption correction described in this work, combined with spherical harmonics correction (ACSH)

  $(g_{\mathbf{h}l} = C_{\mathbf{h}l}T_{\mathbf{h}l}S_{\mathbf{h}l}A_{\mathbf{h}l})$

It is important to note that in each approach the parameters for each part of the scaling model (except $A_{\mathbf{h}l}$) are jointly refined against the integrated intensities and therefore will be different in each approach i.e. $g_{\mathbf{h}l}^{ACSH} \neq g_{\mathbf{h}l}^{SH} \times A_{\mathbf{h}l}$. The combination of the analytical absorption correction with spherical harmonics allows the effect of absorption to be corrected by an accurate analytical model, while still enabling the spherical harmonics model to correct for any residual effects.

3.3 Results

In crystallography, various metrics, such as R-factors [34, 35, 36], correlation coefficients [108], and signal-to-noise ratios are used to evaluate data quality. Additionally, for long-wavelength crystallography peak heights in the phased anomalous difference Fourier maps are important quality indicators [109]. These metrics are used in combination with the success of experimental phasing by single-wavelength anomalous diffraction (SAD) to assess the three different absorption correction strategies and compare them with scaling without absorption correction.

Merging and refinement statistics (based on three datasets for OmpK36 and 22 for Cld) are presented in Table 3.2. As expected, for both samples, all four strategies result in similar resolution ranges, completeness and number of unique reflections. All three approaches to deal with absorption unsurprisingly lead to significant improvements in data quality over the data without correction.

For OmpK36, the analytical absorption correction (AC) gives equivalent merging R-factors to spherical harmonics correction (SH), with an overall R_{merge} of 0.119 for both. Notably, the AC strategy leads to an increase in the mean $I/\sigma(I)$, from 16.42 (SH) to 21.37 (AC) and a stronger anomalous signal, as measured by the anomalous slope (1.69 with AC, as opposed to 1.31 with SH). The anomalous slope [110] is the slope of the central region of a normal probability plot of anomalous differences: a slope greater than one indicates that the anomalous differences are larger than their uncertainties in aggregate. The combination of AC and SH corrections (ACSH) gives further improvements in the merging R-factors, signal-to-noise ratio, as well as the anomalous signal, with the R_{merge} decreasing to 0.105, the mean $I/\sigma(I)$ increasing to 24.92 and the anomalous slope increasing to 1.91. In Figure 3.12 (a), the anomalous peak heights from sulfur atoms for the three correction strategies are compared with no absorption correction for OmpK36. In total 12 sulfur atoms are found, from two methionine residues and two sulfates in the trimeric structure. A significant increase in peak heights is observed with all three absorption correction methods. AC generally gives better results than SH, with the exception of the heights of MET310 in

	No	SH	AC	ACSH
OmpK36 (λ = 3.54 Å)				
Merging statistics of 3 datasets				
Resolution range (Å)	107.4 - 2.34 (2.424 - 2.34)	107.4 - 2.34 (2.424 - 2.34)	107.4 - 2.34 (2.424 - 2.34)	107.4 - 2.34 (2.424 - 2.34)
Multiplicity	10.8 (5.5)	11.0 (5.5)	11.0 (5.5)	11.1 (5.5)
Completeness (%)	98.77 (91.67)	98.85 (92.15)	98.85 (92.12)	98.86 (92.15)
Mean $I/\sigma(I)$	11.99 (1.03)	16.42 (1.58)	21.37 (2.00)	24.92 (2.66)
R_{merge}	0.139 (0.473)	0.119 (0.419)	0.119 (0.458)	0.105 (0.427)
R_{meas}	0.146 (0.525)	0.125 (0.462)	0.125 (0.506)	0.110 (0.472)
R_{pim}	0.043 (0.214)	0.035 (0.185)	0.035 (0.204)	0.031 (0.191)
$CC_{\frac{1}{2}}$	0.996 (0.814)	0.997 (0.896)	0.997 (0.874)	0.998 (0.878)
$CC*$	0.999 (0.947)	0.999 (0.972)	0.999 (0.966)	0.999 (0.967)
Anomalous slope (d ≤3.9 Å)	1.13	1.31	1.69	1.91
Total reflections	654312 (31265)	668732 (31264)	668892 (31264)	672491 (31264)
Unique reflections	60652 (5606)	60652 (5634)	60652 (5633)	60652 (5634)
Refinement statistics				
Work set reflections	60585 (5605)	60631 (5634)	60630 (5632)	60639 (5634)
Free set reflections	3258 (328)	3260 (328)	3260 (328)	3260 (328)
R_{work}	0.219 (0.390)	0.207 (0.338)	0.203 (0.332)	0.199 (0.294)
R_{free}	0.255 (0.386)	0.244 (0.335)	0.240 (0.335)	0.235 (0.303)
PDB code	8QUR	8QUQ	8QVV	8QVS
Cld (λ = 4.13 Å)				
Merging statistics of 22 datasets				
Resolution range (Å)	46.67 - 2.7 (2.797 - 2.7)	46.67 - 2.7 (2.797 - 2.7)	46.67 - 2.7 (2.797 - 2.7)	46.67 - 2.7 (2.797 - 2.7)
Multiplicity	38.8 (23.5)	40.3 (23.5)	41.1 (23.5)	41.1 (23.5)
Completeness (%)	99.43 (97.97)	99.43 (97.97)	99.43 (97.97)	99.43 (97.97)
Mean $I/\sigma(I)$	16.51 (4.83)	20.22 (6.61)	37.43 (13.47)	44.73 (15.68)
R_{merge}	0.205 (0.281)	0.163 (0.240)	0.112 (0.197)	0.095 (0.183)
R_{meas}	0.208 (0.287)	0.165 (0.245)	0.113 (0.201)	0.096 (0.187)
R_{pim}	0.033 (0.056)	0.025 (0.048)	0.017 (0.039)	0.014 (0.037)
$CC_{\frac{1}{2}}$	0.997 (0.986)	0.997 (0.990)	0.999 (0.992)	0.999 (0.993)
$CC*$	0.999 (0.996)	0.999 (0.998)	1 (0.998)	1 (0.998)
Anomalous slope	1.28	1.36	2.48	2.50
Total reflections	531035 (31693)	551553 (31730)	562964 (31747)	563200 (31739)
Unique reflections	13696 (1351)	13696 (1351)	13696 (1351)	13696 (1351)
Refinement statistics				
Work set reflections	13696 (1351)	13696 (1351)	13696 (1351)	13696 (1351)
Free set reflections	686 (76)	686 (76)	686 (76)	686 (76)
R_{work}	0.191 (0.240)	0.176 (0.223)	0.172 (0.210)	0.172 (0.209)
R_{free}	0.234 (0.297)	0.223 (0.285)	0.218 (0.271)	0.218 (0.273)
PDB code	8QUV	8QUU	8QUZ	8QVB

Table 3.2: Merging and refinement statistics from OmpK36 and Cld. Columns represent the four absorption correction methods: spherical harmonics correction (SH), analytical absorption correction (AC), analytical absorption correction combined with spherical harmonics correction (ACSH), no absorption correction (No). Values in brackets are for the outer resolution shell. For the calculation of the anomalous slope, the resolution range is restricted to resolutions below which the anomalous signal is significant in the ACSH processed data, which is 3.9 Å for OmpK36 and the full resolution range for Cld.

chain B and SO4-1 in chain C, which are larger in the SH data. Overall, the ACSH strategy brings further improvements in peak heights, except for the weakest anomalous peak, SO4-2, where AC and ACSH perform similarly. The refinement statistics for all strategies follow a similar trend to the merging statistics, with R-factors being the lowest for ACSH.

SAD phasing was performed as a further test of the efficacy of analytical absorption corrections. Phasing was attempted with one, two out of three and all three datasets available. The results, summarised in Table 3.3, show that the ACSH strategy outperforms the others in requiring only two datasets for successful phasing despite the overall completeness of 89.2% and multiplicity of 8.3. Both AC and SH need all three datasets (98.9% overall completeness, multiplicity of 11.0), while the NO strategy is unsuccessful. The numbers of correct residues automatically built into the experimental electron density maps are identical between the three successful strategies, indicating that the quality of the maps is of similar standard and the lower data completeness used for the ACSH approach has no impact.

For Cld, the merging R-factors, $I/\sigma(I)$ and anomalous slopes are noticeably better for AC compared to SH. All merging statistics show further improvement for the combined ACSH correction. In contrast to OmpK36, where data quality indicators changed little between the SH and AC strategies, for Cld, the analytical absorption correction strategy (AC) gives substantially better data statistics compared to SH. For instance, in terms of the merging R-factors, it is observed a decrease of the R_{merge} from 0.163 with SH to 0.112 with AC and a further decrease to 0.095 with the ACSH treatment. There is also an increase in the overall mean $I/\sigma(I)$ from 20.22 for SH to 44.73 for the ACSH strategy with the high-resolution shell $I/\sigma(I)$ following this trend. The anomalous slope increases from 1.36 when using SH to 2.48 and 2.50 for AC and ACSH, respectively. This represents an improvement of approximately 82% in the anomalous signal when applying analytical absorption corrections compared to SH.

The anomalous peak heights for the different absorption correction strategies for Cld are shown in Figure 3.12 (b). In addition to three methionines and one cysteine per polypeptide

	Strategy	Number of datasets required for phasing	Completeness (overall/high resolution bin)	Multiplicity (overall/high resolution bin)	Refinement R-factor/R free	Number of correct residues automatically built / total number of residues
OmpK36	NO	-	-	-	-	-
	SH	3	98.8 / 88.1	11.0 / 3.9	0.235 / 0.280	1041 / 1041
	AC	3	98.8 / 88.1	11.0 / 3.9	0.227 / 0.274	1041 / 1041
	ACSH	2	89.2 / 71.9	8.3 / 3.3	0.228 / 0.280	1041 / 1041
	ACSH	3	98.8 / 88.1	11.0 / 3.9	0.218 / 0.257	1041 / 1041
Cld	NO	-	-	-	-	-
	SH	3	94.7 / 82.2	5.8 / 2.5	0.259 / 0.336	354 / 376
	AC	2	83.3 / 64.9	4.4 / 2.2	0.266 / 0.348	354 / 376
	AC	3	94.7 / 82.2	5.9 / 2.5	0.260 / 0.320	362 / 376
	ACSH	2	83.3 / 64.9	4.4 / 2.2	0.260 / 0.338	354 / 376
	ACSH	3	94.7 / 82.2	5.9 / 2.5	0.259 / 0.302	362 / 376

Table 3.3: SAD phasing results for OmpK36 (top) and Cld (bottom): statistics from Crank2 for all four absorption correction strategies. While for some strategies only two datasets were needed for successful phasing, the statistics from using three datasets are presented for comparison.

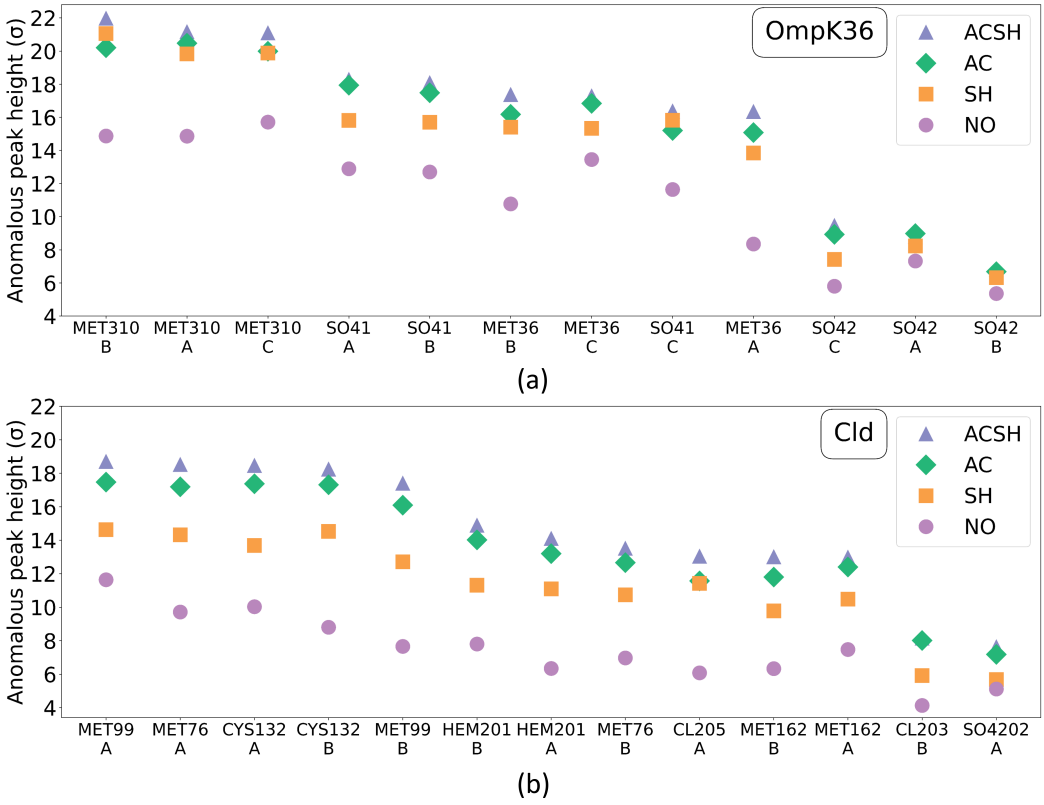


Figure 3.12: Peak heights (>5 σ) in the anomalous difference Fourier maps of anomalous scatterers in OmpK36 (a) and Cld (b) plotted in descending order of peak heights in the ACSH data, generated by ANODE. Raw data is presented in Table 3.4 of Ompk36, and Table 3.5 of Cld.

chain, each Cld monomer also binds an Fe-containing heme ligand and a Cl^- anion. A single SO_4^{2-} anion could be identified for the dimer, bringing the total number of anomalous scatterers to thirteen. SH leads to higher anomalous peak heights compared to no absorption correction. In line with the improved merging statistics, the anomalous signal in AC and ACSH is stronger than that in SH. ACSH gives the highest anomalous peak heights overall. While for OmpK36 the improvements in peak heights given by the AC and ACSH strategies over SH are quite modest, for Cld the increase from SH to AC/ACSH is more substantial. For the largest peaks, MET99 and CYS132, it is observed that increases in peak heights from 14 to 17 and 18 σ for AC and ACSH, respectively. The experimental phasing results for Cld (Table 3.3) show that both the AC and ACSH strategies

Atom	Residue	Chain	Peak heights			
			NO	SH	AC	ACSH
S	MET310	C	15.71	19.88	19.99	**21.09**
S	MET310	B	14.87	21.05	20.2	**21.97**
S	MET310	A	14.86	19.83	20.47	**21.16**
S	MET36	C	13.45	15.33	16.84	**17.28**
S	SO41	A	12.89	15.81	17.94	**18.28**
S	SO41	B	12.70	15.70	17.48	18.08
S	SO41	C	11.64	15.82	15.20	16.37
S	MET36	B	10.77	15.40	16.18	17.37
S	MET36	A	8.35	13.85	15.08	16.34
S	SO42	A	7.32	8.23	8.98	8.81
S	SO42	C	5.80	7.42	8.93	9.46
S	SO42	B	5.36	6.32	6.68	6.82

Table 3.4: Merged anomalous peak heights ($> 5\sigma$) for OmpK36 with various absorption corrections.

Atom	Residue	Chain	Peak heights			
			NO	SH	AC	ACSH
S	MET99	A	11.64	14.63	17.47	**18.70**
S	CYS132	A	10.03	13.68	17.37	**18.46**
S	MET76	A	9.71	14.32	17.19	**18.52**
S	CYS132	B	8.80	14.52	17.31	**18.24**
FE	HEM201	B	7.80	11.31	14.02	**14.89**
S	MET99	B	7.66	12.71	16.09	17.40
S	MET162	A	7.47	10.48	12.40	12.98
S	MET76	B	6.97	10.74	12.66	13.51
FE	HEM201	A	6.34	11.09	13.20	14.10
S	MET162	B	6.33	9.78	11.80	13.01
CL	CL205	A	6.08	11.42	11.57	13.04
S	SO4202	A	5.11	5.68	7.18	7.63
CL	CL203	B	-	5.91	8.01	8.14

Table 3.5: Merged anomalous peak heights ($> 5\sigma$) for Cld with various absorption corrections.

achieve successful phasing using only 2 out of the 22 available datasets, corresponding to an overall completeness of 83.3% and multiplicity of 4.4. In contrast, the SH strategy requires 3 datasets to achieve successful phasing, with a higher completeness of 94.7% and multiplicity of 5.8. This indicates that AC and ACSH can achieve accurate phasing

with lower completeness and multiplicity compared to SH. Notably, as shown in Table 3.2, when merging all 22 datasets, the overall multiplicities for AC and ACSH (both are 41.1) are higher than those for SH (40.3).

These results follow the same pattern seen with the data quality indicators discussed above, where the AC strategy outperforms the SH approach. Experimental phasing is unsuccessful for the Cld data with no absorption corrections, even after merging all 22 datasets.

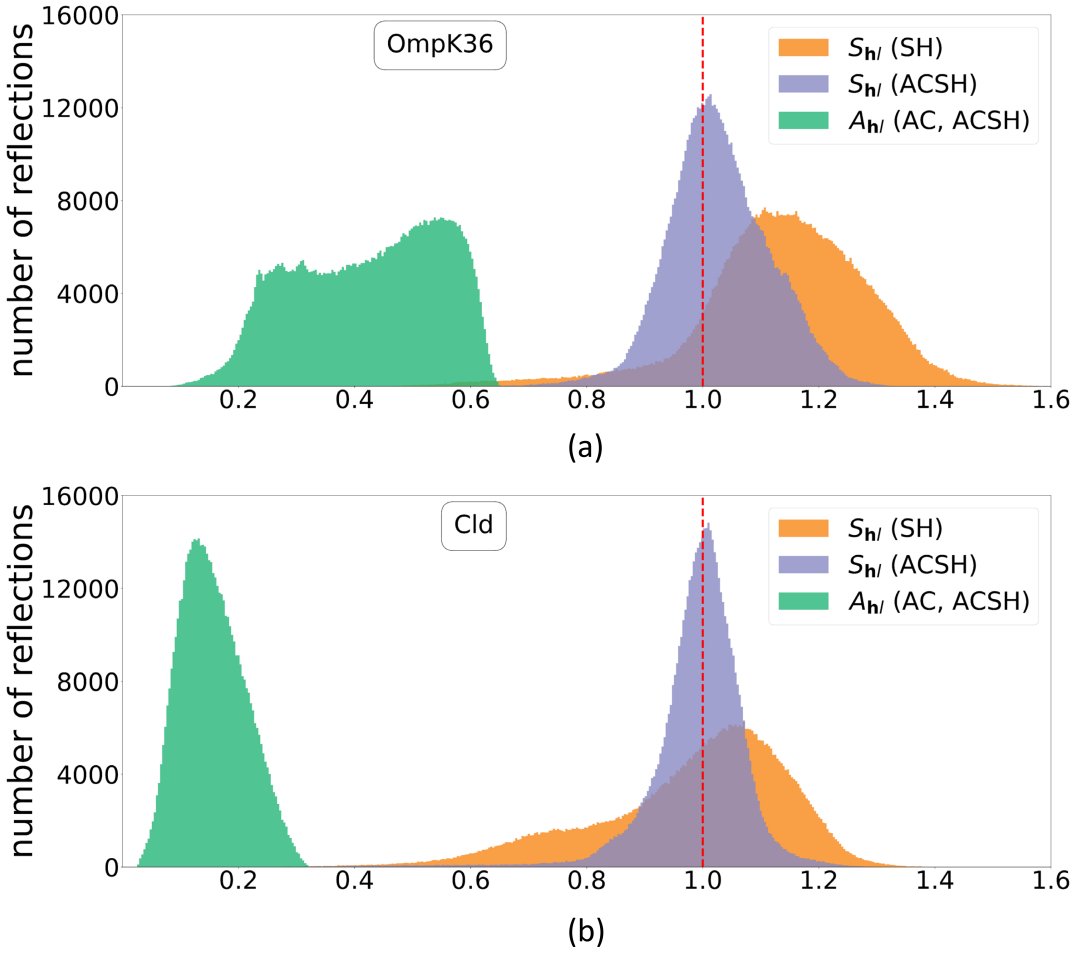


Figure 3.13: Histograms of absorption factors $A_{\mathbf{h}l}$ and spherical harmonics terms $S_{\mathbf{h}l}$ for OmpK36 (a) and Cld (b). $A_{\mathbf{h}l}$ (green) as used in AC and ACSH strategies are on an absolute scale, whereas $S_{\mathbf{h}l}$ for SH (orange) and ACSH (purple) are on a relative scale.

To illustrate the extent of the AC and SH corrections, histograms of the per-reflection analytical absorption correction factors ($A_{\mathbf{h}l}$) and spherical harmonics correction terms ($S_{\mathbf{h}l}$) are presented in Figure 3.13 for OmpK36 and Cld. For both datasets, when employing

the SH correction strategy, the resulting spherical harmonics terms ($S_{\mathbf{h}l}$) are distributed over a large range (0.5 - 1.5). When employing the ACSH strategy, the inclusion of the absorption correction factors ($A_{\mathbf{h}l}$) (shown on the right of Figure 3.13), leads to unimodal $S_{\mathbf{h}l}$ distributions over a narrower range (0.7 - 1.3) centred around 1. As the 'no correction value' for the SH model is $S_{\mathbf{h}l}$ = 1.0, fitting the additional spherical harmonics terms in the ACSH strategy results in further improvement in the internal consistency compared to AC alone, allowing correction for additional systematic effects present in the data.

3.4 Discussion and Conclusion

This chapter presented the fundamental building blocks for the whole thesis as it emphasizes the scientific impact of the analytical absorption correction on crystallography. However, clearly this implementation needs further improvements to be more practical for researchers. It demonstrates the successful application of analytical absorption corrections based on three-dimensional reconstructions from X-ray tomography implemented in AnACor1.0. In addition, it describes the algorithm for calculating the path lengths from 3D models by a standard ray-tracing method. Two very long wavelength experiments from crystals of the proteins OmpK36 and Cld indicate that this approach substantially improves the data quality and the success of experimental phasing compared to the standard scaling protocol based on spherical harmonics. Scaling without any absorption correction is presented as a control and unsurprisingly yields the poorest data quality statistics, and anomalous peak heights and for both samples experimental phasing is unsuccessful. This clearly indicates that data quality is severely affected by absorption effects, demonstrating the need for absorption corrections.

Data from OmpK36, which crystallises in the monoclinic space group *C*2 was collected at a wavelength of λ = 3.54 Å. A clear trend is visible, the analytical absorption correction (AC) is better than the spherical harmonics correction (SH) and the combination of both (ACSH) improves the data even further. While the overall improvements on statistics are small, the fact that the OmpK36 structure could be solved after ACSH correction using only 2/3 of

the data needed for the AC and SH strategies, clearly highlights the importance of such an improvement. For the Cld data ($P1$, $\lambda = 4.13$ Å) the same trend is observed. However, while the difference between AC and ACSH is small, they outperform the spherical harmonics correction. This is in particular reflected in the outcome from experimental phasing, two datasets are sufficient for both AC and ACSH, while three datasets are needed to solve the structure from data corrected by SH. In general, the combined approach of ACSH gives the best results for both samples/wavelengths, as it can model additional systematic effects present in the experimental data.

X-ray absorption increases with the cube of the wavelength, so a change from $\lambda = 1.0$ Å to $\lambda = 4.13$ Å leads to a 70-fold increase in absorption coefficients. The analytical absorption correction compensates for this increase, reflected in the narrow unimodal distribution of the resulting spherical harmonics terms $S_{\mathbf{h}l}$ centred around 1.0 in the two ACSH cases. Both samples used in this chapter crystallise in either monoclinic (OmpK36) or triclinic (Cld) space groups. This in combination with the asymmetry of the cylindrical P12M detector, with an aspect ratio of 2:1, leads to a low overall data multiplicity of five for OmpK36 and only three in the case of Cld, as well as poor data completeness for a single 360° dataset. In contrast to the spherical harmonics, the analytical absorption correction is not dependent on multiple observations, hence ideally suited for crystals in low-symmetry space groups or for radiation-sensitive crystals at long wavelengths.

AnACor1.0 is also able to correct data in multiple crystal orientations and for cases where the sample is larger than the beam. In this chapter, the segmented 3D model is obtained by X-ray tomography on beamline I23 at Diamond Light Source. AnACor1.0 can also be used for analytical absorption corrections for data from other sources, as long as a file with annotated voxels is provided and the relation between the coordinate systems of the 3D model and the diffraction experiment is known. This is suitable for highly absorbing samples in chemical crystallography.

However, AnACor1.0 has a lot of practical constraints, such as long computational running time and inefficient memory usage. For example, it takes about 30 minutes to process

one dataset of Cld, but there are 22 datasets for different Kappa-phi orientations to ensure enough data completeness, which is common in protein crystallography. In the next chapter, Chapter 4, the accelerated version AnACor2.0 is introduced to mitigate these constraints. In addition, the segmentation model is an inevitable part of determining analytical absorption factors. However, the time spent on manual segmentation can take from a few hours to as much as a day for an expert user to complete. To facilitate an automatic and efficient analytical absorption correction pipeline, in Chapter 5, a segmentation pipeline with our AI model AnACorNet is introduced.

Chapter 4

AnACor2.0: A GPU-accelerated open-source software package for analytical absorption corrections in X-ray crystallography

The content of this chapter is accepted with three typographical corrections but not yet in press:

Lu, Y., Adámek, K., Stefanic, T., Duman, R., Wagner, A., and Armour, W. (2024). AnACor2.0: A GPU-accelerated open-source software package for analytical absorption corrections in X-ray crystallography. *Journal of Applied Crystallography* **57**(6), 1984–1995.

4.1 Introduction

As we have seen, during the processing of X-ray crystallography diffraction data collection, the absorption effect is the dominant factor affecting the retrieval of accurate structure factors. The absorption effect is primarily determined by the crystal composition, its shape, and the wavelength of the X-ray beam. In macromolecular crystallography, various data reduction software packages, such as AIMLESS [21], hkl3000 [50], SADABS [82], and DIALS [16], employ spherical harmonics corrections to address absorption effects. This method typically relies on data multiplicity, which is unaffected by the sample's material and geometry. However, its effectiveness diminishes when it is difficult to obtain multiple data sets from different goniometer orientations, such as with radiation-sensitive crystals

in low-symmetry space groups. In such cases, the limited number of symmetry-equivalent reflections hampers the success of spherical harmonics correction. Introducing analytical absorption correction improves data-scaling quality in these scenarios because it doesn't rely on data multiplicity. AnACor1.0 described in Chapter 3 applies this method, which is based on Equation 4.1 below and utilises a 3D model of the sample.

$$A_{\mathbf{h}} = \frac{1}{V}\int_{V} e^{-\mu(L_1(x,y,z)+L_2(x,y,z))}\, dV \tag{4.1}$$

where $L_1(x, y, z)$ and $L_2(x, y, z)$ (hereon referred to as L_1 and L_2) are the incident and diffracted X-ray path lengths to and from each crystal element dV, μ is the absorption coefficient of the crystal and A_h is the inverse absorption factor (referred to as the absorption factor in the following context) [9]. For the implementation of analytical absorption correction on a voxelised 3D model, the integral in Equation 4.1 can be reformulated discretely:

$$A_{\mathbf{h}} = \frac{1}{N}\sum_{n=1}^{N} A_{\mathbf{h}}^{(n)} \tag{4.2}$$

where N is the number of the crystal voxels in the 3D model exposed to the X-ray beam. This is because the crystal volume [96] is the only thing that contributes to the X-ray diffraction. In a crystallographic experiment, it is common for the sample to consist of multiple materials. As a result, the determination of the absorption correction factor $A_h^{(n)}$ for a crystal voxel can be reformulated as follows:

$$A_{\mathbf{h}}^{(n)} = \exp\left[-\sum_{m=1}^{M} \mu_m L_m^{(n)}\right] \tag{4.3}$$

The symbol $L_m^{(n)}$ denotes the combined length of the incident length $L_{m1}^{(n)}$ and the diffracted length $L_{m2}^{(n)}$ as they pass through the material m being diffracted at the crystal voxel n. In AnACor1.0, the path lengths are determined by a ray-tracing method (described in Section 3.2.3), which need to traverse a large number of voxels in the 3D model to obtain accurate results. Macromolecular crystallography often requires examining thousands of reflections, with the number of crystal voxels in the 3D model N reaching into the millions.

This makes processing all reflections and voxels a significant computational obstacle for efficient absorption correction. In the previous chapter, AnACor1.0 performed analytical absorption correction by Python and Numba 0.56.2 [106] to enhance computational efficiency. It employed a Systematic sampling method with a 0.05% sampling ratio, reducing the processing time for a dataset to approximately 40 minutes. However, this is still too slow for quickly analyzing many datasets, especially for large samples with numerous crystal voxels. To mitigate these issues, this section presents AnACor2.0, an innovative software solution designed to streamline analytical absorption correction by incorporating new sophisticated computational strategies.

AnACor2.0 employs sampling methods to process fewer crystal voxels N while maintaining accuracy in the output. It also uses a bisection approach to enhance the standard ray-tracing method for determining path lengths. Instead of traversing every voxel along the diffracting ray, the Bisection method iteratively identifies the middle voxel's material to locate all material boundaries, which are used to calculate the path lengths. This significantly improves time complexity, which is crucial for large samples with many voxels. Additionally, AnACor2.0 includes a module that calculates absorption gridding maps for each crystal voxel n and uses interpolation techniques to determine the path length for a given direction of the diffracting ray. This approach reduces repetitive computation of diffracting rays from similar directions. One of the key characteristic modules of AnACor2.0 is its utilization of NVIDIA's CUDA platform for acceleration. This module utilises the capabilities of parallel computing on GPUs, a computational accelerator that enables concurrent calculations across many processing units. Through these approaches, AnACor2.0 can significantly reduce the amount of computational resources needed, along with computational time taken to produce results.

This chapter explores three standard experimental datasets: Insulin at 3.10 Å, Thermolysin at 3.53 Å, and Thaumatin at 4.13 Å, all in high-symmetry space groups. Insulin is spherical, Thaumatin is pyramidal, and Thermolysin is asymmetrical, demonstrating that AnACor2.0 is versatile and applicable to various sample shapes.

This chapter assesses the performance of different sampling methods and acceleration techniques on analytical absorption correction in these experiments. It employs analytical absorption correction followed by spherical harmonics correction (ACSH) mentioned in Chapter 3 for data-scaling, comparing absolute differences in absorption factors and analyzing relative variations in anomalous peak heights in the anomalous difference Fourier maps of the crystals.

All results are obtained using the Oxford ARC supercomputer [111], on a single node with an Intel Xeon Platinum 8268 CPU with 48 cores. we evaluated GPU performance on NVIDIA V100, A100, and H100 GPUs. Section 4.3 compares the computational time of the acceleration methods with the original baseline presented in AnACor1.0.

AnACor2.0 is publicly released `https://github.com/yishunlu-222/AnACor2.0.git` with GNU General Public License v3.0.

4.2 Methodology

4.2.1 Data preparation and implementation

In selecting the protein samples for this chapter, we aimed for morphological variation and obtaining crystals in high-symmetry space groups, since previously in the last Chapter 3 we had focused only on low-symmetry crystals. Also, high-symmetry crystals were chosen because they provide more symmetry-equivalent reflections, which increase redundancy and reduce the influence of random and systematic errors unrelated to absorption correction. This allows a clearer evaluation of the acceleration strategies, as the observed differences in data quality can more confidently be attributed to the acceleration method itself. The sample crystallization and the diffraction and tomography experiments were finished following a standard procedure [89] by our collaborators, Ramona Duman and Armin Wagner, at the long-wavelength MX beamline I23 at Diamond Light Source, UK.

Crystallisation was performed using the sitting drop vapour-diffusion method at 20°C in Swissci (UVXPO-2 lens) 96-well plates, by mixing 100 nL of protein solution with 100 nL

of crystallisation buffer. Crystals of thaumatin (Sigma, T7638) were obtained from a 50 mg/mL solution of the protein powder suspended in deionised water and a crystallisation buffer consisting of 100 mM ADA, pH 6.5, 750 mM Potassium Sodium Tartrate, dissolved in saturated 5,5'-Dithiobis-(2-nitrobenzoic acid) (DTNB) water, and 25% Glycerol. The crystal used in this chapter had dimensions of $110 \times 84 \times 75$ μm^3 in size. Insulin powder (Sigma, I5500) was dissolved to a concentration of 25 mg/mL in 50 mM Na_2HPO_4, pH 10.5, and 10 mM EDTA, and crystallised by mixing with 20% Ethylene glycol. The crystal used here had dimensions of $35 \times 45 \times 45$ μm^3 in size. To crystallise thermolysin, the protein powder (Sigma, P1512) was dissolved in a buffer consisting of 50 mM MES, pH 6.0, 45% DMSO and 50 mM Sodium Chloride, to a concentration of 50 mg/mL, and mixed with 1.2 M Ammonium Sulphate. The thermolysin crystal selected for this chapter measured $230 \times 70 \times 70$ μm^3 in size.

Sample preparation and data collection for in-vacuum X-ray crystallography followed a published procedure [89]. Tomography data collection was performed at the same wavelength, immediately following the diffraction experiment, as previously described in Chapter 3. All tomography datasets were collected at 100% transmission (no filtering was applied and the full X-ray beam intensity was used). and the beam was adjusted to a size of 700×700 μm^2. For thaumatin, 360° of diffraction data were collected at a wavelength of $\lambda = 4.13$ Å, with a top-hat beam size of 200×200 μm^2, 50% transmission, and a 0.1 s / 0.1° exposure. The tomography dataset consisted of 1800 projections, 20 dark images (no X-ray beam on the sample), and 20 flat-field images (sample out of the beam), recorded with an exposure of 0.2s / 0.1°. The insulin diffraction data was measured at $\lambda = 3.1$ Å as a 360° sweep, with 50% transmission, a beam size of 100×100 μm^2, and an exposure of 0.1 / 0.1°. For tomography, 1800 projections, 20 dark, and 20 flat-field images were collected with an exposure of 0.1s / 0.1°. For the thermolysin diffraction data, measured at $\lambda = 3.54$ Å, 360° of data were collected with a beam size of 350×350 μm^2, 15% transmission, and 0.1s / 0.1° exposure. To ensure data completeness, a kappa goniometer is used with the setting of –70°, where the rod-like crystal of thermolysin was aligned

Sample	Crystal	Liquor	Loop
Insulin at 3.10 Å	0.00745	0.00720	0.00690
Thermolysin at 3.53 Å	0.01312	0.01583	0.01172
Thaumatin at 4.13 Å	0.01926	0.02019	0.01864

Table 4.1: Absorption coefficients of materials in the samples

with the rotation axis. The tomography dataset consisted of 900 projections, 20 dark, and 20 flat-field images, recorded with an exposure of 0.2s / 0.2°. The rotation step size was determined using the Crowther criterion in geometric form: $d = D \cdot \sin(\Delta\theta)$, where D is the crystal dimension perpendicular to the rotation axis and $\Delta\theta$ is the rotation step. For thermolysin, with a cross-sectional diameter of $D = 70\,\mu$m and $\Delta\theta = 0.2^\circ$, the angular sampling yields $d \approx 70 \cdot \sin(0.2^\circ) \approx 0.244\,\mu$m, which is smaller than the detector pixel size of 0.3 µm and thus satisfies the Crowther criterion, with faster data processing. The diffraction data was indexed and integrated with DIALS [16].

The segmented tomographic reconstruction is used as our 3D models of the samples. The tomography data was processed with the SAVU pipeline [91], using standard flat-field correction, followed by ring artefact removal [63], and reconstructed using filtered back projection with TomoPy [92]. The reconstruction datasets were cropped from an initial size of 1600 × 1200 × 1200 voxels to remove unnecessary background and reduce the size of the data. The final dimensions of the tomographic datasets were 1120 × 1001 × 1001 voxels for thaumatin, 470 × 1000 × 1000 voxels for insulin, and 1210 × 1001 × 1001 for thermolysin, with voxel size of 0.3 × 0.3 × 0.3 µm^3. The reconstruction images were subsequently annotated using the segmentation software Avizo (Thermo Fisher), resulting in every pixel labelled as one of the three materials present in the sample: crystal, solvent, loop, or alternatively, background.

To calculate the absorption correction factors, we use a segmented 3D model in an $Array$ data structure, absorption coefficients, and a table of directional vectors for the incident and diffracted X-rays corresponding to the reflections **h**. The absorption coefficients can be determined as described in Chapter 3 or provided as input. The absorption coefficients of Insulin, Thermolysin, and Thaumatin are presented in Table 4.1. Unlike AnACor1.0,

which used Python, AnACor2.0's core computational modules are implemented as C function calls with CPU parallelism via OpenMP. These modules also have a Python interface and can be called directly from Python using `ctypes`. The output is a collection of analytical absorption factors A_h in $JSON$ format, arranged by the order of reflections in the input table. Once the calculations are complete, the analytical absorption correction is applied in the data-scaling process using `dials.scale` in DIALS [16, 112] with the flag of `analytical_correction=True`.

The absorption factors obtained through the standard ray-tracing technique with no sampling are established as the benchmarks for each dataset. The use of mean absolute percentage differences between the absorption factors of acceleration methods and the no-sampling standard method helps to assess the performance differences in acceleration. The differences are calculated as $\frac{abs(A_{acc}-A_{no})}{A_{no}}$ on an absolute scale, where A_{acc} and A_{no} are the absorption factors of the same reflection between accelerated method and no-sampling Standard method. Also, the peak heights in the anomalous difference Fourier maps of experimental datasets, are considered to examine if the final data quality prevails after applying acceleration methods. The published structures, PDBID 4A7E [113] for Insulin, PDBID 1KEI for Thermolysin and PDBID 1RQW for Thaumatin, are used as starting models for the Dimple pipeline (`http://ccp4.github.io/dimple/`). the `-- anode` option [93] is used to calculate anomalous difference Fourier maps with thresholding of 5σ and anomalous peak heights and the option `-- free-r-flags` in the Refmac refinement [94] step ensured the same R-free flags for all acceleration strategies. The peak heights of sulfur atoms in Insulin, Thermolysin and Thaumatin are selected to be compared with no-sampling results. Similar to the absorption factors, the percentage differences of peak heights are calculated as $\frac{abs(H_{acc}-H_{no})}{H_{no}}$, where H_{acc} and H_{no} are the anomalous peak heights of the same atoms between accelerated method and no-sampling Standard method.

4.2.2 Sampling

To accurately calculate an absorption correction, a precise 3D representation of the sample is essential. This requires a high resolution tomographicaly reconstructed volume composed of a large number of voxels. The computing cost associated with calculating path lengths for a large number of crystal voxels is high, along with this, neighbouring diffracting crystal voxels contribute very similar amounts to the overall absorption factor A_h. Given this, sub-sampling the crystal voxels can yield potential computational performance increases. Although voxel sub-sampling significantly reduces computational cost, it does not eliminate the need for high-resolution tomography. Lower-resolution tomograms may fail to capture fine geometric features and sharp boundaries between materials (e.g., crystal vs. mother liquor), leading to segmentation errors and inaccurate path length estimations. In contrast, high-resolution scans preserve spatial fidelity, ensuring that even sampled voxels represent the physical structure accurately. This trade-off ensures that the precision of the absorption correction is retained, while computational efficiency is achieved through sub-sampling. Hence, high-resolution tomography remains essential despite sub-sampling.

On the other hand, Equation 4.2 defines A_h to be the numerical mean of the linear absorption factors of overall crystal voxels. Hence a reduction in the number of summation terms can be obtained by selecting sample crystal voxels whose absorption factor coincides with the average absorption factor of neighbouring crystal voxels. Crystals have a considerable level of structural homogeneity and symmetry, resulting in close absorption effects for adjacent crystal areas. To demonstrate this, in Figure 4.1, six histograms are created to depict the absorption factors for different Systematic sampling ratios of a random reflection of thaumatin. The Kolmogorov-Smirnov (KS) test [114] is employed to evaluate how similar the distribution of each sampling ratio is to the distribution at full sampling (100%). The null hypothesis for this test assumes that the distributions being compared are identical. As the sampling ratio increases, the KS test values decrease, and the p-values increase, suggesting that the sampled distributions become more similar to the full, no-sampled distribution. Notably, p-values exceed 0.95 starting from a sampling ratio of 0.5%,

indicating insufficient evidence to reject the null hypothesis. At a sampling ratio of 1%, the p-value reaches 1, further affirming this trend.

Hence, employing Systematic sampling techniques to select voxels is a viable strategy. This methodology can effectively ensure comprehensive coverage of various crystal locations, facilitating a holistic understanding of the crystal's absorption characteristics. Conversely, this approach also reduces the influence of specific irregularities or disturbances in a particular area and improves the statistical dependability, yielding more resilient and precise estimations of the absorption factors A_h. The methodology employed in our prior research was the utilization of a systematic sampling technique mentioned in Chapter 3. This approach entailed arranging all crystal voxels in a sorted one-dimensional array and selecting a sample for every 2000 crystal voxel (sampling ratio of 0.05%). Nevertheless, it should be noted that the process of Systematic sampling may not always result in the selection of the most representative crystal voxels. In order to evaluate the effectiveness of the current Systematic sampling, three more sampling approaches are proposed: Random sampling, Randomised Systematic sampling and Stratified sampling. The Random sampling strategy selects crystal voxels randomly from a uniform distribution [52]. Compared to Systematic sampling, Randomised Systematic sampling selects a crystal voxel randomly within the interval rather than at the edges of the interval. A Stratified sampling method is also introduced, which uses a K-means clustering approach by starting with a random state. It separates the whole crystal volume into S small regions based on their coordinates and the spatial distances to the centroid of the crystal, where S is the number of sampled crystal voxels. Then, the sampled crystal voxels are the centroids of the small regions. In section 4.3, a comparative evaluation is conducted to evaluate the effectiveness of the sampling strategies in different example datasets.

4.2.3 Ray-tracing by the bisection method

In order to achieve accurate analysis within the three-dimensional model, the standard ray-tracing method must include voxel traversal for every voxel present in the X-ray

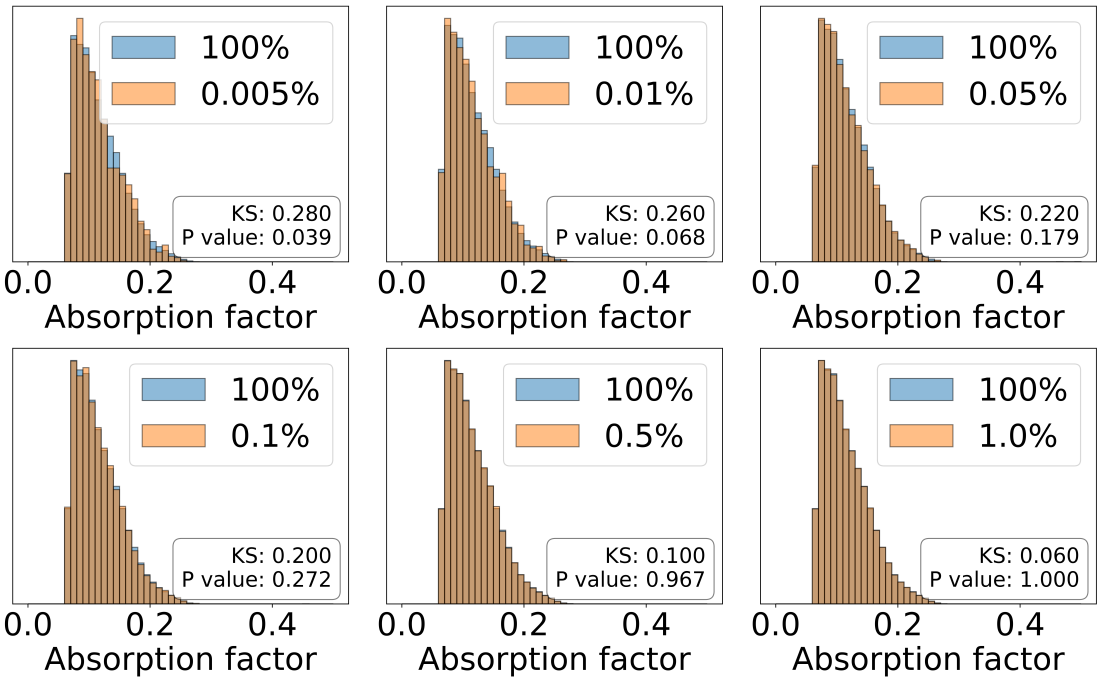


Figure 4.1: Histograms of absorption factors for different Systematic sampling ratios (orange) of a random reflection of thaumatin, compared to that of no-sampling (blue). The overlapping areas of the no-sample and sampled histograms are shown in dark orange. When the ratio rises to 0.5%, the p-values of the Kolmogorov-Smirnov (KS) test [114] become greater than 0.95, failing to reject the null hypothesis, and both histograms mostly overlap.

direction until it meets the end of the tomographic model. In essence, determining the path length requires obtaining information regarding the boundaries of the materials. In MX experiments, the samples comprise of only a small number of components, typically including mother liquor, crystal, and loop, and the sizes of these materials are large, exhibiting distant boundaries. Calculating lengths by traversing stepwise along the X-ray path is computationally demanding. In conventional ray-tracing methods, this approach necessitates a large number of repetitive calculations, especially within regions composed of a single material. A more efficient strategy would involve computing distances only across the boundaries separating different materials.

Hence, in order to optimise computational performance, it is beneficial to calculate the boundary coordinates using a bisection approach rather than traversing all of the voxels

along both the incident and diffracted X-ray paths. Additionally, a bisection approach has the potential to reduce the time complexity from $O(n)$ to $O(log_2 n)$. The bisection approach is further elaborated in Algorithm 2.

Algorithm 2 Algorithm to find the coordinates of boundaries by Bisection method

```
Input:
    OutermostCoord: the outermost coordinate of non-(air/vacuum) voxel in the ray
    CrystalCoord: the crystal coordinate n in Equation 4.3
    VoxelSize: the size of the voxel for the termination condition
    GoingIn: a boolean flag to determine search direction
    TargetClass: the class we are trying to find the boundary of
Output:
    BoundaryCoord: the coordinate of the boundary found by bisection
if GoingIn then
    startCoord ← OutermostCoord
    endCoord ← CrystalCoord
else
    startCoord ← CrystalCoord
    endCoord ← OutermostCoord
end if
while EuclideanDistance(startCoord, endCoord) ≥ VoxelSize do
    middleCoord ← midpoint(startCoord, endCoord)
    if label(middleCoord) is not TargetClass then
        startCoord ← middleCoord
    else
        endCoord ← middleCoord
    end if
end while
boundaryCoord ← middleCoord
return BoundaryCoord
```

The outermost coordinates are determined by the intersections of the X-rays with the plane of the model, which can be referred to as a cuboid with six planes. These coordinates can be computed using Equation 3.6. The Bisection method is capable of determining the coordinates of boundaries, however it lacks the ability to differentiate between inner and outer limits. Algorithm 2 contains the GoingIn procedure, which enables the Bisection method to determine whether the resulting boundaries are classified as inner or outer. Only the crystal's outer boundary is determined, as the ray traverses from the coordinate within the crystal. This method is based on the assumption that there are no air/vacuum

gaps between the crystal and the loop but only liquor in between them. It is sufficient to consider only the borders of the crystal, the loop and the interface between the sample and the surrounding air/vacuum. The default ordering to determine the boundaries of the Bisection method involves the following sequence: the crystal outer boundary, the air/vacuum boundary (which separates the sample from the air/vacuum), the loop inner boundary, the loop outer boundary, and, subsequently, the boundaries of other materials. The boundary determination process ceases once it has successfully acquired all material boundaries except for the mother liquor. The computation of final path lengths remains consistent with the standard ray-tracing method. The placement of the crystal and the loop can be random, resulting in liquor regions of varying sizes between them that cannot be predetermined. Calculating the dimensions of all these regions is computationally demanding, especially for smaller liquor regions. This variability adds complexity to the computational process. Therefore, the determination of the path length through the mother liquor is the subtraction of the total path length to the path lengths through the crystal, loop, air/vacuum, and any additional elements. Porosity (internal air or vacuum regions) is not currently accounted for in our model. If such regions exist within the sample, they could lead to overestimation of the absorption correction factor, result in an overestimation of the path length through mother liquor, and correspondingly its respective attenuation effect.

4.2.4 Gridding interpolation for multiple datasets

In the aforementioned approaches, the overall computing runtime scales linearly with the number of reflections, as both the ray-tracing and Bisection methods are performed for every individual reflection h. When processing multiple datasets, many diffracted X-rays share similar directional vectors, leading to redundant computations.

To improve computational efficiency, a precomputed grid of angular-dependent attenuation exponents $G_{(\theta,\phi)}$ is introduced, where each value is defined as:

$$G_{(\theta,\phi)} = -\sum_{m=1}^{M} \mu_m L_m^{(n)}$$

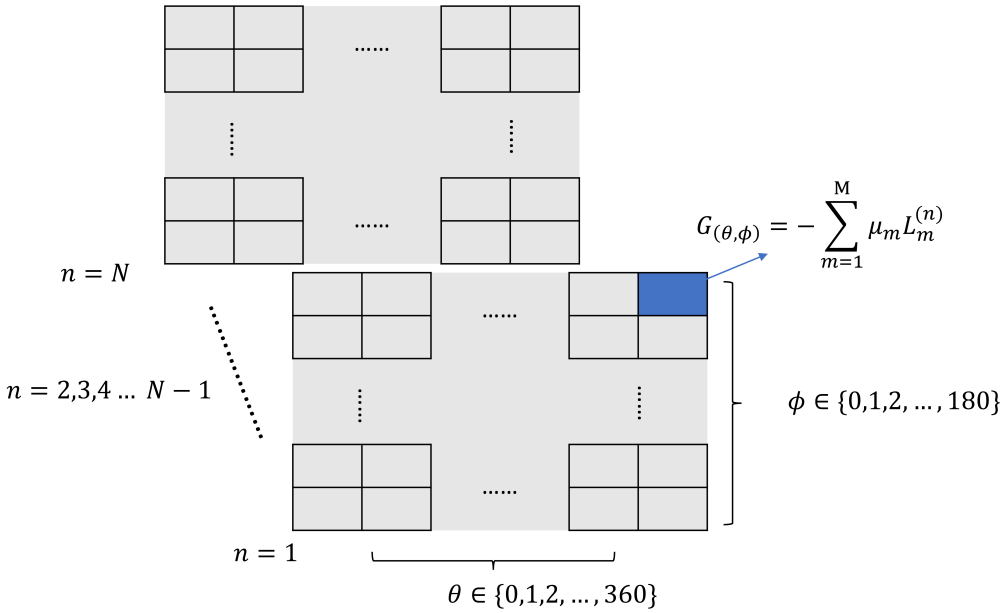


Figure 4.2: Illustration of Gridding interpolation algorithm. There are N absorption grids the same number as the crystal voxels with a shape of (360,180). Each grid point is an angular-dependent exponent of absorption factor.

Here, μ_m is the absorption coefficient and $L_m^{(n)}$ the path length through material m at voxel n for the given ray direction (θ, ϕ) in the spherical coordinate system. As illustrated in Figure 4.2, each of the N crystal voxels is indexed by n, and is associated with an angular grid that spans 360 values for θ and 180 values for ϕ (with 1° increments in both directions). This results in a 2D grid per voxel, where each entry corresponds to the exponent used in computing the absorption factor $A_h^{(n)}$, and enable efficient interpolation during subsequent calculations.

To mitigate edge effects during interpolation, continuity is ensured by wrapping $\frac{1}{12}$ of the grid data from one edge to the opposite edge, resulting in an expanded interpolation grid of dimensions $(420, 210)$. Once constructed, the exponent in the absorption factor $A_h^{(n)}$ can be determined by combining the incident and diffracted ray vectors and applying nearest-neighbour interpolation on the corresponding $G_{(\theta,\phi)}$ grid.

Interpolations are implemented using the GNU Scientific Library (GSL) [115], which supports high-precision numerical operations. As each voxel requires a separate angular

grid, memory requirements become significant—approximately 1 MB per voxel using double precision. Therefore, sampling strategies described earlier are essential to make this approach tractable for full datasets.

4.2.5 CUDA implementation

The processing power of GPUs can now be used in scientific environments, as demonstrated by the extensive NVIDIA GPU-accelerated libraries available [116]. GPUs, especially those like the NVIDIA V100, A100 and H100, are designed for high-performance parallel computing, enabling thousands of operations to be performed simultaneously. This is particularly beneficial for tasks such as ray-tracing, which involve extensive computational workloads.

To implement ray-tracing in AnACor2.0 on NVIDIA GPUs, the CUDA programming language is used. CUDA (Compute Unified Device Architecture) is NVIDIA's parallel computing platform and programming model based on C/C++, which enables developers to harness the power of GPUs efficiently.

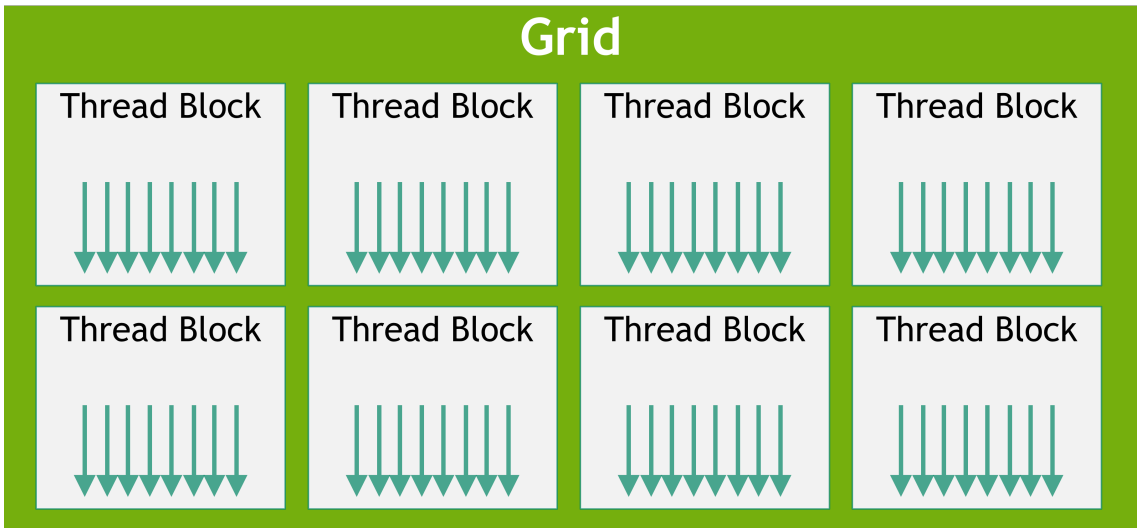


Figure 4.3: CUDA programming grids of thread blocks [116]

In CUDA, a thread is the smallest unit of execution that performs a sequence of instructions independently, and multiple threads run concurrently on the GPU to divide and process

tasks efficiently. The computations are organized into grids of thread blocks, where each block contains multiple threads, as illustrated in Figure 4.3. The threads within the same thread block can execute on the same streaming multiprocessor (SM) and cooperate to perform shared tasks. Each thread within a block has a unique identifier, accessible as `threadIdx.x`, which allows the thread to know its position within the block. Similarly, the total number of threads within a block is defined as `blockDim.x`. These identifiers are crucial for parallel computation, as they help determine which portion of the task each thread is responsible for.

The overall grid of thread blocks is identified by `blockIdx.x`, which indicates the block's position within the grid. By using these indices (`threadIdx.x`, `blockDim.x`, and `blockIdx.x`), CUDA allows the program to efficiently divide and manage workloads, ensuring that all threads operate concurrently across the GPU's many cores. This setup is essential for efficient executions of ray-tracing calculations, as it maximizes the parallel processing capability of the GPU.

To fully utilize the resources of a GPU, the implementation must expose sufficient parallelism—meaning it needs to break down the computation into many smaller tasks that can run concurrently. In the AnACor2.0 ray-tracing module, parallelism is exposed in several ways: the reflections being calculated, the paths of incident and diffracted X-rays, and the individual voxels each ray traverses. By parallelizing across all these aspects, the implementation maximizes the number of concurrent threads. The GPU implementation largely follows the steps outlined in Section 3.2.3, with necessary modifications to optimize for GPU execution. The details are shown in Algorithm 3.

Algorithm 3 Ray-Tracing Absorption Calculation in a CUDA kernel

Require: 3D segmentation model (d_label_list), Crystal coordinates (d_coord_list), Precalculated faces (d_face), Precalculated angles (d_angles), Precalculated increments ($d\_increments$), Output array (d_result_list), Reflection index ($index$), Maximum length on the 3D segmentation model ($total\_length$)
Ensure: Absorption value for each ray is stored in d_result_list
  Initialize thread ID: $id \leftarrow$ blockIdx.x
  Determine if the ray is incoming: $is\_ray\_incoming \leftarrow id\&1$
  Calculate position index: $pos \leftarrow id >> 1$
  Load coordinates from d_coord_list and face information from d_face
  Load angles θ and ϕ from d_angles using $index$ and $is\_ray\_incoming$
  Load traversal increments from $d\_increments$
  Calculate the number of iterations: $nIter \leftarrow \lceil total\_length/\text{blockDim.x} \rceil$
  for each iteration f from 0 to $nIter - 1$ **do**
    Calculate linear position: $lpos \leftarrow f \times \text{blockDim.x} + \text{threadIdx.x}$
    Compute new coordinates (x, y, z) using traversal increments and angles θ and ϕ
    if (x, y, z) are within voxel boundaries **then**
      Retrieve voxel label from d_label_list
      Update material counters based on the label value
    end if
  end for
  Calculate total ray length based on Euclidean Distance
  Compute absorption factors based on voxel material counts
  Store partial absorption value in shared memory: $s\_absorption[\text{threadIdx.x}]$
  Perform reduction to sum absorption values across threads in the block
  if threadIdx.x $== 0$ **then**
    Write the final absorption value to d_result_list for the current ray
  end if

A key challenge in GPU computing is the data transfer between CPU memory and GPU memory, as the bandwidth between these two components is relatively slow. To mitigate this bottleneck, the 3D segmentation model is transferred to the GPU memory once at the start, and all reflections are computed in parallel on the GPU. This reduces the need for frequent data transfers, improving overall performance.

Once the 3D segmentation model is transferred to the GPU, AnACor precalculates several parameters required for the ray tracing algorithm, including the rotation of incident and refracted rays, angles θ and ϕ, and traversal increments for each reflection. These calculations involve a large number of transcendental function calls (such as trigonometric functions), which are handled by the GPU's special function units (SFUs). SFUs have

lower throughput compared to the floating point units (FPUs) available on the GPU, so separating these computations with ray-tracing computation is essential. The precalculated values are shared among all rays within a given reflection vector **h**.

To further optimize the computation, each ray on a crystal voxel n is assigned to a single thread block. Threads within a thread block step along the ray's path in chunks, iterating through the 3D segmentation model, as illustrated in Algorithm 3. The number of iterations depends on the size of the 3D segmentation model. Each thread counts the number of different material voxels it encounters along the path. Similar to the serialized algorithm, the traversal ends when the boundary of the 3D segmentation model is reached. The partial absorption factors calculated by individual threads are then aggregated through a reduction operation to recover the total absorption factor, $A_h^{(n)}$, for the ray from that crystal voxel n. For improved efficiency and memory usage, each thread uses FP32 (32-bit floating point) precision during these calculations, allowing the GPU to process larger datasets within its memory.

4.3 Results

Figure 4.4 compares the mean percentage differences in absorption factors of various sampling strategies (a) and acceleration methods (b) against no sampling results. These illustrate the difference across sampling ratios from 0.001% to 1%, with error bars representing one standard deviation. Figure 4.4 (a) shows that when the sampling ratio is at least 0.01%, all mean differences are smaller than 2%, with Systematic sampling generally having smaller mean differences and less deviation. After the sampling ratio exceeds 0.5%, the differences become nearly zero across all sampling methods, with minor deviations. The exceptions are Stratified sampling in Thaumatin and Thermolysin. The poorer performance of Stratified sampling observed in Thaumatin and Thermolysin is likely due to the distinct morphological features of these crystals. Thaumatin is more isotropic but smaller in size, while Thermolysin has a highly elongated, rod-like geometry. In both cases, uniform stratified slices may not capture the spatial variability in absorption paths

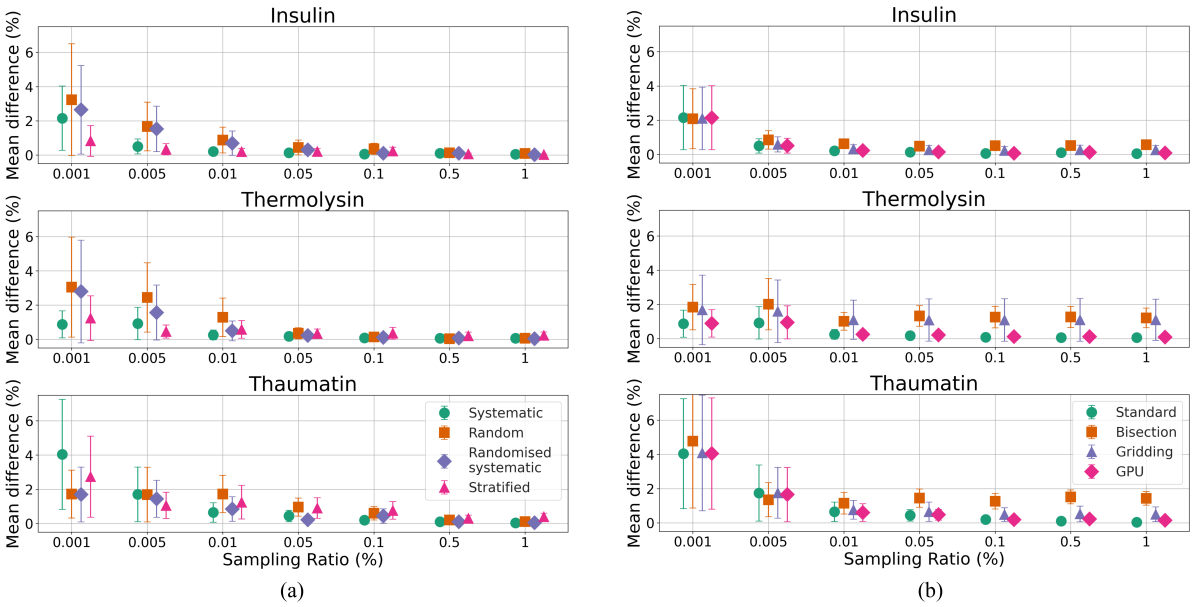


Figure 4.4: Mean absorption factors differences (%) between sampling and no sampling (a), and between acceleration methods and no sampling (b) for test crystal datasets across various sampling ratios. The sampling methods in (b) are all Systematic. The error bars represent one standard deviation.

as effectively. For Thermolysin in particular, a large proportion of the absorption variation is concentrated along the long axis, and stratified slices may sample insufficiently along this dimension. Conversely, Systematic sampling, which evenly spans the sorted range of path lengths, is less sensitive to spatial orientation and provides better coverage for such anisotropic or small crystals. These findings align with the Kolmogorov-Smirnov (KS) test results in Figure 4.1, where the P value approaches 1 for the 0.5% and 1.0 % sampling ratios.

Randomised Systematic sampling closely aligns with Systematic sampling but shows greater variances. However, it consistently outperforms Random sampling for ratios larger than 0.01%. For insulin crystals, which are roughly spherical, Stratified sampling proves more effective, showing smaller differences at low sampling ratios. In contrast, for Thermolysin and Thaumatin crystals, which grow as large rods and bipyramids, respectively, Stratified sampling performs less favourably at high sampling ratios. This highlights the significant impact of crystal size and shape on the success of Stratified sampling strategies.

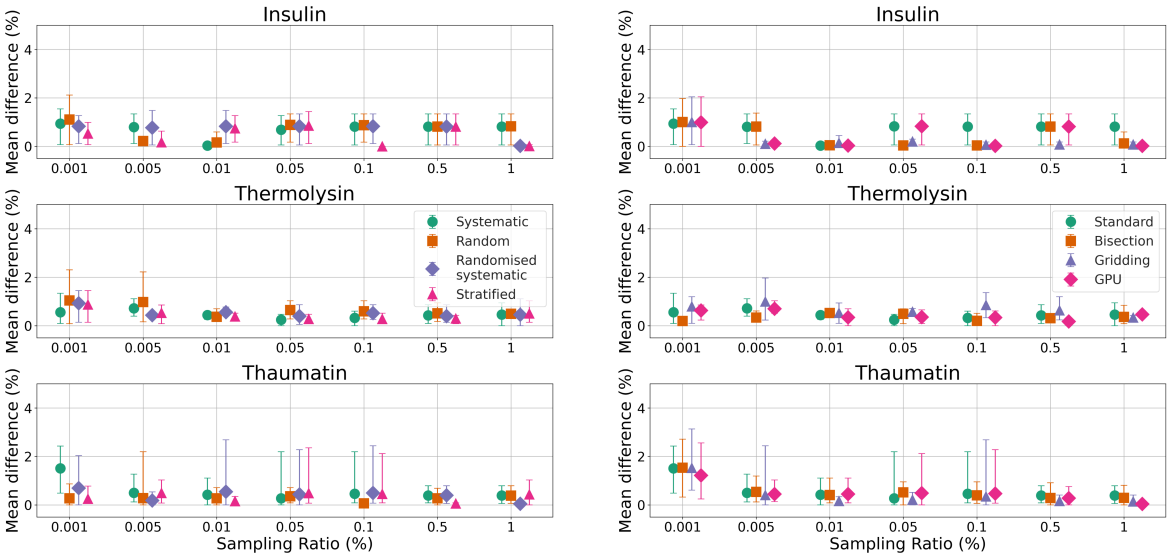


Figure 4.5: Mean anomalous peak height differences (%) of sulfur atoms between sampling and no sampling (a), and between acceleration methods and no sampling (b) for test crystal datasets across various sampling ratios. The error bars represent the maximum and minimum differences.

In this analysis of acceleration techniques for computational sampling, we focus on the Systematic sampling method because its absorption factor differences have smaller deviations across various sampling ratios relying less on the crystal shape. Figure 4.4 (b) illustrates the absolute mean differences between results from no-sampling and those achieved using acceleration methods. The Bisection and Gridding methods exhibit consistent differences once the sampling ratio exceeds 0.01%, showing uniform deviations across different datasets. Conversely, the GPU acceleration method, which uses a technique similar to the Standard method, displays smaller deviations. Specifically, in the cases of Insulin and Thaumatin, the Bisection method show higher deviations than Gridding. However, for Thermolysin, Gridding results in larger deviations, suggesting that the performance differences between Bisection and Gridding depend on the crystal's shape.

Absolute percentage differences between the anomalous peak heights of no sampling and those determined by sampling and acceleration methods are displayed in Figure 4.5, showing the mean differences of sulfur atoms with peak heights above 10σ. For sampling ratios larger than 0.01%, all sampling and acceleration methods display mean peak height

differences within 1% compared to the no-sampling results. This indicates that methods with sampling ratios above 0.01% can achieve similar peak height results. Except for Insulin, all sampling methods perform similarly at the same sampling ratios. Although there are differences in the absorption factors of Bisection and Gridding methods, their peak height differences perform similarly to standard and GPU methods at the same sampling ratios, except for the Gridding method of Thermolysin. For Insulin, the thresholds of $\leq 0.01\%$, the differences between sampling methods either stabilise around the mean of 1% with similar maximum and minimum values across various sampling ratios or become very close to zero. Conversely, the large differences in absorption factors in the Gridding method do not affect the peak heights shown in Figure 4.5 (b). The anomalous peak height differences remain around 1% with much smaller error bars. In addition, Thaumatin exhibits the largest error bars (defined as the range between maximum and minimum values) at sampling ratios of 0.05% and 0.1% in Figure 4.5(b), with maximum absolute differences exceeding 2% for the sulfur atom SG_A:CYS66, despite the mean differences remaining very small.

Atom	Residue	Chain	Peak Heights		
			No	SH	ACSH (100%)
S	CYS19	B	14.61	17.59	**17.39** ↓
S	CYS6	A	12.28	13.92	**14.18**
S	CYS20	A	12.03	13.61	**13.86**
S	CYS11	A	11.85	13.27	**13.44**
S	CYS7	B	8.98	11.16	**11.08** ↓
S	CYS7	A	8.25	10.23	10.44

Table 4.2: Anomalous peak heights of Insulin for no absorption correction (No), spherical harmonics correction (SH), and analytical absorption correction (100% full sampling) followed by spherical harmonics correction (ACSH).

The data presented in the Tables 4.2-4.4 reveal that for Insulin, the anomalous peak heights achieved with analytical absorption correction followed by spherical harmonics (ACSH) are quite similar to those obtained with spherical harmonics alone (SH), although both methods outperform the scenario without any correction (No). However, in the cases of Thermolysin and Thaumatin, where the wavelengths used are larger than 3.5 Å ACSH demonstrates a

Atom	Residue	Chain	Peak Heights		
			No	SH	ACSH (100%)
S	MET205	A	12.52	18.88	**21.54**
ZN	ZN405	A	9.77	15.00	**16.12**
S	MET120	A	7.69	11.97	**12.70**
S	SO41003	A	8.30	11.29	**11.72**
S	SO41006	A	6.74	7.26	**7.84**
CL	CL406	A	4.65	6.05	6.75
S	SO41004	A	4.31	5.72	6.40
S	SO41005	A	-	6.00	6.34
S	SO41007	A	4.92	6.75	6.11
S	SO41008	A	-	5.15	5.44

Table 4.3: Anomalous peak heights of Thermolysin for no absorption correction (No), spherical harmonics correction (SH), and analytical absorption correction (100% full sampling) followed by spherical harmonics correction (ACSH).

Atom	Residue	Chain	Peak Heights		
			No	SH	ACSH (100%)
S	CYS56	A	16.35	16.48	**16.61**
S	CYS9	A	15.01	15.26	**16.07**
S	CYS145	A	13.74	14.55	**15.64**
S	MET112	A	13.86	14.68	**15.44**
S	CYS149	A	13.87	14.37	**14.81**
S	CYS77	A	13.32	13.89	14.73
S	CYS193	A	13.16	14.08	14.45
S	CYS204	A	12.45	12.71	13.17
S	CYS71	A	11.51	12.00	12.62
S	CYS164	A	12.18	12.16	12.40
S	CYS66	A	12.01	12.14	12.29
S	CYS121	A	11.97	12.15	12.29
S	CYS126	A	10.57	10.74	11.53

Table 4.4: Anomalous peak heights of Thaumatin for no absorption correction (No), spherical harmonics correction (SH), and analytical absorption correction (100% full sampling) followed by spherical harmonics correction (ACSH).

significant improvement over SH. Such improvements are particularly more significant in key residues like MET205 in Thermolysin and CYS9 in Thaumatin, respectively. This indicates that while SH is sufficient for shorter wavelengths, ACSH provides substantial benefits in enhancing peak heights and data quality at longer wavelengths, particularly

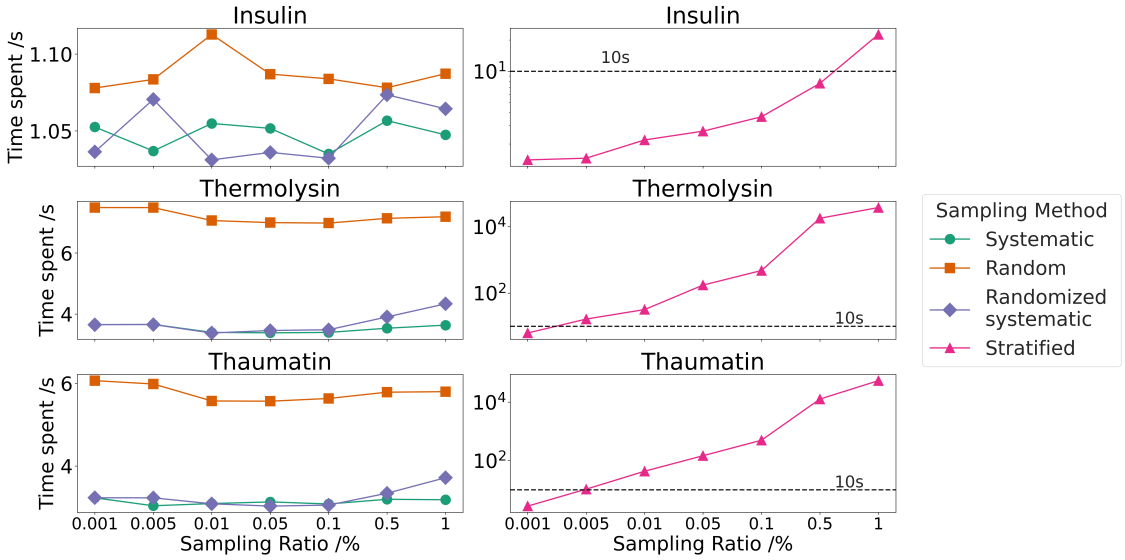


Figure 4.6: Average time spent on processing sampling methods for 10 runs. They are all determined on the same node with an Intel Xeon Platinum 8268 CPU with 48 cores

above the 3.5Å threshold. These findings align with the results in Chapter 3.

As shown in Figure 4.6, the processing time for Systematic and Randomised Systematic sampling is generally shorter and shows little difference compared to Random sampling. Both Systematic and Randomised Systematic sampling times increase with higher sampling ratios, while Random sampling time remains unaffected by the sampling ratios. Although more crystal voxels lead to longer processing times for all sampling methods, they still remain under 10 seconds. In contrast, the time required for Stratified sampling increases exponentially with higher sampling ratios. This is particularly evident for Thermolysin and Thaumatin crystals, which have substantially more voxels ($\approx$ 30 million, $\approx$ 20 million) than insulin crystals ($\approx$ 2 million voxels). This variation highlights the impact of crystal shape and sampling strategy on the efficiency of the sampling process.

In analyzing the balance between accuracy and computational speed provided by acceleration methods, Figure 4.7 presents detailed comparisons of computational time, using a sampling ratio of 0.5%. Figure 4.7 highlights two key findings: firstly, acceleration methods significantly shorten computational times across sampling ratios compared to the baseline mentioned in Chapter 3; secondly, it evaluates the efficiency in handling an

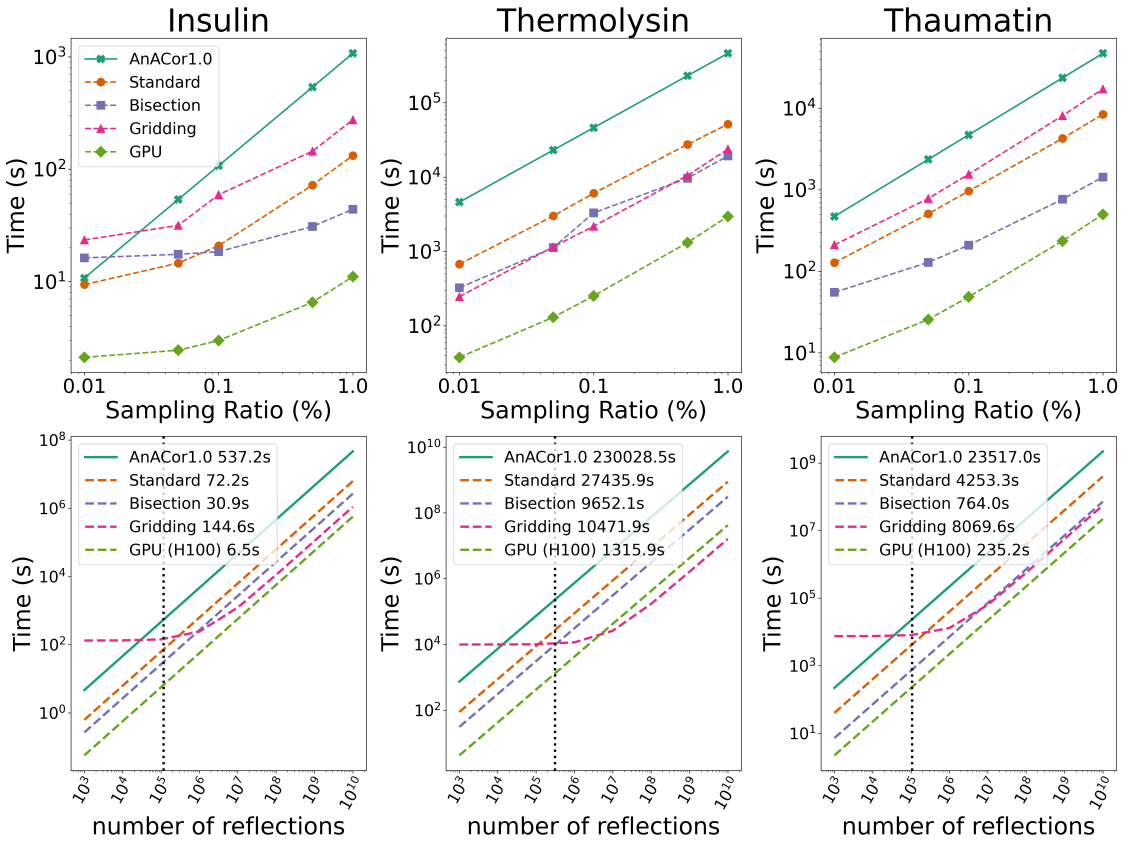


Figure 4.7: Top: Computational time taken by different acceleration methods across sampling ratios. Bottom: Computational time taken by acceleration methods with a Systematic sampling ratio of 0.5% to process increasing numbers of reflections. The black dotted line indicates the number of reflections in each experimental dataset, with computational times presented in the legend.

increased number of reflections. The comparative analysis on the top of Figure 4.7 reveals that acceleration techniques boost computational speed by at least 5 times at sampling ratios $> 0.01\%$. Among these, the GPU-based approach stands out for its exceptional time efficiency, taking around $\frac{1}{30}$,$\frac{1}{180}$ and $\frac{1}{90}$ of the time required by the baseline for Insulin, Thermolysin and Thaumatin, respectively. On CPUs, the Bisection method demonstrates superior speed over the Standard and Gridding methods for sampling ratio $> 0.1\%$. Notably, the performance of the Gridding method varies with sample shape and the number of reflections: it is faster for Thermolysin, while for Insulin and Thaumatin, the Gridding method is less efficient than the Standard method. The lower section of Figure 4.7 illustrates cross-over points between the Gridding method and other CPU-based methods.

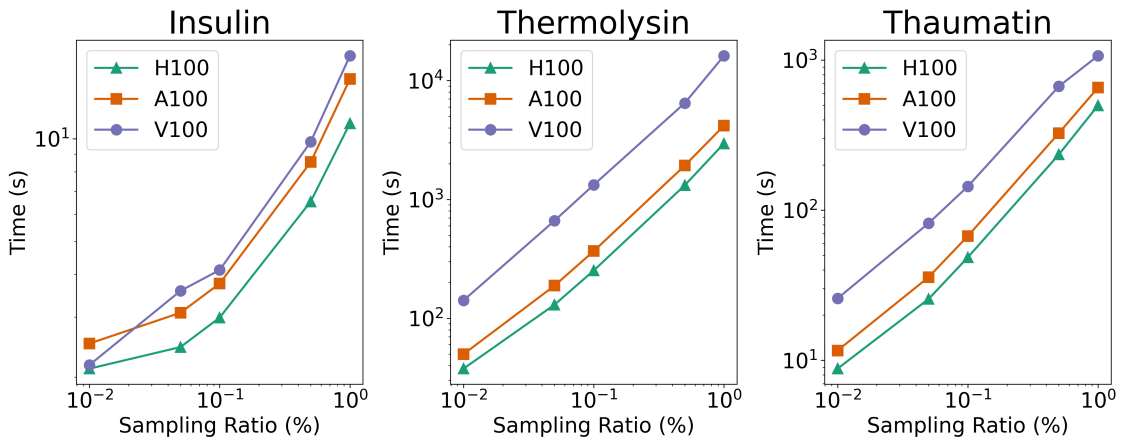


Figure 4.8: Computational time taken by different NVIDIA computational cards.

Notably, with an increasing number of reflections, as seen in Thermolysin and Thaumatin datasets, the performance of the Gridding method approaches that of the GPU method. This emphasise the Gridding method has the advantages when dealing with large datasets if advanced GPUs are not available.

Figure 4.8 compares the performance of three modern NVIDIA computational cards, H100, A100, and V100, across different sampling ratios. The results indicate similar trends in time expenditure among these GPU models, highlighting the consistent benefits of GPU acceleration across a range of computational hardware. This comparison underscores the importance of choosing the right acceleration method and computational hardware based on the specific requirements of the diffraction experiment, balancing speed against the need for accuracy in the final results.

4.4 Discussion

The previous chapter (Chapter 3) details the utilization of a ray-tracing algorithm for path length calculations in tomography reconstructions. This method is versatile and applicable to various fields for the determination of X-ray diffraction path lengths when a 3D model and its orientation are available. It then demonstrated the validity of the ray-tracing approach for analytical absorption corrections based on long-wavelength data from proteins crystallised

in monoclinic and triclinic space groups. This chapter has demonstrated the effectiveness of acceleration methods in AnACor2.0 for improving the efficiency of analytical absorption corrections over AnACor1.0 (presented in Chapter 3). Additionally, the data presented here not only confirms the previous findings on low-symmetry space groups discussed in Chapter 3, but also extends these findings to crystals with high-symmetry space groups, including cubic (Insulin), hexagonal (Thermolysin), and tetragonal (Thaumatin). More specifically, it also shows that a combination of analytical absorption correction and spherical harmonics yields substantial improvements in data quality, over spherical harmonics, for data collected at wavelengths larger than 3.5 Å(Thermolysin and Thaumatin) as illustrated in Tables 4.3 and 4.3.

Our findings indicate that the Systematic sampling used in AnACor1.0 consistently yields stable results with minimal differences and variance compared to no-sampling approaches, across increasing sampling ratios. Interestingly, Stratified sampling, employing a k-means clustering algorithm, can outperform Systematic sampling for crystals with more spherical shapes, such as Insulin, even for small sampling ratios. However, it becomes less practical for crystals with a large number of voxels due to the exponentially increasing sampling time and the challenges of achieving global optimization with the clustering algorithm. Random and Randomised Systematic sampling don't show clear advantages over the other sampling methods, so they are removed from practical use in AnACor2.0. While Stratified sampling is recommended for small crystals and spherical crystals, Systematic sampling remains the default option in AnACor2.0.

The analysis also reveals that the deviations between sampled and no-sampled absorption factors diminish beyond a 0.5% sampling ratio threshold in the test crystal datasets, advocating for the use of sampling-based calculations. For all the sampling methods, the anomalous peak heights results, for sampling ratios larger than 0.01%, show mean differences smaller than 1% compared to no sampling. The default sampling ratio is set to 0.5% to ensure accuracy, as confirmed by the P-value of the Kolmogorov-Smirnov (KS) test of over 0.95. The sampling ratio can be adjusted accordingly to prioritise computational

speed.

The introduction of acceleration techniques in this chapter has led to a remarkable increase in computational efficiency, improving the performance up to 180 times over AnACor1.0. mentioned in Chapter 3, by using NVIDIA GPUs. Although the Standard and GPU methods use the same underlying ray-tracing algorithm, the GPU method improves memory usage and data transfer efficiency by using `float32` precision for core computations. The results show that the difference between the `float64` precision used in the standard method and the `float32` precision used in the GPU method is very small. By employing `float32`, the GPU can handle larger datasets within its memory, reducing the frequency of data transfers between the CPU and GPU, which are often a significant bottleneck.

GPU acceleration with NVIDIA's H100 and a sampling ratio of 0.5% reduces the processing time of Insulin and Thaumatin to a few minutes (Figure 4.7), maintaining an absorption factor difference below 0.5%, compared to 100% sampling results (Figure 4.4).

If GPU acceleration is not available, AnACor2.0 also offers the Bisection and Gridding methods for improved CPU performance. The Bisection method emerges as a fast option, reducing the time complexity from $O\left(n\right)$ to $O\left(log_2 n\right)$. Meanwhile, the Gridding method is particularly adept at handling large datasets, offering an interpolation approach to reduce computational time. The results reveal that the bisection algorithm consistently shows the largest differences from no-sampling outcomes in the Insulin and Thaumatin cases. This is attributed to its approximation approach to the standard ray-tracing method, which assumes fixed relative locations for different materials and significant spacing between their boundaries. The Gridding method enhances computational efficiency by pre-calculating absorption factors in a spherical coordinate system with 1-degree increments between grids and employing nearest-neighbour interpolation during the inference stage. The efficiency of the Gridding method surpasses other methods when the number of reflections reaches a certain threshold, as depicted in Figure 4.7, however, it is sample-dependent. For instance, in the case of Insulin and Thaumatin, with a smaller number of reflections, the advantage of the Gridding method is reduced, as fewer computations of similar path lengths

are needed. However, the Gridding method can introduce errors because of the nearest-neighbour interpolation when there is a large path length difference between adjacent gridding points, as illustrated in the Thermolysin case. In Figure 3.5, if the direction of the ray rotates anti-clockwise, the path lengths through the solvent increase significantly, causing inaccurate interpolation. Therefore, the Gridding method is more suitable when the number of reflections is large (more than 10 datasets with different orientations), reducing the per-reflection computational cost, and when the crystal shape is more isotropic (closer to spherical), as the path length variations between adjacent grid points become smaller, minimizing interpolation errors.

The mean absorption factor differences between the Bisection and Gridding methods are larger than those of GPU acceleration. However, as illustrated in Figure 4.5(b), this is not reflected in the anomalous peak height differences of sulfur atoms, which are similar across the acceleration methods. Hence, a user can benefit from any of the acceleration methods presented, with their choice determined by the computational hardware available to them.

4.5 Conclusion

Analytical absorption corrections based on a ray-tracing approach improve the data quality for macromolecular crystallography at very long wavelengths. AnACor2.0 leverages numerical algorithms and GPU parallelism to significantly increase computational efficiency for calculating analytical absorption factors.

AnACor2.0 can calculate absorption factors in 6.5 seconds for Insulin, 1315.9 seconds for Thermolysin, and 235.2 seconds for Thaumatin, with GPU acceleration. This achieves computational efficiency improvements of over 90x, 175x, and 100x, respectively, compared to AnACor1.0. The deviations in absorption factors are minimal compared to the no-sampling results with AnACor1.0's standard ray-tracing method, at 0.09% for Insulin and Thermolysin, and 0.16% for Thaumatin. Additionally, the mean anomalous peak heights of sulfur atoms show deviations of only 0.82% for Insulin, 0.17% for Thermolysin,

and 0.28% for Thaumatin. In our research, the segmented 3D model was created using X-ray tomography at beamline I23, Diamond Light Source. Importantly, AnACor2.0's utility extends beyond data from this source and will be integrated within Dials in the future. It can facilitate analytical absorption corrections for any data set, provided that a voxel-annotated file is available and the relationship between the coordinate system of the 3D model and the diffraction experiment is clearly defined.

Chapter 5

Automatic segmented tomography reconstruction in crystallography

5.1 Introduction

Segmented tomography reconstruction has proven effective in accurately reproducing the three-dimensional details of a sample, including the crystal, mounting loop, and surrounding mother liquor mentioned in Chapter 3. The most common reconstruction methods are Filtered Back Projection (FBP) and iterative reconstruction [62]. However, segmenting the resulting 3D model is typically performed manually, a process that can take an expert up to a day to complete. Automatic segmentation, which involves assigning each voxel to the correct material, is challenging because it not only requires accurate classification but also needs to handle noisy inputs. This noise is often caused by artefacts such as streaks and rings that occur during tomography reconstruction.[63]. Semi-automatic segmentation techniques can assist in synchrotron radiation experiments. These include intensity thresholding, which separates regions based on differences in intensity [117]; region growing, which expands a region from a ”seed” point to include adjacent points with similar intensities [118]; and topological watershed, which separates regions based on gradient differences at the edges [119]. While these methods can effectively differentiate between regions, they still require human intervention to assign the correct material to each segmented region. Additionally, these techniques are sensitive to noise as mentioned above.

Recently, deep learning methods for segmentation in synchrotron tomography reconstruction experiments have gained significant traction [120, 121, 122, 123]. Among these methods, Convolutional Neural Networks (CNNs) have emerged as the preferred approach

due to their ability to automatically learn spatial hierarchies of features from input images. CNNs are particularly well-suited for image analysis tasks, with their architecture typically comprising several layers of convolutional filters that scan across the input data to capture local patterns such as edges, textures, and shapes. These convolutional layers are often followed by pooling layers, which reduce spatial dimensions, thereby enhancing translation invariance and reducing computational complexity.

One of the most prominent CNN architectures in this domain is U-Net [58], specifically designed for pixel-level segmentation, where each pixel of the input image is classified into a specific material. U-Net features a symmetric architecture with a contracting path (encoder) and an expanding path (decoder). The encoder captures contextual information through repeated convolutions followed by pooling operations, while the decoder upscales feature maps and applies transpose convolutions to enable precise localization. Crucially, U-Net incorporates skip connections between corresponding layers of the encoder and decoder, allowing it to combine high-level abstract features with low-level detailed features. This design enables U-Net to achieve high accuracy in segmenting complex structures, even with limited training data, which is a common challenge in tomography reconstruction experiments in synchrotrons.

The tasks of medical image segmentation and synchrotron tomography reconstruction share significant similarities, particularly in their focus on volumetric image analysis, where the primary objective is to accurately describe complex structures within noisy and heterogeneous data. Both fields face challenges in segmenting intricate and overlapping features, necessitating robust architectures. While CNNs excel at capturing local details and low-level information, they often struggle with incorporating global context [124]. This limitation has been addressed by the Vision Transformer (ViT) architecture, which has recently demonstrated remarkable success in various medical image segmentation tasks [125]. ViT models partition an image into a sequence of patches and model their dependencies using self-attention mechanisms. This approach may not be as effective as the convolution operations employed by CNN models for extracting local features

within receptive fields. Therefore, it is straightforward that CNN and ViT can be combined through U-net (encoder-decoder) structures to mitigate the local representation weaknesses. In this study, a 3D segmentation model called AnACorNet is proposed based on the previous literature on the combination of CNN and ViT [76, 126, 127]. Due to the limitation of the training dataset, a simulation dataset is designed and then fed into training the segmentation model. It is proven that after introducing a synthetic simulation dataset to the training process, the accuracy of AnACorNet_RS is improved significantly. The SAM-2 refinement further enhances segmentation accuracy by addressing residual artefacts and refining the boundaries, leading to more precise segmentation results. As a result, with scaling methods AAC and ACSH mentioned in Chapter 3, we obtain absorption correction outcomes for the Insulin and Thermolysin datasets that closely match the manual segmentation, while achieving even better performance for the Thaumatin dataset. This demonstrates the effectiveness of combining synthetic simulation data and SAM-2 refinement in improving segmentation quality and absorption correction accuracy across various datasets.

5.2 Methodology

5.2.1 Principles of simulating tomography projection images

One of the contributions of this study is the generation of the synthetic dataset. The simulation process consists of projecting the sample to the detector and then performing tomography reconstruction on the projection images by filtered back-projection (FBP). In a real tomography reconstruction experiment, there are significant edge effects at the boundaries between different materials.

This is because the sharp edge of the crystal and the mounting loop generate significant phase contrast due to Fresnel diffraction, particularly at material boundaries with large refractive index gradients. These edge effects are amplified by the finite propagation distance (i.e., sample-to-detector distance) and the small pixel size of the detector, which together make the phase contrast more detectable in the tomography experiment. In order

to achieve a high-similarity simulation dataset, the physical principles of X-ray tomography imaging need to be considered.

The overall process can be treated as a wavefield emitted by a source propagating through an arbitrary number of objects, and finally, its intensities are captured at the virtual imaging plane (detector in reality) at a certain distance from the source.

Incident wavefield

A point source emits a monochromatic wavefield $u_0(\mathbf{x}, z_1)$ with wavelength λ, where $\mathbf{x}$ represents the 2D coordinates perpendicular to the X-ray incident axis z, and z_1 is the distance to the first object in the beam path. The intensity distribution of the wavefield is given by $I_0(\mathbf{x}, z_1) = |u_0(\mathbf{x}, z_1)|^2$. For a spherical wave, when z_1 is sufficiently large (as in synchrotron common setups), the spherical phase profile can be approximated to [128]:

$$u_0(\mathbf{x}, z_1) = \sqrt{I_0(\mathbf{x}, z_1)}\, e^{jkz_1} \tag{5.1}$$

where $k = \frac{2\pi}{\lambda}$ is the wavenumber.

Wavefield propagations

When an X-ray beam propagates through non-vacuum objects, its intensity is attenuated, and the beam undergoes a phase shift. This behavior is described by the 3D complex refractive index of the object i at the 2D coordinate $\mathbf{x}$, located a distance z from the X-ray source. The refractive index is represented as [129]:

$$n_i(\mathbf{x}, z) = 1 - \delta_i(\mathbf{x}, z) + j\beta_i(\mathbf{x}, z), \tag{5.2}$$

where $\delta_i(\mathbf{x}, z)$ corresponds to the real part of the refractive index, representing the phase shift, and $\beta_i(\mathbf{x}, z)$ is the imaginary part, representing absorption within the object.

The propagation function at the exit plane of object i can be determined by integrating along the z-direction. This is expressed as [129]:

$$T_i(\mathbf{x}) = \exp\left(jk \int n_i(\mathbf{x}, z)dz\right) = e^{-k(B_i(\mathbf{x}) - j\varphi_i(\mathbf{x}))}, \tag{5.3}$$

where

$$B_i(\mathbf{x}) = \int \beta_i(\mathbf{x}, z)dz \quad \text{and} \quad \varphi_i(\mathbf{x}) = \int \left[1 - \delta_i(\mathbf{x}, z)\right] dz.$$

Here, $B_i(\mathbf{x})$ represents the cumulative local absorption of the X-ray as it propagates through object i, and $\varphi_i(\mathbf{x})$ represents the total phase shift induced by the refractive index variation. Therefore, the relationship between the wavefield $u_{i-1}(\mathbf{x}, z_i)$ at the entrance plane of the i-th object and $u_i(\mathbf{x}, z_i)$ at the exit plane can be described as:

$$u_i(\mathbf{x}, z_i) = T_i(\mathbf{x}) u_{i-1}(\mathbf{x}, z_i). \tag{5.4}$$

In the case where the X-ray propagates through air or vacuum, the wavefield does not experience material attenuation but still undergoes spreading, diffraction, and phase evolution as it propagates through free space. This free-space propagation can be modelled using the angular spectrum formalism between two parallel planes separated by a distance Δz [130]. Therefore, the 2D Fourier transform of the wavefield, denoted by $\tilde{u}(\xi) = \mathcal{F}[u(\mathbf{x})]$, describes the wavefield in terms of 2D spatial frequencies ξ. The free-space propagator is given by:

$$\tilde{u}(\xi, z + \Delta z) = \tilde{P}(\xi, \Delta z)\tilde{u}(\xi, z), \tag{5.5}$$

where the propagator $\tilde{P}(\xi, \Delta z)$ can be written as:

$$\tilde{P}(\xi, \Delta z) = \exp\left\{jk\Delta z \left[1 - (\lambda\xi)^2\right]^{1/2}\right\} \quad \text{or} \quad \tilde{P}_F(\xi, \Delta z) = \exp(jk\Delta z)\exp(-j\pi\lambda\Delta z\xi^2). \tag{5.6}$$

In this equation, the second form $\tilde{P}_F(\xi, \Delta z)$ is the Fresnel approximation, derived under the assumption of small angles (parabolic approximation). The Fresnel approximation is suitable when the distance between the object and the detector is large compared to the

wavelength and the feature size of the object, which is common in most X-ray imaging applications. However, it should not be used when dealing with very small distances or high-resolution imaging where the full angular spectrum method is required for accuracy. Thus, the wavefield at a distance Δz behind the i-th object can be calculated using the recursive relation:

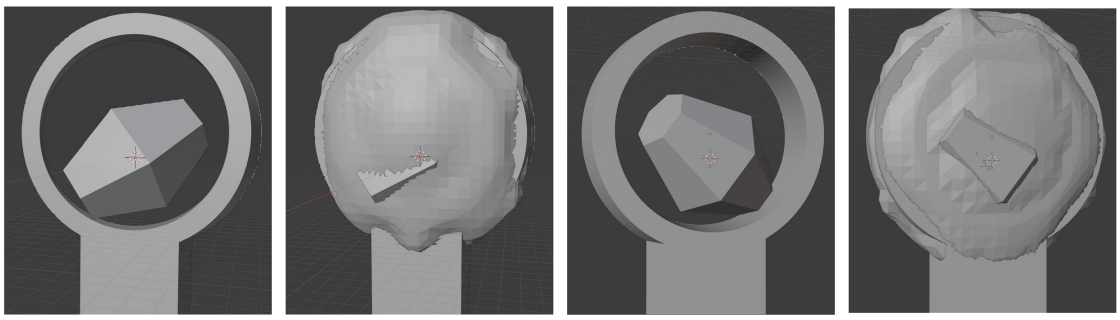
$$u_i(\mathbf{x}, z_i + \Delta z) = \mathcal{F}^{-1}\left\{\tilde{P}(\xi, \Delta z)\mathcal{F}\left[u_{i-1}(\mathbf{x}, z_i)T_i(\mathbf{x})\right]\right\}. \tag{5.7}$$

In this context, the sample and detector can be treated as different instances of object i, allowing for a recursive propagation of the wavefield from the X-ray source to the detector plane.

5.2.2 Implementation details

The sample in the simulation process is constructed and designed using a 3D mesh software called *Blender*, which has Python API to allow automation. The simulated mounting loop is created manually by comparing the dimensions of a real MiTiGen mounting loop used in Beamline I23 at Diamond Light Source. The crystals are simulated using an open-source repository [131] to create different sizes and shapes with a varying number of vertices. The crystals are simulated using an open-source repository [131] to generate diverse 3D crystal morphologies across a range of sizes and shapes, using low-polygon meshes defined by edge lengths, vertices, and faces. The method focuses on replicating the external geometry of crystals—such as rod-like, plate-like, or equant forms, without modeling internal atomic details. While crystallographic space groups define the internal symmetry and unit cell geometry, they do not uniquely determine the macroscopic crystal shape, which is strongly influenced by growth conditions and surface energetics. Therefore, the simulation approach does not directly use space group symmetry. Instead, it supports the indirect incorporation of space group–related information, such as crystal system (e.g., monoclinic or orthorhombic) or known aspect ratio tendencies, to guide the geometry generation. The simulation is implemented in `Python` using the Blender API, enabling efficient batch generation of

mesh models for use in tomography simulation, data augmentation, or machine learning applications. The mother liquor is generated using the *Liquid diffusion* module in *Blender*, which simulates how virtual liquids interact with their environment and other objects. An example simulation is shown in Figure 5.1. The simulation output from *Blender* is in *Mesh* format, and the 3D volume of a tomography reconstruction is in *Array* format. Hence, the simulation dataset from *Blender* is voxelized using *Open3D* [132].



(a) (b) (c) (d)

Figure 5.1: A example of simulation results by *Blender* with front ((a)-(b)) and back ((c)-(d)) views. The mother liquor simulations are shown at (b) and (d)

Then the simulated model of the sample is forwarded to the tomography projection model mentioned in the previous section §5.2.1. The incident wavefield is assumed to be the tomography projection images taken without any objects present, representing only the background signal at the corresponding wavelengths. The size of the background tomography projection image dictates the dimension of the virtual detector. The simulated sample is placed in the middle of the virtual detector and then it performs the simulation of free-space propagation of the incident beam between the source and the sample.

To simulate the propagation in Equation 5.3, hyperparameters δ and β, corresponding to local absorption and phase shift respectively, are needed. The conversion between β in Equation 5.3 and the linear absorption coefficient μ is $\beta = \frac{\mu\lambda}{4\pi}$. The absorption coefficients for both the simulated crystal and the surrounding mother liquor are adjusted by introducing random perturbations. These perturbations are within a range of $\pm 20\%$, centred around

the baseline absorption coefficients provided in the standard dataset mentioned in Chapter 4. These adjustments are applied at three different wavelengths as listed in Table 4.1: specifically, for Insulin at 3.10 Å, Thermolysin at 3.54 Å, and Thaumatin at 4.13 Å. The purpose of this random variation is to simulate realistic experimental conditions where absorption coefficients may fluctuate due to factors such as sample heterogeneity or experimental noise. The absorption coefficients of the mounting loop at three different wavelengths, which is made of Kapton, are determined from the CXRO database [133]. The parameter δ corresponding to the phase shift of the mounting loop are also obtained from the tables while those of the crystal and the mother liquor are difficult to measure in a standard tomography experiment setup.

The δs mainly depend on the atomic composition and electron densities of the object, and they can be calculated by the following equation [134]:

$$\delta = \frac{r_e \lambda^2 N_A Z \rho}{2\pi A} \tag{5.8}$$

where r_e is the classical electron radius (usually 2.817×10^{-15} m), λ is the X-ray wavelength in meters, N_A is Avogadro's number (6.022×10^{23} mol^{-1}), Z is the weighted average atomic number of the protein, ρ is the protein density (usually 1.35g/cm^3 [135]), and A is the atomic mass in grams per mole, which can be found in the Protein Data Bank (PDB). PDB IDs are 4A7E [113] for Insulin, 1KEI for Thermolysin and 1RQW for Thaumatin.

As the sample is in the 3D *Array* format, each slice perpendicular to the **z** axis (incident X-ray direction) can be treated as the 2D plane **x** and dz is turned to be Δz, which is the separation distance between slices. This separation distance is $0.3\mu m$ which is the same as in Chapter 3 and Chapter 4. Finally, an overall propagation function $T_i(\mathbf{x})$ after the incident beam interacts with the simulated sample is constructed.

After simulating the later free-space propagation between the sample and the virtual detector, the intensities of the wavefield are recorded as synthetic tomography projection images. By rotating the simulated sample for 180° with 0.2° as an increment, there is

a total of 900 synthetic tomography projection images in a dataset. In order to achieve computational efficiency, the generation of tomography projection images consists of Multiprocessing and GPU acceleration. We multiprocesses projection images at different angles concurrently and utilise *CuPy* [136] to accelerate the Fast Fourier Transform (FFT) in the free-propagation equation and compute the propagation function concurrently. The reconstruction process is completed using the *TomoPy* [137] software, which is GPU-accelerated.

5.2.3 Segmentation model

A complex U-Net architecture with hybrid large-kernel attention blocks and Vision Transformer (ViT) modules is proposed, as illustrated in Figure 5.2. The network follows an encoder-decoder structure based on a 3D U-Net, enhanced with Transformer-based attention modules. The input to the model is a volumetric tomogram of shape $B \times C \times D \times H \times W$, where B is the batch size, C is the number of input channels, and D, H, W are the spatial dimensions.

The encoder path consists of four levels of downsampling, each followed by a sequence of 3D Trans-UNet blocks. At the deepest level, two 3D Trans-UNet blocks operate at $D/16 \times H/16 \times W/16$ resolution, capturing global contextual information. Each 3D Trans-UNet block integrates convolutional layers with large-kernel deformable attention modules and 3D Vision Transformer (ViT) units, allowing the network to model both local detail and long-range dependencies. Residual connections and skip connections between the encoder and decoder levels help preserve spatial features and mitigate information loss due to downsampling.

The decoder mirrors the encoder, using upsampling layers followed by 3D Trans-UNet blocks to progressively recover the spatial resolution. Feature maps from the encoder are concatenated at each decoder level via skip connections. Final feature refinement is performed using two $3 \times 3 \times 3$ residual convolutional blocks, followed by a $1 \times 1 \times 1$ convolution to produce the class scores for each voxel.

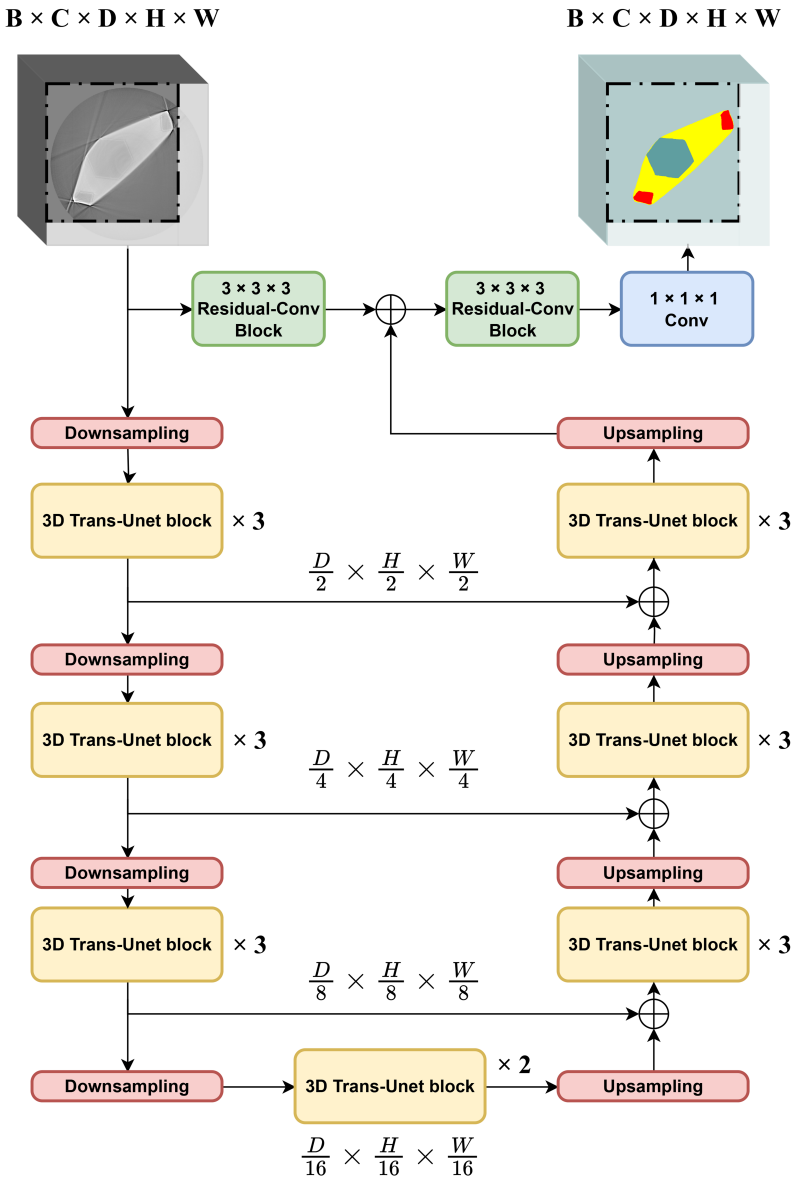


Figure 5.2: The overall architecture of 3D segmentation model.

The model segments each input volume into four distinct classes: background, crystal, mother liquor, and mounting loop. The hybrid attention modules and ViT layers are critical in coping with the severe noise and blurring introduced by the filtered back-projection (FBP) reconstruction, enhancing both localization and class boundary accuracy. This design balances detail preservation and semantic understanding, enabling effective segmentation even in challenging tomographic conditions.

Hybrid large-kernel attention with deformable convolution (HLKA)

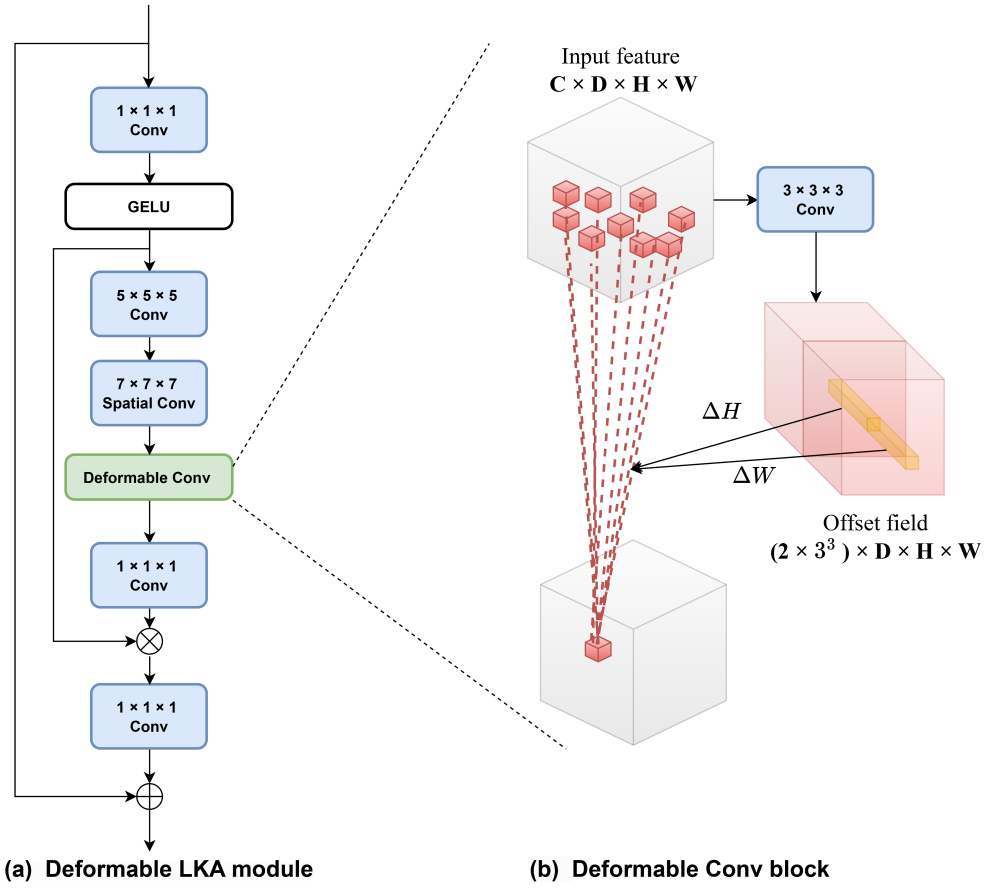


Figure 5.3: Illustration of hybrid large-kernel attention with deformable convolution (HLKA) module. Blank blocks represent layers with no learnable parameters, while blue blocks indicate components consisting of a single neural network layer.

The self-attention mechanism [72, 138] treats images as 1D sequences, neglecting their inherent 2D structure. The receptive field is the region of the input image that contributes to a particular output value. In convolutional neural networks, it determines the amount of spatial context or information from the input image that a specific layer of the feature map can capture and interpret. Large-kernel convolution can achieve a similar receptive field as the self-attention mechanism, allowing it to capture long-range dependencies in the image. Moreover, it can be made computationally efficient by decomposing through a depth-wise convolution ($DW\text{-}Conv$), a depthwise dilated convolution ($DW\text{-}D\text{-}Conv$), and a $1 \times 1 \times 1$ convolution [139].

The standard large-kernel convolution is effective at capturing broad spatial context but has fixed sampling locations (fixed kernel grid), making it less adaptable to local variations and irregularities. In contrast, the deformable convolution (DC) can dynamically adjust its sampling points, making it more flexible and capable of accurately capturing fine-grained details, even in complex or noisy regions [140].

The architecture of hybrid large-kernel attention with deformable convolution (HLKA) module and its corresponding deformable convolution block is shown in Figure 5.3 (a) and (b) respectively. The HLKA module begins with a $1 \times 1 \times 1$ convolution followed by the GELU activation function, as in previous literature [139]. This is followed by a $5 \times 5 \times 5$ depthwise convolution and a $7 \times 7 \times 7$ depthwise convolution with a dilation rate of 3. The deformable convolution layer is then applied, which adjusts its kernel dynamically to focus on relevant local details. The overall module can be computed as:

$$Tmp = GELU(\text{Conv}_{1\times1}(Inp)) \tag{5.9}$$

$$\text{Attention} = \text{DW-D-Conv}(\text{DW-Conv}(\text{DC}(Tmp))) \tag{5.10}$$

$$\text{Output} = \text{Conv}_{1\times1}(\text{Attention} \otimes Tmp) + Inp \tag{5.11}$$

Here, $Inp \in \mathbb{R}^{C\times H\times W}$ is the input feature from the last layer. Attention $\in \mathbb{R}^{C\times H\times W}$ denotes the attention map. The value in the attention map indicates the importance of each feature. $\otimes$ means element-wise product. Different from common attention methods, HLKA doesn't require additional normalization functions such as sigmoid or Softmax. These normalization functions tend to neglect high-frequency information, thereby decreasing the performance of self-attention-based methods [141].

Deformable convolution (DC)

The deformable convolution block is shown in Figure 5.3 (b). A standard $3 \times 3 \times 3$ convolution is applied to generate a dynamic offset field, which shifts the positions of the sampling points within the convolution operation. For each location in the input feature

map, the network learns an offset field of size $(2 \times 3^3) \times D \times H \times W$, where $3^3 = 27$ corresponds to the number of points in the convolution kernel (since the kernel is $3 \times 3 \times 3$, it has 27 positions). To maintain computational efficiency, only 2D spatial information is dynamically learned in the direction of H and W (deformed along the height and width dimensions), which accounts for the factor of 2 [126]. These learned offsets, denoted as ΔH and ΔW, allow both positive and negative shifts in the sampling grid for each spatial location in the input feature map. This process can be expressed as:

$$Y(p_0) = \sum_{p_n \in \mathcal{R}} W(p_n) \cdot X(p_0 + p_n + \Delta p_n)$$

where p_0 is the reference position in the input, $\mathcal{R}$ is the set of relative positions in the $3 \times 3 \times 3$ convolution kernel, and $W(p_n)$ are the weights corresponding to each position p_n. The feature map X is sampled at positions $p_0 + p_n$ shifted by the learned offsets $\Delta p_n = (0, \Delta H, \Delta W)$. This deformable convolution allows the model to adapt its 2D spatial focus dynamically, thereby making it more robust to noise and irregularities in each slice.

AnACorNet

The input to the model is 3D data composed of a list of 2D slices from tomographic reconstruction. This data is generated using Filtered Back Projection (FBP), which often results in large redundant regions around the edges of each slice. The area of the detector exceeds the sample size, leading to vacuum or air regions at the start and end of the slice list. To achieve data augmentation and enhance the computational efficiency, as illustrated in Figure 5.2, $\frac{3}{4}$ of the 3D input data is randomly cropped before entering the model. The architecture follows a U-Net style structure combined with Trans-Unet blocks for feature extraction. The cropped input first passes through a 3D residual-conv block, which includes two $3 \times 3 \times 3$ convolutions with a residual connection. Following this, downsampling is applied multiple times, each followed by 3 successive 3D Trans-Unet blocks. The downsampling layer is performed by convolution with kernel size and stride size of 2 and

doubling the channel size. The central bottleneck includes another set of 2 Trans-Unet blocks before upsampling begins. The upsampling path by transpose convolution mirrors the downsampling path, with the corresponding number of Trans-Unet blocks applied at each dimension. Skip connections between matching dimension in the downsampling and upsampling paths ensure the propagation of fine-grained features. After processing the input data through the initial 3D residual-conv block, the output is combined with the input data using a residual connection. This combination helps retain essential features from the original input. The combined result is then passed through another 3D residual-conv block. Finally, a $1 \times 1 \times 1$ convolution is applied to adjust the number of channels so that it matches the number of segmentation classes

The Trans-Unet blocks integrate a ViT layer followed by residual convolution, they are used as the core computing blocks as shown in Figure 5.2. First, the 3D input is flattened into a 1D sequence for the ViT-like architecture, with layer normalization [74] and positional embedding applied to ensure that the model is aware of the original spatial structure even after the data has been flattened. The hybrid large-kernel attention (HLKA) module follows, capturing both global and local dependencies. After a residual connection, the sequence is reshaped back into a 3D feature map for convolutional processing. The residual-conv block shown in Figure 5.4(b) applies a $3 \times 3 \times 3$ convolution, batch normalization, and Leaky ReLU activation [142], followed by another convolution and batch normalization [143]. A residual connection ensures gradient flow, which helps mitigate the vanishing gradient problem by allowing gradients to bypass layers during back-propagation, thus improving the training of deep networks. A final $1 \times 1 \times 1$ convolution is then applied to reduce the dimensionality, making the output suitable for the next stage [73].

Refining details using SAM-2

Voxel-wise segmentation involves predicting the most probable material or class for each individual voxel in an image. This is inherently a probabilistic task, meaning that the model assigns probabilities to each possible class and selects the one with the highest likelihood.

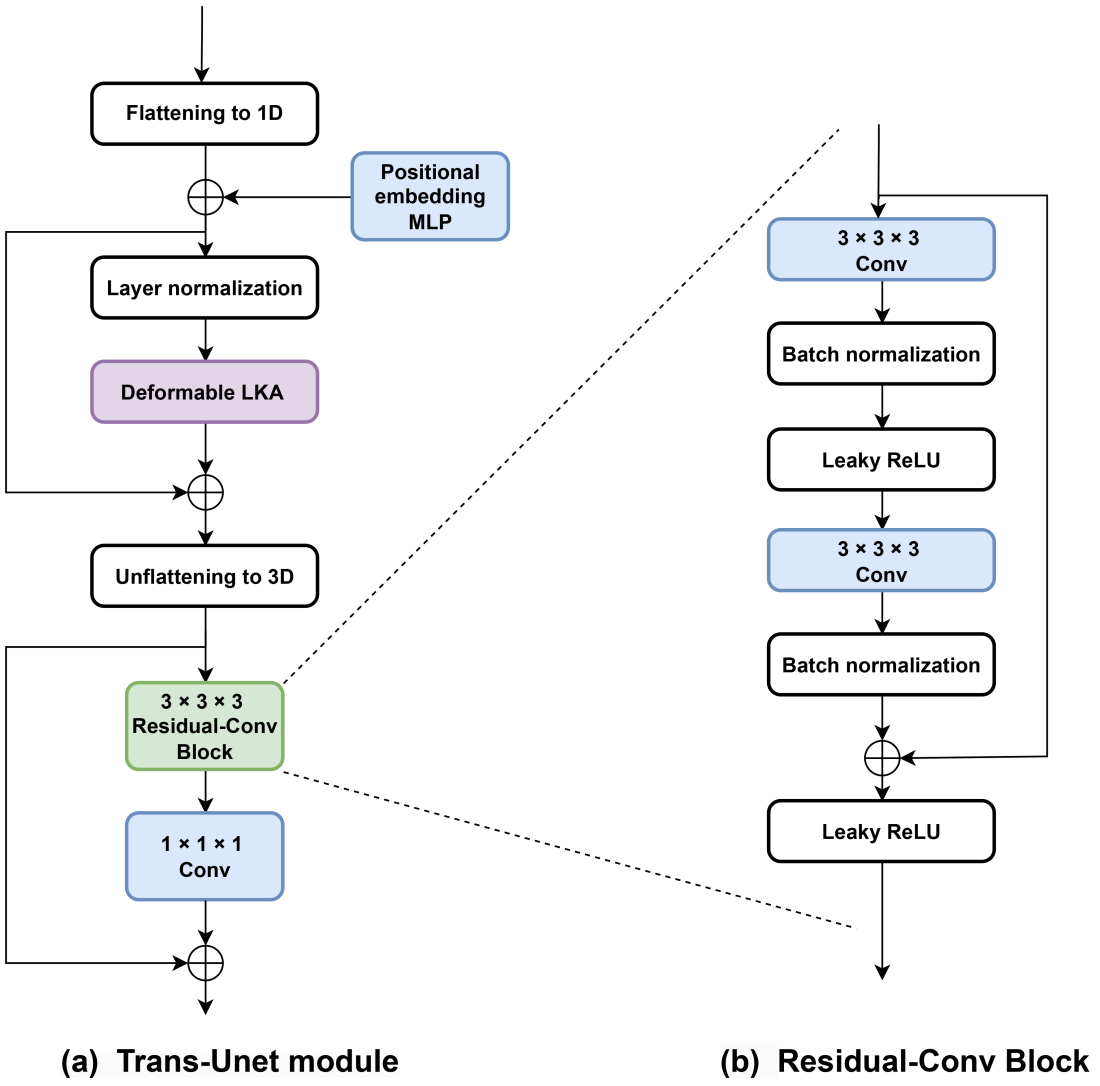


Figure 5.4: Illustration of ViT module in AnACorNet.

In many applications, this method works well and provides sufficiently accurate results. However, in scenarios like absorption correction, where extreme precision is required, the limitations of this probabilistic approach become more apparent. Small errors that might be negligible in standard segmentation metrics can lead to critical parts of the segmentation being missed. These omissions can significantly compromise the accuracy of the absorption correction, resulting in errors that could affect the overall quality and reliability of the final outcome. Also, during the training of the segmentation model, input data is often down-scaled through interpolation to enhance computational efficiency. While

this down-scaling speeds up the training process and reduces the computational load, it introduces a new challenge. When the data is subsequently up-scaled back to its original size, the interpolation process can lead to pixel-wise inaccuracies. These inaccuracies may not significantly affect the overall segmentation metrics but can result in slight deviations that impact the model's performance in tasks requiring high precision, such as absorption correction.

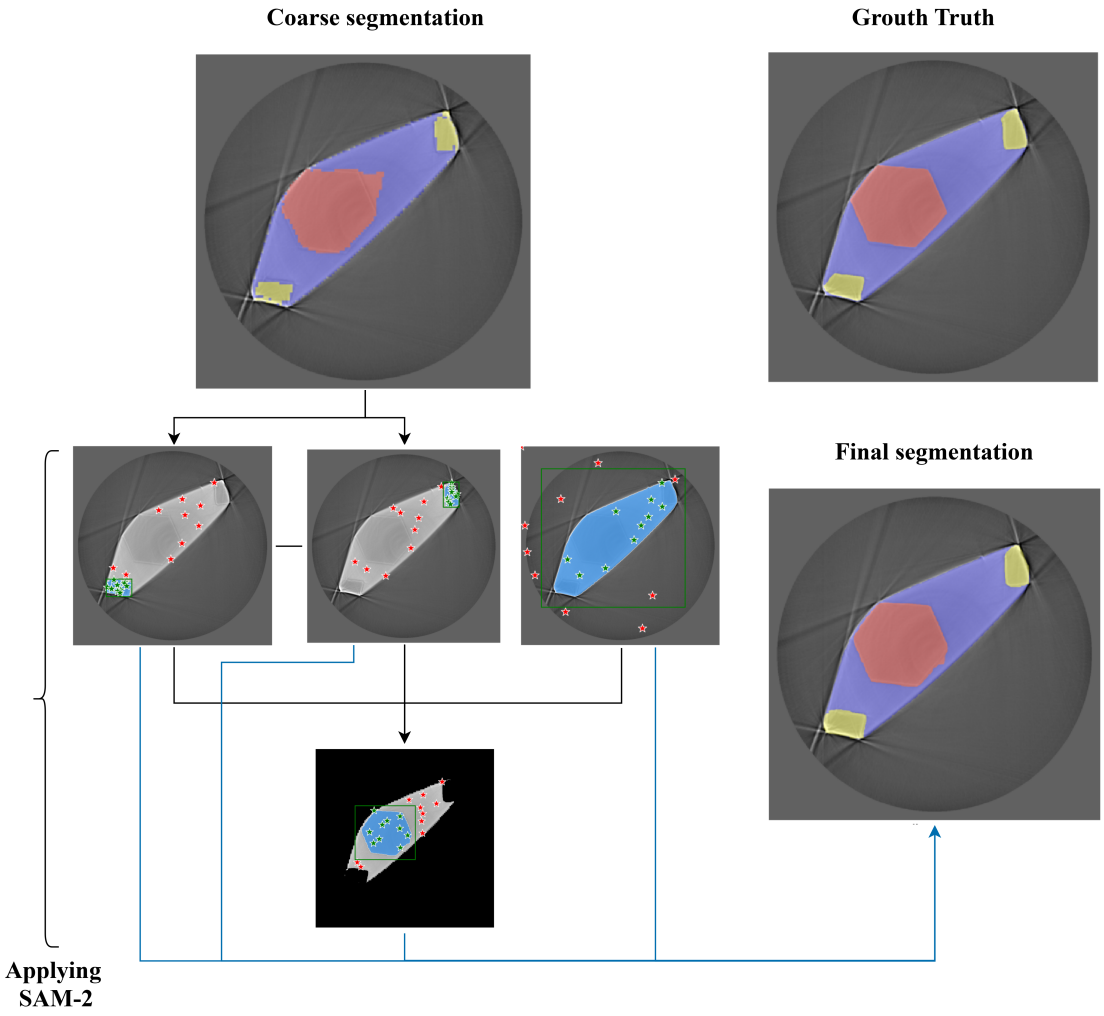


Figure 5.5: Wolkflow of applying SAM-2 to refine the segmentation details

Segment-anything-2 model (SAM-2) [144] is an advanced and super-large image segmentation model comprising 224.4 million parameters. It can utilise user-selected points and bounding boxes as prompts to achieve pixel-wise accurate segmentation. In light of the

challenges discussed, incorporating SAM-2 into the workflow can significantly enhance segmentation accuracy.

The workflow of applying SAM-2 is illustrated in Figure 5.5. In the process of applying SAM-2, the initial coarse segmentation results from the trained model are used to generate points and bounding box prompts for SAM-2. Specifically, 10 points are randomly sampled from the morphologically eroded area of interest, and a bounding box that is scaled to be 1.2 times larger than the area of interest is used as the prompt to SAM-2. This approach helps to provide more context to the model, improving the accuracy of the refined segmentation. The most ambiguous segmentation details typically occur at the noisy boundaries, where the streak effect due to phase shift is most significant. This is especially problematic for segmenting materials like crystals and the mounting loop, which are particularly susceptible to these errors. The workflow begins by examining the area of the mounting loop, as this region may be split into two separate areas in some slices. To identify this, the Connected Components Algorithm [145] is employed to determine whether there is a single region or multiple regions of the mounting loop in each slice. This step is crucial for correctly identifying and segmenting the loop, especially in complex scenarios where noise and artefacts might cause misidentification. Then, the background region is also determined by entering the points and bounding boxes of the non-background objects.

To further enhance SAM-2's performance, all regions except the mother liquor and the crystal are removed, as illustrated in Figure 5.5. This preprocessing step focuses the model on the most critical areas, reducing the influence of irrelevant structures and noise. The area of mother liquor is the region where it is not crystal, the mounting loop and the background. Finally, morphological closing is applied to all segmented materials to fill small holes and gaps, ensuring a smoother and more coherent final segmentation.

5.3 Experiments

5.3.1 Experimental setup

To examine the similarity and quality of the simulated images compared to the real images, the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are used as comparison metrics. SSIM is a perception-based model that considers changes in structural information, luminance, and contrast to measure the similarity between two images. The SSIM index between two images x and y is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5.12}$$

where μ_x and μ_y are the means of x and y, σ_x^2 and σ_y^2 are the variances of x and y, σ_{xy} is the covariance of x and y, and C_1 and C_2 are constants to stabilize the division.

On the other hand, PSNR measures the quality of the synthetic image by comparing it to the real-experiment image. It is expressed in terms of the Mean Squared Error (MSE) between the images x and y. The PSNR is defined as:

$$\text{PSNR}(x, y) = 10 \cdot \log_{10} \left(\frac{(\text{MAX}_I)^2}{\text{MSE}} \right) \tag{5.13}$$

where MAX_I is the maximum possible pixel value of the image, and the Mean Squared Error (MSE) is given by

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [x(i, j) - y(i, j)]^2. \tag{5.14}$$

Regarding the segmentation results, an NVIDIA RTX 3090 GPU is used for training and inference. 12 real datasets from manual segmentation and 108 selected synthetic datasets are used for training the model. Stochastic gradient (SGD) was employed with a base learning rate of 0.01 and a weight decay of 3×10^{-5}. During training, a combination of Cross-Entropy Loss and weighted Dice Loss is used to optimize the model. Cross-entropy loss measures the pixel-wise classification accuracy by comparing the predicted probability distribution to the true labels. The Dice Loss, on the other hand, measures the overlap

between the predicted and ground truth segmentations, focusing on the intersection-over-union of the predicted and actual regions. To address class imbalance, the Dice Loss is weighted such that less frequent or more important classes are given larger weights, ensuring that segmentation errors in those classes have a proportionally larger impact on the total loss. The weighted Dice Loss can be expressed as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{k=1}^{K} w_k |A_k \cap B_k|}{\sum_{k=1}^{K} w_k (|A_k| + |B_k|)},$$

where K is the number of material classes, A_k and B_k represent the predicted and ground truth segments for class k, and w_k is the weight assigned to class k. This weighting of crystal, mounting loop and mother liquor is $3 : 1 : 1$, and this weighting ratio was derived from the inverse of their relative voxel frequencies in the training dataset. Specifically, the crystal region accounts for approximately one-third of the number of voxels compared to either the mounting loop or the mother liquor.

The total loss function combining Cross-Entropy and weighted Dice Loss is then defined as:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{Dice}},$$

where $\mathcal{L}_{\text{CE}}$ is the Cross-Entropy Loss, and α and β are weighting coefficients that balance the contribution of each loss term. In this study, α and β are set as 0.1 and 0.9 respectively. For evaluation, the overlap accuracy of each material i is used to assess the performance of the segmentation. This accuracy, often quantified by metrics such as Intersection over Union (IoU) or Dice Coefficient, provides a robust measure of the segmentation performance for each material, taking into account both precision and recall. The accuracy for material i is given by:

$$\text{Accuracy}_i = \frac{2|A_i \cap B_i|}{|A_i| + |B_i|},$$

where A_i is the set of predicted segments for material i and B_i is the set of ground truth segments for material i.

Finally, we calculated the absorption factors by a sampling rate of 0.1%, which is proved to have the same effect with full sampling as discussed in §4.2.1. Then we compared the merging statistics and anomalous peak heights in the Fourier map for the datasets of Insulin at 3.01 Å , Thermolysin at 3.53 Å and Thaumatin at 4.03 Å . This comparison aims to assess the impact of segmentation on the final scaling results. The scaling strategy AAC and ACSH used here follows the method outlined in Chapter 3. AAC applies only analytical absorption corrections while ACSH performs analytical absorption correction followed by a spherical harmonic correction to mitigate systematic errors, such as segmentation artefacts. The calculation of anomalous peak heights was carried out using the procedure detailed in section §4.2.1.

5.3.2 Simulation results

A qualitative comparison of projection images is presented in Figure 5.6 between simulation results ((a)-(c) and (g)-(i)) and real experimental data ((d)-(f) and (j)-(l)) for two Thermolysin samples, labelled Sample A and Sample B. Each column corresponds to a different time point in the simulation or experiment: the initial projection (top row), the projection at the point of minimum PSNR (middle row), and the final projection image (bottom row). All images are linearly scaled to the same 0-255 grayscale range. However, differences in visual contrast between the simulated and experimental images are apparent. These differences do not arise from mismatched dynamic range, but rather from variations in how the intensity values are distributed. Simulated projections typically exhibit smooth attenuation profiles and sharp crystal boundaries, while experimental data include detector artefacts, noise, and background non-uniformity, which compress the effective use of the dynamic range. No additional histogram normalization or contrast equalization was applied, as the goal of this figure is to reflect the raw intensity characteristics used during training and evaluation. Preserving these differences ensures that the model is exposed to realistic variance in intensity distributions and allows a direct comparison of simulation fidelity against unprocessed experimental data. Addressing this contrast variation with

post-processing might improve visual consistency but would obscure the raw conditions under which the segmentation model must perform.

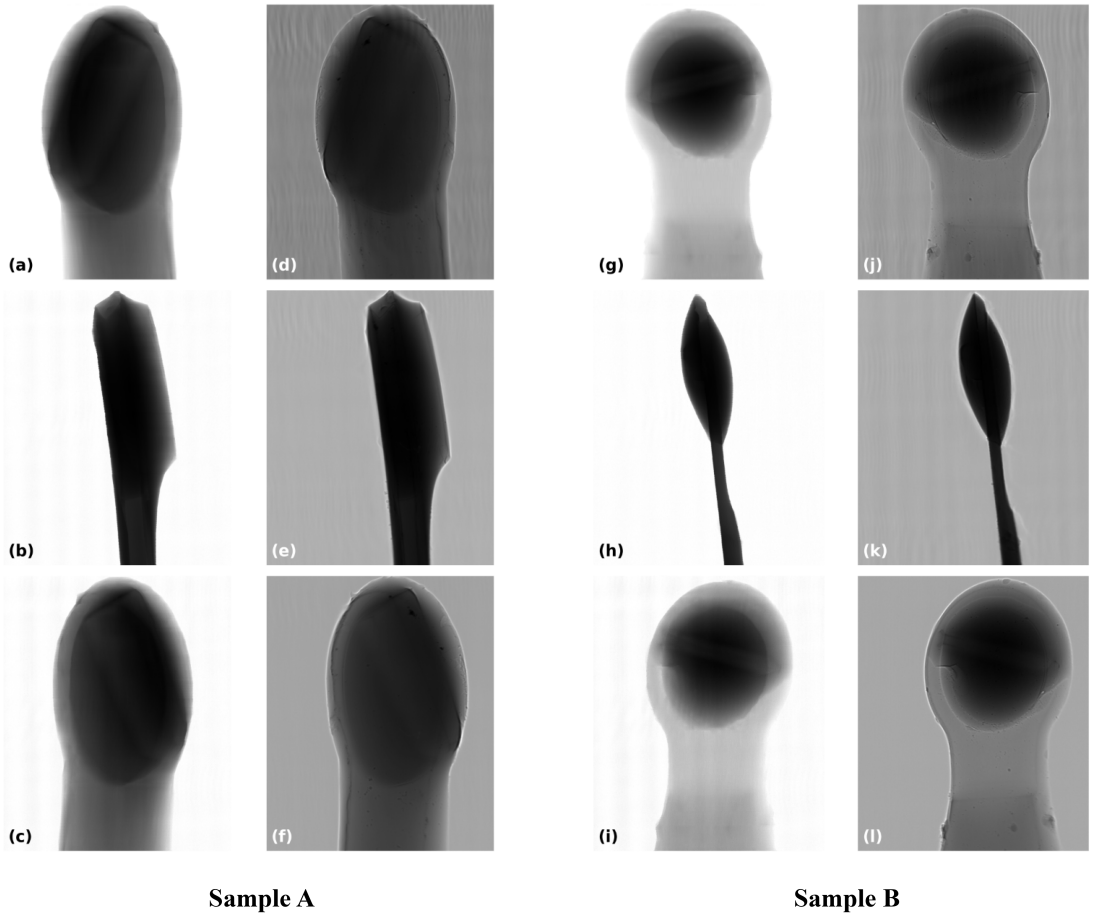


Figure 5.6: Qualitative comparison of projection images between simulation results ((a-c) and (g-i)) and real experimental results ((d-f) and (j-l)) for two Thermolysin samples. The three rows represent the initial image, the image at the point where the PSNR is at its minimum peak (second minimum in Sample A), and the final image, respectively

The performances on SSIM (Structural Similarity Index), PSNR (Peak Signal-to-Noise Ratio), and MSE (Mean Squared Error) across a list of projection images of two real Thermolysin samples A and B are shown in Figure 5.7. The simulation results are obtained from the segmentations of the real data. In Sample A, the SSIM scores fluctuate between 0.86 and 0.92, indicating a consistently high level of structural similarity throughout the projections, with minor variations. The PSNR scores for Sample A vary between 28 and 30, reflecting good image quality across the projections. The MSE for Sample A remains

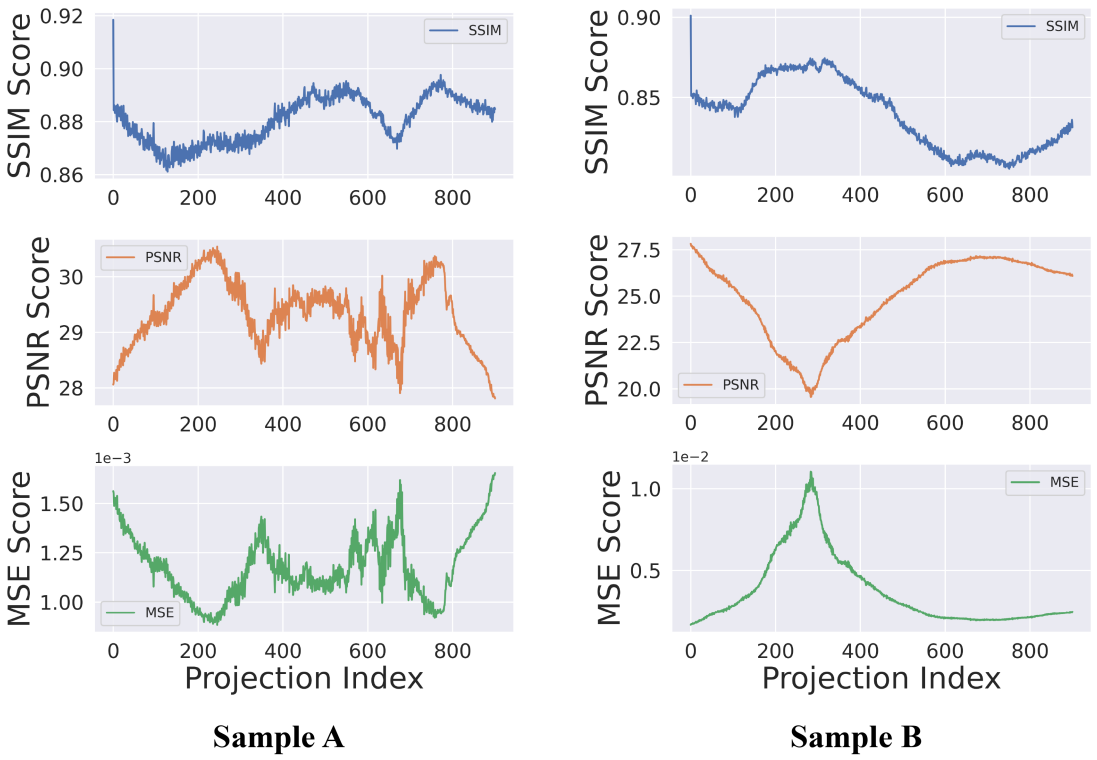


Figure 5.7: Quantitative comparison of projection images between simulation results and real experimental results for two Thermolysin samples.

low, mostly below 0.002, suggesting minor discrepancies between the synthetic and real images. In contrast, Sample B exhibits lower SSIM scores, ranging from 0.85 to 0.90. The PSNR scores in Sample B range from 20 to 27.5 and the MSE for Sample B is higher, with peaks reaching around 0.01. Overall, despite Sample B's metrics being slightly worse, they still reflect a good level of performance in terms of image simulation quality.

Notably, there is a significant drop in SSIM right after the first projection image for both samples in Figure 5.7. The flat-field correction used to remove the background in the real experiment relies on background images recorded before the experiment, where no object is present. As a result, the initial projection images experience minimal impact from background noise, as shown in Figure 5.6. However, due to the instability of the X-ray beam during the experiment, the effectiveness of flat-field correction varies for subsequent projections. While the first projection benefits from a more accurate background subtraction, the later images undergo imperfect flat-field correction, affecting their SSIM

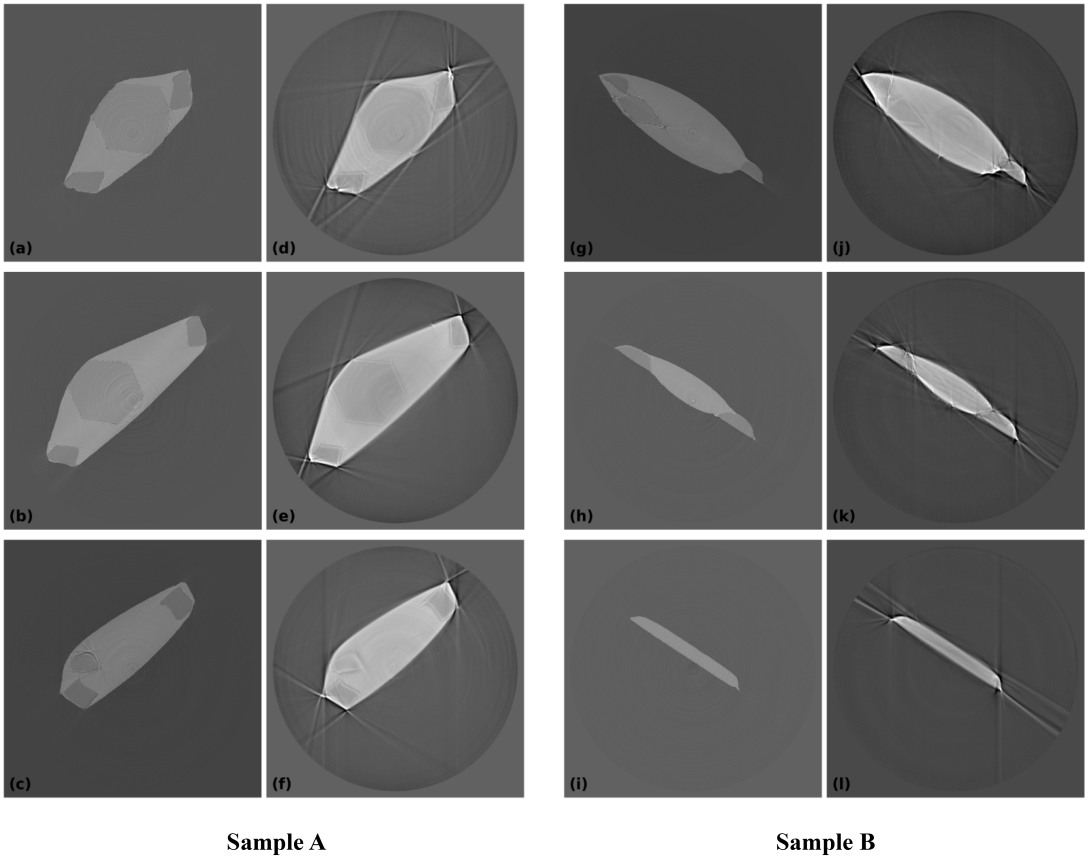


Figure 5.8: Qualitative comparison of reconstruction slice images between simulation results ((a-c) and (g-i)) and real experimental results ((d-f) and (j-l)) for two real Thermolysin samples A and B. The three rows represent slice index of 200,500 and 800 for Sample A and 400, 600 and 800 for Sample B.

scores and introducing variations in the background.

Figure 5.9 shows the same metrics across different slices of the reconstruction by FBP for two Thermolysin samples, A and B. In Sample A, the SSIM score begins around 0.95 and shows a gradual decrease, reaching a low point around 0.85 near slice index 700, before increasing again towards the later slices, indicating that it shows good structural similarity across slices of the reconstruction. In Sample B, the SSIM score follows a similar pattern but with slightly lower values overall. It starts near 0.9, decreases more steeply to around 0.82 near slice index 400, and exhibits a fluctuating pattern in the later slices. Overall, the PSNR and MSE of both samples show large variations and poor performance. This is because of differences in background regions in the reconstruction slices, as illustrated in

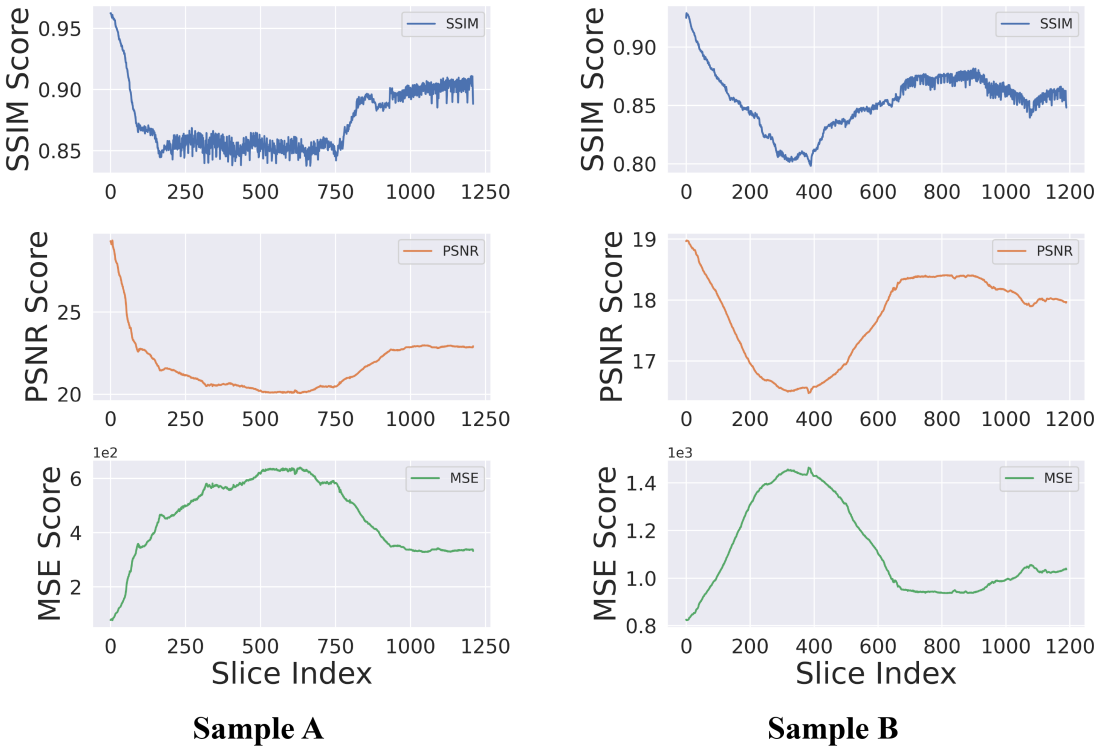


Figure 5.9: Comparison of reconstruction images (by FBP) between simulation results and real experimental results for two Thermolysin samples.

Figure 5.8. They are less indicative of true image quality because they are significantly influenced by large pixel-wise intensity variations. The SSIM, which focuses more on structural similarity, provides a more reliable indication of image quality in this context. Despite the fluctuations in SSIM scores, both samples show relatively good structural similarity between the synthetic and real images.

Figure 5.10 presents synthetic samples (C, D, E, F) at projection angles of 0°, 60°, and 90°, effectively illustrating the preservation of the spatial relationships between different materials and structural details, such as the fluid property of the mother liquor. The phase contrast between the crystal and the surrounding mother liquor is clearly evident, which would contribute to the streak effect observed around the edges during reconstruction. Moreover, the sharp edges visible at all angles demonstrate the simulation technique's accuracy in modelling real experimental conditions.

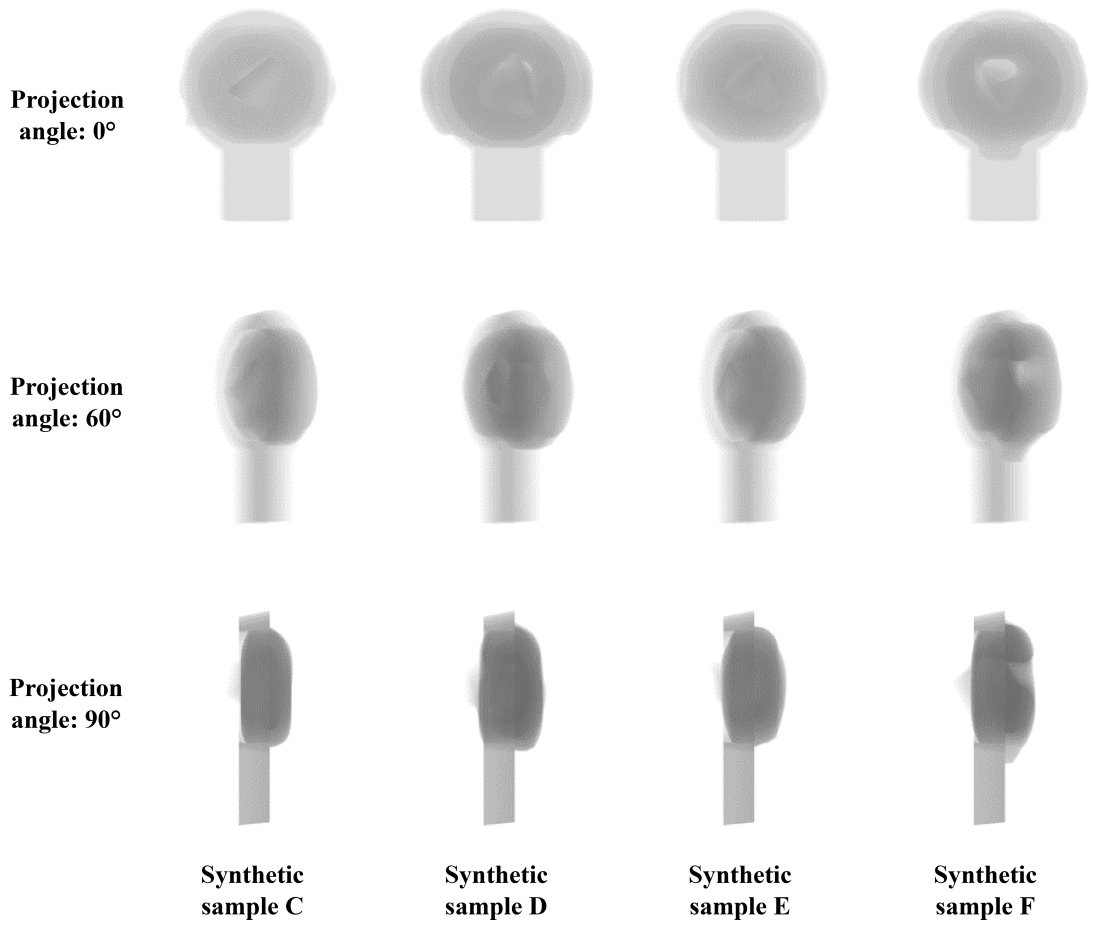


Figure 5.10: Synthetic projection images of four synthetic samples (C, D, E, F) at various projection angles (0°, 60°, and 90°).

Then, the projection images are reconstructed using Filtered Back-Projection (FBP), as shown in Figure 5.8. The figure presents reconstructed slices of synthetic samples (C, D, E, F) at slice indices 300, 400, and 500, highlighting several important features. The streak artefacts are consistently preserved across all samples, demonstrating the simulation method's effectiveness in accurately capturing these details. The spatial relationships between different materials are clearly maintained, representing the distinct separations and interactions between components are accurately reconstructed. Furthermore, the blurring observed around certain material boundaries reflects the method's capability to replicate experimental artefacts, emphasizing its effectiveness. This combination of

preserved features and artefacts confirms the simulation method's fidelity in modelling realistic X-ray tomography imaging experiments.

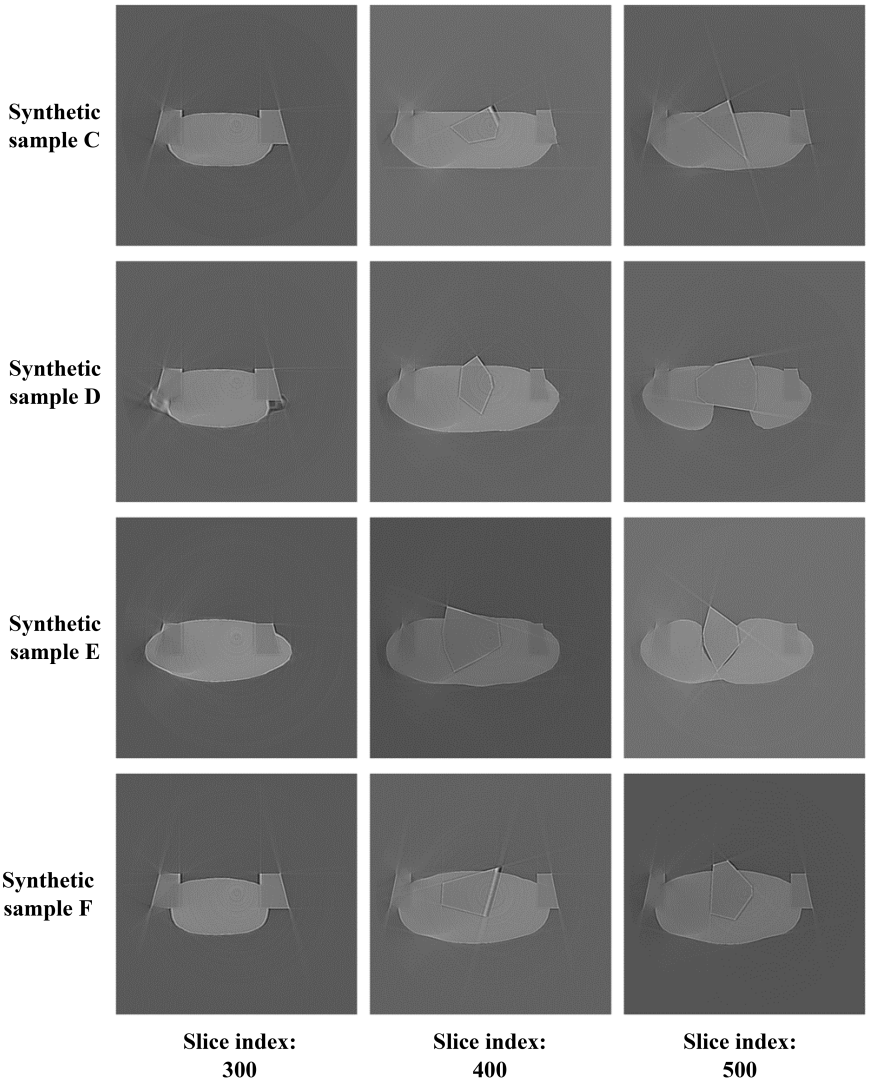


Figure 5.11: Synthetic reconstruction slice images of four synthetic samples (C, D, E, F) at various slice indices (300, 400, and 500).

5.3.3 Segmentation results

The results of the segmentation performance of Insulin, Thermolysin and Thaumatin using different techniques are presented in Table 5.1. In this comparison:

- **AnACorNet_R**: indicates the model trained solely on the real dataset.

- **AnACorNet_R_SAM** represents the model trained on the real dataset and further refined using the SAM-2 model.
- **AnACorNet_RS** denotes the model trained on a combination of both real and synthetic datasets.
- **AnACorNet_RS_SAM** refers to the model trained on both real and synthetic datasets, followed by refinement with the SAM-2 model.

The details of these three dataset are the same as section §4.2.1. The model with the largest accuracy in identifying crystals was selected for evaluation.

Table 5.1: Comparison of different segmentation models for Thermolysin and Thaumatin dataset in terms of accuracy (Dice Loss) for each class and Cross-entropy (CE) loss.

Method	Accuracy (Dice Loss)				CE Loss
	Background	Mother Liquor	Loop	Crystal	
Insulin					
AnACorNet_R	99.93% (0.0008)	57.57% (0.4505)	49.18% (0.4167)	16.46% (0.9514)	0.1138
AnACorNet_R_SAM	99.86% (0.0013)	75.13% (0.3621)	64.58% (0.2765)	32.71% (0.7615)	0.1120
AnACorNet_RS	99.93% (0.0010)	81.69% (0.3364)	61.20% (0.2781)	45.42% (0.2929)	0.1034
AnACorNet_RS_SAM	99.94% (0.0008)	80.44% (0.2239)	82.32% (0.1462)	81.57% (0.2025)	0.0371
Thermolysin					
AnACorNet_R	99.95% (0.0014)	82.68% (0.1770)	43.48% (0.4336)	80.43% (0.3134)	0.1839
AnACorNet_R_SAM	99.90% (0.0013)	88.15% (0.1258)	82.50% (0.1241)	88.26% (0.1394)	0.0706
AnACorNet_RS	99.88% (0.0013)	93.93% (0.1014)	78.10% (0.1426)	90.06% (0.0813)	0.0541
AnACorNet_RS_SAM	99.89% (0.0013)	93.47% (0.0712)	88.11% (0.0938)	96.54% (0.0382)	0.0327
Thaumatin					
AnACorNet_R	99.94% (0.0014)	86.51% (0.1989)	52.52% (0.3652)	69.42% (0.2892)	0.1341
AnACorNet_R_SAM	99.85% (0.0012)	93.80% (0.0978)	87.06% (0.1093)	88.38% (0.0776)	0.0360
AnACorNet_RS	99.82% (0.0012)	95.15% (0.0945)	81.40% (0.1286)	92.91% (0.0589)	0.0368
AnACorNet_RS_SAM	99.88% (0.0011)	93.38% (0.0716)	90.45% (0.0858)	96.42% (0.0415)	0.0222

Table 5.1 provides a detailed comparison of different segmentation models applied to the Insulin, Thermolysin, and Thaumatin dataset, focusing on accuracy across four categories (Background, Mother Liquor, Loop, and Crystal) and evaluating the models based on Accuracy, Dice Loss and Cross-Entropy (CE) Loss. In analyzing the Insulin dataset, it is clear that `AnACorNet_RS`, which utilizes both synthetic and real data for training, significantly outperforms `AnACorNet_R`, which is trained solely on real data. This advantage is particularly evident in challenging materials such as the Loop and Crystal categories. `AnACorNet_RS` achieves an accuracy of 45.42% in the Crystal category with a reduced

Dice Loss of 0.2929, by contrast to `AnACorNet_R`, which only reaches an accuracy of 16.46% and a Dice Loss of 0.9514. The incorporation of synthetic data in `AnACorNet_RS` clearly contributes to better generalization and segmentation performance. Furthermore, when the SAM-2 is applied to both models, the differences become even more prominent. `AnACorNet_R_SAM` shows improvements over `AnACorNet_R`, but `AnACorNet_RS_SAM` further elevates the performance across most categories, achieving an accuracy of 81.57% in the Crystal category with a Dice Loss of 0.2025 and the lowest Cross-Entropy Loss of 0.0371 among all models. This demonstrates the synergistic effect of combining synthetic data with SAM, leading to a more robust and accurate model.

In the Thermolysin dataset, `AnACorNet_RS` already exhibits enhanced performance over `AnACorNet_R` across multiple categories, particularly in the Loop and Crystal regions. When SAM-2 is incorporated, `AnACorNet_RS_SAM` achieves the highest accuracy in the Crystal category at 96.54% with a minimal Dice Loss of 0.0382, far surpassing the performance of `AnACorNet_R_SAM`. This indicates that the combined use of synthetic data and SAM significantly boosts the model's ability to accurately segment even the most challenging structures.

Similarly, for the Thaumatin dataset, `AnACorNet_RS` demonstrates superior accuracy over `AnACorNet_R`, especially after the application of SAM. `AnACorNet_RS_SAM` stands out with high accuracy in almost all categories, achieving 90.45% in the Loop category with a Dice Loss of 0.0858, and 96.42% accuracy in the Crystal category with a Dice Loss of 0.0415. This further emphasizes that the integration of SAM-2 into a model trained with both synthetic and real data results in a significant enhancement of segmentation performance.

Overall, the comparison shows that `AnACorNet_RS` trained with both synthetic and real data is consistently superior to `AnACorNet_R` trained solely on real data. The addition of SAM to both models further improves their performance, with `AnACorNet_RS_SAM` consistently emerging as the best-performing model across all datasets. This comprehensive analysis underscores the importance of synthetic data and refinement by SAM in achieving more

accurate and efficient segmentation in X-ray crystallography.

Table 5.2: Comparison of running time for different segmentation models across Insulin, Thermolysin and Thaumatin dataset.

Method	**Inference running Time**		
	Insulin	**Thermolysin**	**Thaumatin**
AnACorNet	9.24 seconds	25.59 seconds	20.85 seconds
AnACorNet + SAM	≈35 minutes	≈70 minutes	≈60 minutes
Manual Labor	≈≥ 4 hours	≈≥ 4 hours	≈≥ 4 hours

Table 5.2 provides a comparison of inference running times for different segmentation methods applied to the Insulin, Thermolysin, and Thaumatin datasets. The basic AnACorNet model can complete the segmentation tasks within one minute for each dataset: 9.24 seconds for Insulin, 25.59 seconds for Thermolysin, and 20.85 seconds for Thaumatin. However, after applying SAM-2, the processing time increases substantially. The AnACorNet + SAM variant takes approximately 35 minutes for the Insulin dataset, 70 minutes for Thermolysin, and 60 minutes for Thaumatin. Despite the increased computational cost, this approach offers enhanced segmentation performance, as previously noted in the accuracy table, making it a valuable option when high precision is required. On the other hand, manual segmentation by an expert user is highly time-consuming, taking at least 4 hours per dataset. This shows the efficiency advantage of automated methods like AnACorNet and its variants, which can significantly reduce the time and effort required for such tasks.

5.3.4 Absorption correction results

Figure 5.12 presents comparative histograms of absorption factors between manual segmentation (ground truth) and segmentations from AnACorNet_RS and AnACorNet_RS_SAM for three substances: Insulin, Thermolysin, and Thaumatin. The histograms illustrate that AnACorNet_RS_SAM provides a closer alignment with the ground truth across all substances, as indicated by the significant overlap between the model's segmentation and the ground truth in each subplot. This alignment is particularly well-supported by

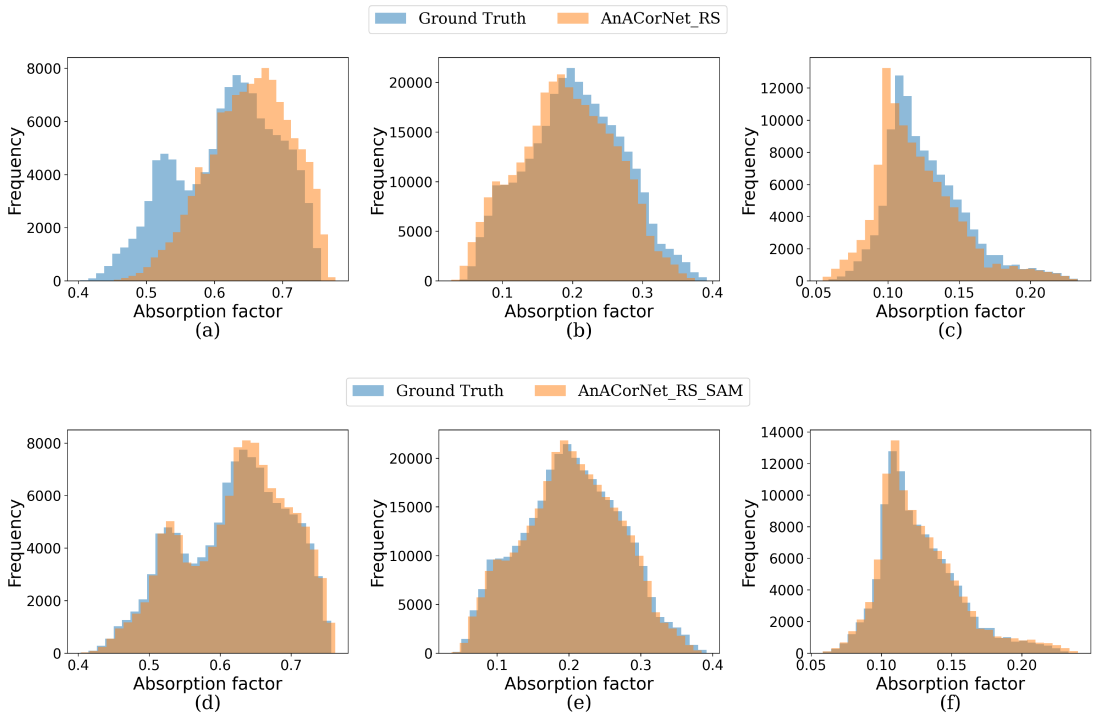


Figure 5.12: Comparative histograms of absorption factors between manual segmentation (ground truth) and segmentations from AnACorNet_RS and AnACorNet_RS_SAM. (a,d), (b,e), and (c,e) are from Insulin, Thermlysin and Thaumatin respectively.

the quantitative results in Table 5.1, where AnACorNet_RS_SAM demonstrates superior performance in terms of Dice Loss across different classes, especially for challenging classes like the Loop and Crystal.

Although AnACorNet_RS does not perform as well as AnACorNet_RS_SAM for Thermolysin and Thaumatin, as shown in subplots (b) of Thermolysin and (c) of Thaumatin, its segmentation results still provide reasonable approximations to the ground truth with similarly shaped histograms. This is reflected in the Table 5.1, where AnACorNet_RS achieves satisfactory Dice Loss values for these dataset. However, in the case of Insulin (subplot (a)), AnACorNet_RS shows a noticeable discrepancy compared to the ground truth, which is confirmed by the statistical data in Table 5.1, indicating less accurate segmentation for this particular sample.

The merging statistics for the Insulin, Thermolysin, and Thaumatin datasets, as presented in Tables 5.3, 5.4, and 5.5, provide further insights into the performance of AnACorNet_RS

Metric	Method	Manual	AnACorNet_RS	AnACorNet_RS_SAM
Resolution limit	AAC		55.26 - 2.05 (2.09 - 2.05)	
	ACSH		55.26 - 2.05 (2.09 - 2.05)	
Completeness (%)	AAC		100.0 (98.3)	
	ACSH		100.0 (98.3)	
Multiplicity	AAC		22.4 (9.0)	
	ACSH		22.4 (9.0)	
I/sigma	AAC	23.6 (1.3)	19.2 (1.1)	23.7 (1.3)
	ACSH	24.4 (1.3)	23.9 (1.3)	24.4 (1.3)
Rmerge	AAC	0.096 (0.754)	0.109 (0.756)	0.096 (0.753)
	ACSH	0.092 (0.760)	0.093 (0.760)	0.092 (0.760)
Rmeas	AAC	0.098 (0.798)	0.111 (0.801)	0.098 (0.798)
	ACSH	0.094 (0.805)	0.095 (0.805)	0.094 (0.805)
Rpim	AAC		0.019 (0.262)	
	ACSH		0.018 (0.264)	
CC half	AAC	1.000 (0.775)	0.999 (0.768)	1.000 (0.775)
	ACSH		1.000 (0.763)	
Anomalous correlation	AAC	0.806 (-0.141)	0.735 (-0.152)	0.805 (-0.139)
	ACSH	0.796 (-0.165)	0.800 (-0.169)	0.797 (-0.164)
Anomalous slope	AAC	0.857	0.760	0.862
	ACSH	0.864	0.800	0.865
Total reflection number	AAC	114443 (2139)	114549 (2139)	114442 (2139)
	ACSH	114453 (2139)	114460 (2139)	114452 (2139)
Unique reflection number	AAC		5106 (238)	
	ACSH		5106 (238)	

Table 5.3: Comparison of merging statistics of Insulin results from Ground Truth, AnACorNet_RS, and AnACorNet_RS_SAM for both AAC and ACSH scaling methods mentioned in Chapter 3. The values in brackets represent high-resolution statistics.

and AnACorNet_RS_SAM compared to the ground truth.

For Insulin, the metrics between the ground truth and AnACorNet_RS_SAM are nearly identical across both AAC and ACSH scaling methods, indicating that the segmentation differences have minimal impact on the merging results, as shown in Table 5.3. Specifically, metrics such as *Rmerge*, *Rpim*, *CC half*, *Anomalous correlation*, and *Anomalous slope* show negligible differences between AnACorNet_RS_SAM and the ground truth, suggesting a high-quality segmentation. For example, the *Rmerge* under ACSH scaling is consistent

between the ground truth and AnACorNet_RS_SAM (0.092), while the *CC half* remains perfect (1.000) under both scaling methods, confirming the fidelity of AnACorNet_RS_-SAM. This is consistent with the histogram (a) in Figure 5.12, where AnACorNet_RS_SAM closely matches the ground truth distribution for both scaling methods.

In contrast, while AnACorNet_RS for Insulin also shows very similar values to the ground truth for most metrics, there are minor deviations. Under AAC scaling, the *Rmerge* is slightly higher (0.109 vs. 0.096) compared to both the ground truth and AnACorNet_RS_-SAM. Additionally, the *Anomalous slope* is lower (0.760 vs. 0.857) and the *I/sigma* value is also reduced (19.2 vs. 23.6), indicating large discrepancies with the manual segmentation. However, after applying the ACSH scaling method, these discrepancies are reduced. The *Rmerge* and *Anomalous slope* values improve significantly, showing closer alignment with the manual segmentation results.

Similar to Insulin, the Thermolysin results, as shown in Table 5.4 show that the metrics between the ground truth and AnACorNet_RS_SAM are highly consistent across both AAC and ACSH scaling methods. Key metrics like *Rmerge*, *I/sigma*, and *Anomalous slope* show minimal differences, indicating that segmentation quality is well-preserved. Under AAC scaling, the *Rmerge* for AnACorNet_RS_SAM closely matches the ground truth (0.133 vs. 0.134), and *CC half* remains stable at 0.997. On the other hand, AnACorNet_RS exhibits some minor deviations from the ground truth, similar to the pattern observed in Insulin. Specifically, the *Rmerge* is slightly higher (0.141 vs. 0.134) and the *Anomalous slope* is marginally lower. However, after applying ACSH scaling, these discrepancies are significantly reduced, demonstrating that spherical harmonics effectively address segmentation artefacts, improving the data quality to levels comparable with the manual segmentation results.

In Table 5.5, the Thaumatin results demonstrate high consistency between the ground truth and AnACorNet_RS_SAM across both AAC and ACSH scaling methods, with minimal variation. For instance, under AAC scaling, the *Rmerge* and *I/sigma* for AnACorNet_RS_-SAM closely matches the ground truth (0.084 vs. 0.082) and (37.6 vs 37.7) respectively.

Metric	Method	Manual	AnACorNet_RS	AnACorNet_RS_SAM
Resolution limit	AAC		129.26 - 2.31 (2.35 - 2.31)	
	ACSH		129.26 - 2.31 (2.35 - 2.31)	
Completeness (%)	AAC		96.5 (90.2)	
	ACSH		96.5 (90.2)	
Multiplicity	AAC		21.3 (9.4)	
	ACSH		21.3 (9.4)	
I/sigma	AAC	26.6 (7.9)	26.0 (7.8)	26.5 (7.9)
	ACSH	25.5 (5.6)	25.9 (5.8)	25.6 (5.6)
Rmerge	AAC	0.134 (0.345)	0.141 (0.352)	0.133 (0.344)
	ACSH	0.102 (0.367)	0.104 (0.370)	0.102 (0.365)
Rmeas	AAC	0.137 (0.364)	0.144 (0.371)	0.135 (0.363)
	ACSH	0.105 (0.387)	0.106 (0.390)	0.104 (0.385)
Rpim	AAC		0.027 (0.113)	
	ACSH		0.021 (0.120)	
CC half	AAC	0.996 (0.943)	0.997 (0.939)	0.997 (0.944)
	ACSH		0.998 (0.944)	
Anomalous correlation	AAC	-0.159 (-0.521)	-0.210 (-0.520)	-0.167 (-0.524)
	ACSH	0.006 (-0.527)	-0.015 (-0.526)	-0.002 (-0.530)
Anomalous slope	AAC	1.096	1.081	1.092
	ACSH	0.819	0.839	0.818
Total reflection number	AAC	308760 (6189)	308944 (6189)	308778 (6189)
	ACSH	309125 (6189)	309143 (6189)	309137 (6189)
Unique reflection number	AAC		14513 (656)	
	ACSH		14513 (656)	

Table 5.4: Comparison of merging statistics of Thermolysin results from Ground Truth, AnACorNet_RS, and AnACorNet_RS_SAM for both AAC and ACSH scaling methods mentioned in Chapter 3. The values in brackets represent high-resolution statistics.

However, they become better under ACSH scaling, where *Rmerge* becomes smaller,*I/sigma* and *Anomalous slope* become larger. AnACorNet_RS shows more discrepancies, similar to patterns seen in Insulin and Thermolysin when using AAC scaling. The *Rmerge* is slightly higher (0.086 vs. 0.082), and the *Anomalous slope* and *I/sigma* are also slightly reduced compared to manual segmentation. After applying ACSH scaling, these differences are reduced and become negligible.

Figure 5.13 presents the peak heights for individual atoms in the datasets of Insulin

Metric	Method	Manual	AnACorNet_RS	AnACorNet_RS_SAM
Resolution limit	AAC		150.73 - 2.70 (2.75 - 2.70)	
	ACSH		150.96 - 2.70 (2.75 - 2.70)	
Completeness (%)	AAC		99.2 (90.8)	
	ACSH		99.2 (90.8)	
Multiplicity	AAC		13.9 (5.4)	
	ACSH		13.9 (5.4)	
I/sigma	AAC	37.7 (18.9)	35.5 (17.5)	37.6 (19.0)
	ACSH	58.0 (28.9)	57.3 (28.3)	60.9 (30.7)
Rmerge	AAC	0.082 (0.098)	0.086 (0.094)	0.084 (0.094)
	ACSH	0.061 (0.082)	0.061 (0.083)	0.058 (0.074)
Rmeas	AAC	0.085 (0.108)	0.089 (0.103)	0.087 (0.103)
	ACSH	0.063 (0.090)	0.063 (0.091)	0.060 (0.082)
Rpim	AAC	0.022 (0.043)	0.023 (0.041)	0.022 (0.041)
	ACSH	0.016 (0.036)	0.016 (0.037)	0.015 (0.033)
CC half	AAC		0.997 (0.992)	
	ACSH		0.998 (0.993)	
Anomalous correlation	AAC	0.572 (0.315)	0.551 (0.363)	0.565 (0.302)
	ACSH	0.812 (0.547)	0.810 (0.529)	0.838 (0.617)
Anomalous slope	AAC	2.781	2.615	2.835
	ACSH	4.142	4.070	4.348
Total reflection number	AAC	105385 (1873)	105392 (1873)	105381 (1873)
	ACSH	105278 (1873)	105280 (1873)	105264 (1873)
Unique reflection number	AAC		7580 (345)	
	ACSH		7580 (345)	

Table 5.5: Comparison of merging statistics of Thaumatin results from Ground Truth, AnACorNet_RS, and AnACorNet_RS_SAM for both AAC and ACSH scaling methods mentioned in Chapter 3. The values in brackets represent high-resolution statistics.

(a), Thermolysin (b), and Thaumatin (c), comparing the results from AnACorNet_RS, AnACorNet_RS_SAM, and the Ground Truth (manual segmentation) under AAC scaling. Consistent with the merging statistics, AnACorNet_RS_SAM closely aligns with the Ground Truth (green circles) for all atoms across the Insulin and Thermolysin datasets, with slightly poorer performance on Thauatmin. In contrast, AnACorNet_RS generally follows the trend of the Ground Truth but shows greater variations in peak height values. This discrepancy reflects the differences observed in the merging statistics, where metrics such

as *Rmerge*, *I/sigma*, and *Anomalous slope* slightly deviate from the manual segmentation results. This is particularly noticeable in Thaumatin.

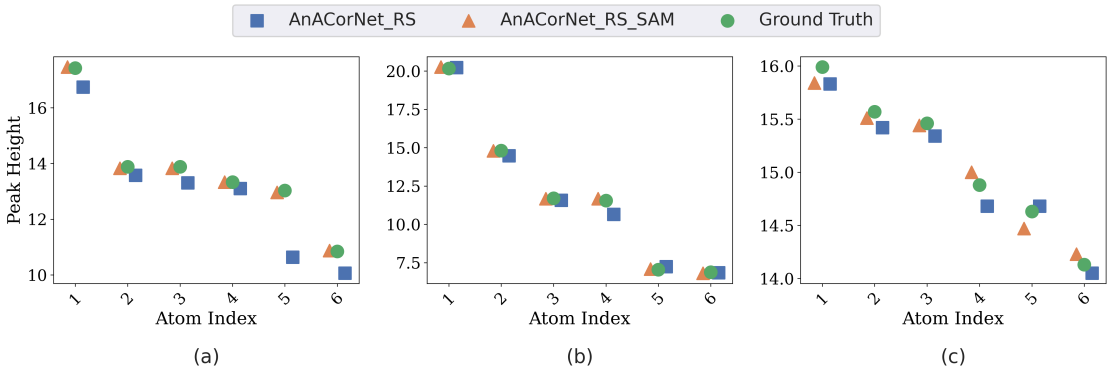


Figure 5.13: Anomalous peak height (σ) differences of top 6 highest peaks between manual segmentation (ground truth) and segmentations from AnACorNet_RS and AnACorNet_-RS_SAM of AAC scaling. (a), (b) and (c) are from Insulin, Thermlysin and Thaumatin respectively.

After applying ACSH, as illustrated in Figure 5.14, both AnACorNet_RS and AnACorNet_RS_SAM produce peak heights closely matching those obtained through manual segmentation. Notably, in the case of the Thaumatin dataset, AnACorNet_RS_SAM even achieves higher values for the first three highest peaks compared to the Ground Truth. This observation aligns with the improvements observed in the merging statistics of Thaumatin, as detailed in Table5.5.

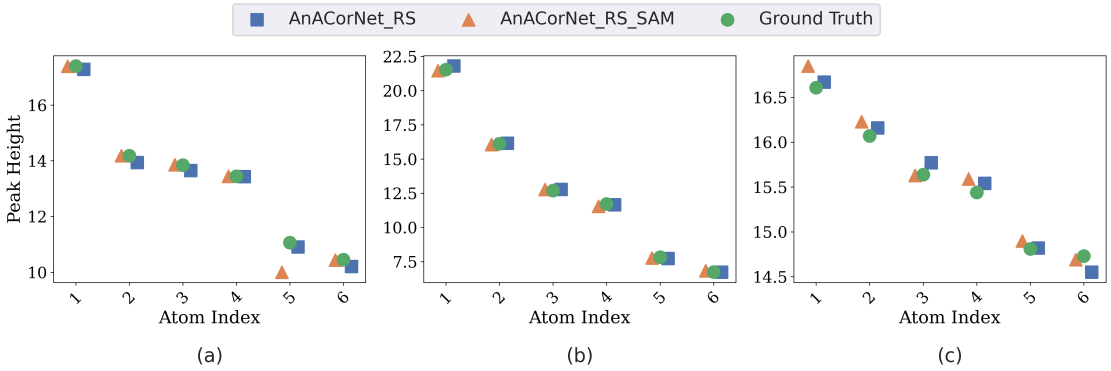


Figure 5.14: Anomalous peak height (σ) differences of top 6 highest peaks between manual segmentation (ground truth) and segmentations from AnACorNet_RS and AnACorNet_-RS_SAM. (a), (b) and (c) are from Insulin, Thermlysin and Thaumatin respectively.

5.4 Discussion

The results from our experiments provide a comprehensive evaluation of the proposed segmentation approach, particularly focusing on the performance of AnACorNet variants in comparison to the ground truth. The quantitative assessment of the image simulation quality using SSIM, PSNR, and MSE indicates a high level of structural similarity between the simulated and real images, especially for real Thermolysin samples (Sample A and Sample B). In Sample A, the SSIM scores fluctuate between 0.86 and 0.92, reflecting consistently high structural similarity throughout the projections. However, Sample B exhibits slightly lower SSIM scores ranging from 0.85 to 0.90. The noticeable drop in SSIM right after the first projection image in both samples is primarily attributed to the limitations of the flat-field correction used in the real experiments.

Although the initial projection images appear clean due to flat-field correction using pre-recorded background (flat-field and dark-field) images, this method assumes that imaging conditions remain constant throughout the scan. In reality, changes such as beam intensity fluctuations, detector drift, or sample movement occur over time, causing the flat-field reference to become outdated. As a result, the correction becomes less accurate in later projections, leading to background variations and lower SSIM scores.

The reconstruction results further reinforce the robustness of the simulation, with SSIM scores for both samples indicating good structural similarity. Despite fluctuations in the PSNR and MSE values, the SSIM scores provide a more reliable measure of image quality, underscoring the structural accuracy of the simulations. Qualitative comparisons, as illustrated in Figures 5.6 and 5.8, reveal that the simulated images maintain a close resemblance to the real experimental results, validating the effectiveness of our simulation approach in preserving structural details. In addition, the tomography projection images and reconstruction slice images of synthetic samples C, D, E and F, as shown in Figures 5.10 and 5.11, demonstrate the preservation of phase shifts, streak effects, and blurred edges, which are commonly observed in real X-ray tomography reconstruction experiments.

The segmentation performance across the Insulin, Thermolysin, and Thaumatin datasets

shows that incorporating synthetic data into the training process significantly enhances model accuracy. The AnACorNet_RS model, trained on both real and synthetic datasets, demonstrates a marked improvement in segmenting challenging regions such as Loop and Crystal, achieving Dice Loss values that are consistently lower than those of AnACorNet_R trained solely on real data. The integration of SAM-2 further boosts performance, with AnACorNet_RS_SAM achieving the highest accuracy across all categories and datasets, particularly for the Crystal class. This model's superior performance is evidenced by its ability to reduce Dice Loss substantially and to maintain lower Cross-Entropy Loss values, highlighting the effectiveness of using both synthetic data and SAM refinement to improve segmentation quality.

The efficiency of the segmentation methods is also noteworthy. Table 5.2 shows that while AnACorNet offers rapid inference times, the inclusion of SAM-2 increases processing time significantly. However, this trade-off is justified by the enhanced segmentation accuracy, as AnACorNet_RS_SAM consistently outperforms the other methods. This highlights the potential of SAM-2 as a valuable post-processing step when precision is important, despite its higher computational cost (about 30 minutes in Nvidia RTX 3090). Furthermore, the automated methods provide a substantial reduction in manual labour time (usually > 4 hours), showcasing their practical utility in high-throughput environments.

For Thermolysin and Thaumatin, although AnACorNet_RS shows minor deviations in metrics like *Rmerge* and *Anomalous slope*, it still provides a satisfactory approximation to the ground truth. This suggests that while AnACorNet_RS_SAM achieves optimal performance, AnACorNet_RS can still offer a reliable segmentation solution in cases where computational resources or time constraints limit the use of SAM-2.

The analysis of anomalous peak heights in Figure 5.13 demonstrates that AnACorNet_-RS_SAM provides a close replication of the ground truth measurements, with minimal deviation across all datasets. This indicates that AnACorNet_RS_SAM can effectively capture the intricate details of peak height features, ensuring that the segmentation's impact on subsequent data processing steps is minimal. While AnACorNet_RS also follows the

general trend of the ground truth, its slightly greater variation points out the added value of SAM-2 in refining the segmentation to achieve higher precision.

The comparison of absorption factors, as presented in Figure 5.12, shows that AnACorNet_-RS_SAM aligns more closely with the ground truth than AnACorNet_RS, particularly for Insulin and Thermolysin. The histograms demonstrate that AnACorNet_RS_SAM achieves a distribution that significantly overlaps with the ground truth, confirming its superior segmentation accuracy. The quantitative merging statistics in Tables 5.3, and 5.4 and anomalous peak heights in Figure 5.13 and 5.14 support this observation. They reveal that AnACorNet_RS_SAM produces metrics almost indistinguishable from the ground truth across Insulin and Thermolysin for both AAC and ACSH scaling methods. In the case of Thaumatin, AnACorNet_RS_SAM even achieve better performance after applying ACSH as shown in Table 5.5.

AnACorNet_RS exhibits deviations in metrics under AAC scaling across all datasets. However, after applying ACSH, the differences compared to the ground truth become minimal. This indicates that while AAC alone may not match the effectiveness of manual segmentation, the enhancements observed with ACSH scaling demonstrate that the use of spherical harmonics helps mitigate segmentation artefacts, improving data quality and aligning AnACorNet_RS more closely with manual segmentation values. Although AnACorNet_RS_SAM achieves the best performance, AnACorNet_RS remains a reliable segmentation option with ACSH scaling, as described in Chapter 3, particularly when computational resources or time constraints limit the use of SAM-2.

These findings emphasize the importance of integrating synthetic data and advanced post-processing techniques like SAM-2 in improving segmentation accuracy in X-ray crystallography. The approach not only enhances structural similarity in simulated images but also provides robust and reliable training data for accurate segmentations, crucial for absorption correction.

5.5 Conclusion

In this study, we demonstrated the effectiveness of using synthetic data combined with SAM-2 refinement for improving 3D volume segmentation in X-ray crystallography. Our results show that the AnACorNet_RS_SAM model outperforms other approaches in terms of segmentation accuracy, as evidenced by lower Dice Loss and Cross-Entropy Loss values, and a closer alignment with the ground truth in merging statistics and anomalous peak heights after analytical absorption correction. The integration of synthetic data in training enhances the model's ability to generalize to real-world examples, while the SAM-2 refinement further improves segmentation precision. This combined approach not only maintains high structural similarity in simulated data but also significantly reduces manual labour for segmentation, making it a valuable tool for high-throughput long-wavelength crystallography analysis. Combining with the work in section §4, a fast and comprehensive pipeline for absorption correction in long-wavelength crystallography can be introduced.

Chapter 6

Conclusion and Future Works

In this thesis, we present a novel analytical absorption correction technique for X-ray long-wavelength macromolecular crystallography. By using ray-tracing methods on segmented tomography reconstructions, the incident and the diffracted path lengths of X-ray traversal through the sample can be determined, which are crucial to calculating accurate analytical absorption factors. For practical use in Beamlines, such as Beamline I23 at Diamond Light Source, various advanced methods are explored to ensure accuracy and computational efficiency. The following sections summarize the main findings of each chapter, highlight their contributions to the field, and propose avenues for future work.

6.1 Summary of Major Contributions

6.1.1 Chapter 3: AnACor1.0 - Ray-tracing Analytical Absorption Corrections

In Chapter 3, we presented AnACor1.0 , a ray-tracing analytical absorption correction method designed for long-wavelength macromolecular crystallography. The code of this software is 100% my contribution, with thanks to James Beilsten-Edmands (Diamond Light Source) for the example usage of the DIALS API. Compared to conventional methods, this approach improved the strength of the anomalous signal by 35% through the incorporation of segmented 3D models of the sample, including the crystal, mounting loop, and mother liquor. Also, a novel method of experimental absorption coefficient determination was introduced to ensure accurate absorption coefficients for the materials in the sample. The ray-tracing method proved particularly effective in accounting for crystals in low-symmetry space groups, where standard spherical harmonics correction does not perform well. The

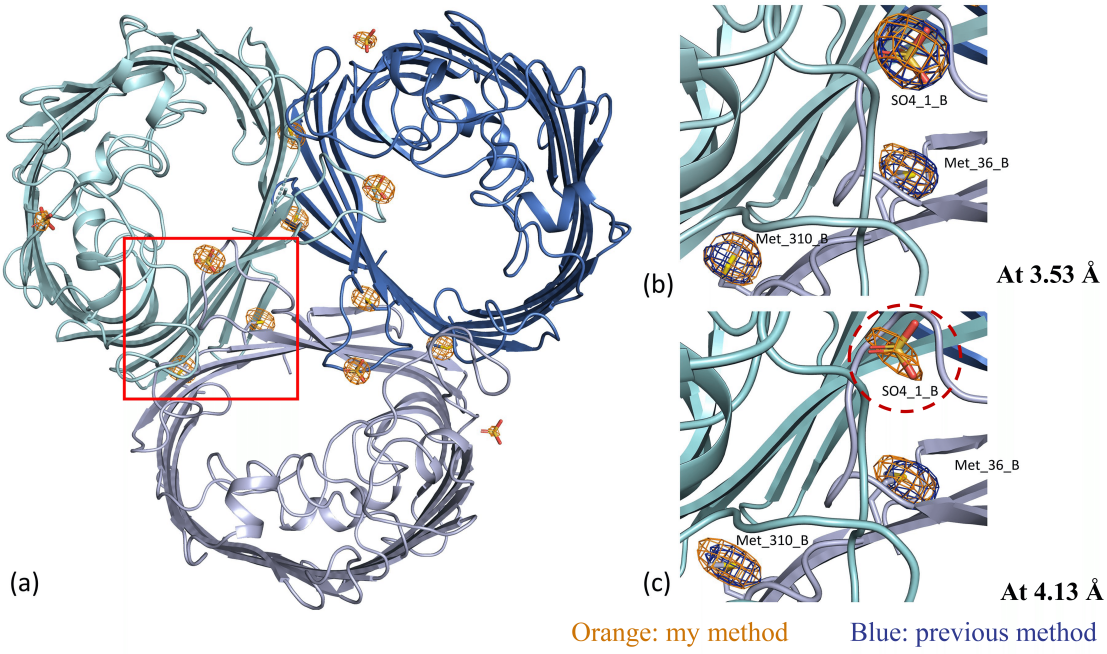


Figure 6.1: Anomalous difference maps of OmpK36 at 3.53 Å (b) and 4.13 Å (c), contoured at the 5σ level. These maps were calculated using anomalous differences (ΔF_{ano}) and phases from molecular replacement. (a) is the reference map with a larger view. Blue meshes indicate anomalous signal corresponding to sulfur atoms detected using the previous method, while orange meshes represent the signal obtained using the method introduced in this thesis. At 3.53 Å the new method produces stronger anomalous signals. At 4.13 Å it successfully preserves the signal from the sulfate group SO4_1_B, which is missing in the previous method.

results showed that using AnACor1.0 led to reduced systematic errors and improved quality of diffraction data, which enhanced the performance of merging statistics and anomalous peak heights.

However, the effectiveness of AnACor1.0 depends on the availability of an accurate, high-resolution 3D model of the sample components (obtained through segmented tomography reconstruction), including the crystal, mounting loop, and surrounding mother liquor. If the reconstruction or segmentation quality is significantly compromised, the accuracy of the absorption correction will be reduced. The method is most valid under experimental conditions where single-wavelength data is used and the anomalous signal is sufficiently strong. While it has demonstrated improvements in long-wavelength single-wavelength anomalous

diffraction (SAD) experiments, its performance under multi-wavelength anomalous diffraction (MAD) or broader bandwidth conditions has not yet been evaluated. Moreover, the accuracy of the correction may degrade in cases with extreme beam divergence, complex sample morphologies not well captured by tomography, or strong scattering artifacts that affect the segmentation.

Future Work: Future developments of AnACor1.0 could explore its applicability beyond macromolecular crystallography, particularly in fields that require precise modelling of X-ray absorption and geometric ray paths. One promising direction is Small Molecule Single Crystal (SMSC) Diffraction, where accurate absorption correction remains essential—especially for crystals containing heavy atoms or irregular morphologies. Traditional methods, such as spherical or empirical multi-scan corrections, may not adequately model complex crystal shapes or mounting artefacts. AnACor1.0's voxel-based ray-tracing approach could offer more accurate corrections by explicitly modelling real 3D crystal geometries and mounting materials, which is particularly valuable in high-precision SMSC studies such as charge-density analysis.

6.1.2 Chapter 4: AnACor2.0 - GPU-Accelerated Analytical Absorption Correction

In Chapter 4, we introduced AnACor2.0, a GPU-accelerated version of the analytical absorption correction method. By leveraging CUDA-based implementations and numerical CPU-based accelerations, this version significantly reduces computational time. The code of this software is 100% my contribution, with thanks to Karel Adamek (University of Oxford) for the example usage of the Reduction CUDA algorithm. This enables several capabilities that were previously impractical: for example, correction of full 3D absorption grids for entire crystals with tens of thousands of voxels across hundreds of datasets can now be completed within minutes, compared to hours or even days using CPU-based methods. This scalability makes the method feasible for routine use in large-scale macromolecular crystallography, where multiple datasets from different goniometer angles are common.

Future Work: Further optimization could involve extending the CUDA implementation to support multi-GPU and distributed computing environments, enabling the processing of even larger datasets. Currently, AnACor2.0 is limited to a single GPU. Additionally, a ROCm-based version (AMD's GPU programming framework) could be developed to ensure compatibility with AMD hardware. Another direction is to incorporate an adaptive error estimation model for the CPU-based version, improving result accuracy when GPU resources are not available.

6.1.3 Chapter 5: Automatic Segmented Tomography Reconstruction

In Chapter 5, we described the development and application of an automatic segmented tomography reconstruction technique for crystallography. This chapter presented the use of machine learning models for automatically segmenting X-ray tomographic images to generate accurate 3D representations of crystal samples. The model was trained using a combination of real and synthetic simulation data, which emulated the conditions and variations observed in real tomography experimental data.

The simulated datasets were designed to replicate experimental imaging conditions, including parameters such as propagation distance, X-ray wavelength, and detector pixel size in I23 Beamline, at Diamond Light Source. Importantly, the simulated dataset included phase contrast by modelling X-ray propagation using Fresnel diffraction. Incorporating these effects into the training data improved the model's robustness and its ability to accurately segment features in phase-influenced experimental reconstructions.

We ensured that the model remained robust across a range of scenarios, including variations in crystal morphology, the presence or absence of mounting loops, differences in crystal attenuation coefficients, and changes in X-ray wavelength and phase contrast levels. When segmentation performance degrades, which is indicated by low Dice scores on validation sets or noticeable visual discrepancies compared to experimental data, retraining or fine-tuning the model may be required. This is especially important when encountering new imaging conditions or sample types not well captured in the original training dataset.

Additionally, we applied SAM-2 for post-refinement of the segmented results. SAM-2 enhances the accuracy of segmentation by refining boundaries and correcting minor discrepancies that may occur during the initial automated process. This automation significantly reduces the human effort required for manual segmentation. Also, the use of synthetic data enables rapid iterations and refinement of the segmentation model without relying solely on time-consuming and labour-intensive manual annotations.

Future Work: Although SAM-2 has shown effectiveness in refining segmentation, its processing speed remains a limitation. Future research could explore traditional machine learning algorithms, such as Random Forest, which may offer faster processing times for segmentation tasks. Additionally, the current model employs a large number of parameters, leading to increased computational complexity. Future efforts could focus on developing lighter models that maintain segmentation accuracy while improving efficiency and reducing computational load. By optimizing the model architecture and exploring simpler algorithms, it may be possible to achieve a more efficient and scalable solution for automatic tomography segmentation.

6.2 General Future Directions

This thesis has established the foundations for efficient and accurate absorption correction and segmentation in long-wavelength macromolecular crystallography. Looking forward, several broader directions can be pursued:

1. **Integration with Automated Data Pipelines**: Future work could integrate the absorption correction and segmentation methods presented in this thesis into a fully automated and optimized crystallography data analysis pipeline. This would facilitate real-time corrections and improve the efficiency of structural biology studies.

2. **Validation on Diverse Samples and Experimental Conditions**:

   Future work should include testing the proposed methods across a broader range of crystal types. So far, the techniques have primarily been evaluated on protein

crystals with relatively simple morphologies and limited space group diversity. Expanding to crystals of other substances with different symmetries, shapes, or internal complexities will help further validate their robustness and adaptability.

This includes validating the techniques on radiation-sensitive crystals, which are prone to have local heating, and possible structural degradation under long-wavelength X-ray exposure. In such cases, we expect to observe subtle changes in the shape of the crystal, including microcracking or shrinkage due to solvent loss, which may in turn affect segmentation accuracy and absorption path estimation. In this study, heterogeneous materials such as the crystal, loop, and surrounding mother liquor were explicitly modelled and segmented.

However, future validation should also consider more complex forms of heterogeneity, such as partially disordered regions within the crystal (where the density becomes diffuse, blurred, or irregular). Also, inclusions, such as salts or precipitated agents, are small, dense and non-uniform regions within the crystal and may be introduced during crystallization or freezing. These features can create unpredictable attenuation profiles and may reduce the accuracy of both segmentation and absorption correction.

3. **Interdisciplinary Applications**: The methods presented in this thesis can also be adapted for other fields, such as material sciences and medical imaging, where long-wavelength tomography is becoming increasingly relevant or the absorption effect is significant. Exploring these interdisciplinary applications could open up new possibilities for the developed techniques.

Bibliography

[1] H. E. K. Mortimer Abramowitz, Kenneth R. Spring and M. W. Davidson, "Basic principles of microscope objectives," *BioTechniques*, vol. 33, no. 4, pp. 772–781, 2002.

[2] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Ábrányi Balogh, J. Brandão-Neto, A. Carbery, G. Davison, A. Dias, T. D. Downes, L. Dunnett, M. Fairhead, J. D. Firth, S. P. Jones, A. Keeley, and M. A. Walsh, "Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease," *Nat Commun*, vol. 11, no. 1, p. 5047, 2020.

[3] W. H. Bragg and W. L. Bragg, "The reflection of x-rays by crystals," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 88, no. 605, pp. 428–438, 1913.

[4] I. J. Pickering, R. C. Prince, T. Divers, and G. N. George, "Sulfur k-edge x-ray absorption spectroscopy for determining the chemical speciation of sulfur in biological systems," *FEBS Letters*, vol. 441, no. 1, pp. 11–14, 1998.

[5] K. El Omari, R. Duman, V. Mykhaylyk, C. M. Orr, M. Latimer-Smith, G. Winter, V. Grama, F. Qu, K. Bountra, H. S. Kwong, M. Romano, R. I. Reis, L. Vogeley, L. Vecchia, C. D. Owen, S. Wittmann, M. Renner, M. Senda, N. Matsugaki, Y. Kawano, T. A. Bowden, I. Moraes, J. M. Grimes, E. J. Mancini, M. A. Walsh, C. R. Guzzo, R. J. Owens, E. Y. Jones, D. G. Brown, D. I. Stuart, K. Beis, and A. Wagner, "Experimental phasing opportunities for macromolecular crystallography at very long wavelengths," *Commun. Chem.*, vol. 6, p. 219, Oct. 2023.

[6] A. Wagner, R. Duman, K. Henderson, and V. Mykhaylyk, "In-vacuum long-wavelength macromolecular crystallography," *Acta Cryst.*, vol. D72, pp. 430–439, Mar. 2016.

[7] W. A. Hendrickson and M. M. Teeter, "Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur," *Nature*, vol. 290, no. 5802, pp. 107–113, 1981.

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[9] G. Albrecht, "The absorption factor in crystal spectroscopy," *Rev. Sci. Instrum.*, vol. 10, pp. 221–222, 1939.

[10] E. N. Maslen, "Absorption corrections," in *International Tables for Crystallography* (E. Prince, ed.), vol. C, ch. 6.3.3, pp. 600–608, Wiley, 3rd ed., 2006.

[11] O. Aurelius, R. Duman, K. El Omari, V. Mykhaylyk, and A. Wagner, "Long-wavelength macromolecular crystallography - first successful native sad experiment close to the sulfur edge," *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 411, pp. 12–16, November 15 2017.

[12] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, "Crystallography Open Database – an open-access collection of crystal structures," *Journal of Applied Crystallography*, vol. 42, pp. 726–729, Aug 2009.

[13] "X-ray diffraction," 2020. Retrieved November 27, 2020, from `https://phys.libretexts.org/@go/page/4511`.

[14] W. Kabsch, "XDS," *Acta Cryst.*, vol. D66, pp. 125–132, Feb 2010.

[15] T. G. G. Battye, L. Kontogiannis, O. Johnson, H. R. Powell, and A. G. W. Leslie, "iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM," *Acta Cryst.*, vol. D67, pp. 271–281, Apr 2011.

[16] G. Winter, D. G. Waterman, J. M. Parkhurst, A. S. Brewster, R. J. Gildea, M. Gerstel, L. Fuentes-Montero, M. Vollmar, T. Michels-Clark, I. D. Young, N. K. Sauter, and G. Evans, "DIALS: implementation and evaluation of a new integration package," *Acta Cryst.*, vol. D74, pp. 85–97, Feb 2018.

[17] W. contributors, "Miller index," 2021. Retrieved January 20, 2021, from `https://en.wikipedia.org/w/index.php?title=Miller_index&oldid=1001590663`.

[18] M. M. Ripoll, "Crystallography. direct and reciprocal lattices." Retrieved July 31, 2021, from `https://www.xtal.iqfr.csic.es/Cristalografia/parte_04-en.html`.

[19] V. K. Pecharsky and P. Y. Zavalij, *Properties, sources, and detection of radiation*. 2009.

[20] E. F. Garman, "Radiation damage in macromolecular crystallography: what is it and why should we care?," *Acta Cryst.*, vol. S66, pp. 339–351, Apr 2010.

[21] P. R. Evans and G. N. Murshudov, "How good are my data and what is the resolution?," *Acta Cryst.*, vol. D69, pp. 1204–1214, Jul 2013.

[22] J. Beilsten-Edmands, G. Winter, R. Gildea, J. Parkhurst, D. Waterman, and G. Evans, "Scaling diffraction data in the DIALS software package: algorithms and new approaches for multi-crystal scaling," *Acta Cryst.*, vol. D76, pp. 385–399, Apr 2020.

[23] K. D. Cowtan. `http://www.ysbl.york.ac.uk/~cowtan/fourier/fourier.html`. Retrieved July 18, 2024.

[24] "The nobel prize in chemistry 1985. nobelprize.org. nobel prize outreach ab 2021." Retrieved July 30, 2021, from `https://www.nobelprize.org/prizes/chemistry/1985/summary/`.

[25] A. L. Patterson, "A direct method for the determination of the components of interatomic distances in crystals," *Zeitschrift für Kristallographie-Crystalline Materials*, vol. 90, no. 1-6, pp. 517–542, 1935.

[26] G. Taylor, "The phase problem," *Acta Cryst.: Biological Crystallography*, vol. D59, no. 11, pp. 1881–1890, 2003.

[27] M. G. Rossmann and D. M. Blow, "The detection of sub-units within the crystallographic asymmetric unit," *Acta Crystallographica*, vol. 15, no. 1, pp. 24–31, 1962.

[28] W. Hendrickson and M. Teeter, "Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur," *Nature*, vol. 290, no. 5802, pp. 107–113, 1981.

[29] E. Dodson, "Is it jolly SAD?," *Acta Cryst.*, vol. D59, pp. 1958–1965, Nov 2003.

[30] M. Schiltz and G. Bricogne, "Exploiting the anisotropy of anomalous scattering boosts the phasing power of SAD and MAD experiments," *Acta Cryst.*, vol. D64, pp. 711–729, Jul 2008.

[31] C. Giacovazzo and D. Siliqi, "Phasing via SAD/MAD data: the method of the joint probability distribution functions," *Acta Cryst.*, vol. D60, pp. 73–82, Jan 2004.

[32] J. P. Rose, B.-C. Wang, and M. S. Weiss, "Native SAD is maturing," *IUCrJ*, vol. 2, pp. 431–440, Jul 2015.

[33] L. M. F. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, "Before and after alphafold2: An overview of protein structure prediction," *Frontiers in bioinformatics*, vol. 3, p. 1120370, 2023.

[34] M. S. Weiss and R. Hilgenfeld, "On the use of the merging R factor as a quality indicator for X-ray data," *J. Appl. Cryst.*, vol. 30, pp. 203–205, Apr 1997.

[35] K. Diederichs and P. A. Karplus, "Improved r-factors for diffraction data analysis in macromolecular crystallography," *Nat. Struct. Biol.*, vol. 4, pp. 269–275, 1997.

[36] M. S. Weiss, "Global indicators of X-ray data quality," *J. Appl. Cryst.*, vol. 34, pp. 130–135, Apr 2001.

[37] A. T. Brünger, "Free r value: a novel statistical quantity for assessing the accuracy of crystal structures," *Nature*, vol. 355, no. 6359, pp. 472–475, 1992.

[38] International Union of Crystallography, "R factor," 2006. Accessed: 2025-04-27.

[39] R. G. Howells, "A graphical method of estimating absorption factors for single crystals," *Acta Cryst.*, vol. 3, pp. 366–369, Sep 1950.

[40] J. de Meulenaer and H. Tompa, "The absorption correction in crystal structure analysis," *Acta Cryst.*, vol. 19, pp. 1014–1018, Dec 1965.

[41] N. W. Alcock, G. S. Pawley, C. P. Rourke, and M. R. Levine, "An improvement in the algorithm for absorption correction by the analytical method," *Acta Cryst.*, vol. A28, pp. 440–444, Sep 1972.

[42] R. C. Clark, "The absorption-correction factor of multifaceted crystals," *Acta Cryst.*, vol. A49, pp. 692–697, Sep 1993.

[43] R. C. Clark and J. S. Reid, "The analytical calculation of absorption in multifaceted crystals," *Acta Cryst.*, vol. A51, pp. 887–897, Nov 1995.

[44] G. T. DeTitta, "ABSORB: An absorption correction program for crystals enclosed in capillaries with trapped mother liquor," *J. Appl. Cryst.*, vol. 18, pp. 75–79, Apr 1985.

[45] A. C. T. North, D. C. Phillips, and F. S. Mathews, "A semi-empirical method of absorption correction," *Acta Cryst.*, vol. A24, pp. 351–359, May 1968.

[46] T. Furnas, *Single Crystal Orienter Instruction Manual*. Milwaukee, 1957.

[47] C. Katayama, N. Sakabe, and K. Sakabe, "A statistical evaluation of absorption," *Acta Cryst.*, vol. A28, pp. 293–295, May 1972.

[48] C. Katayama, "An analytical function for absorption correction," *Acta Cryst.*, vol. A42, pp. 19–23, Jan 1986.

[49] R. H. Blessing, "An empirical correction for absorption anisotropy," *Acta Cryst.*, vol. A51, pp. 33–38, Jan 1995.

[50] W. Minor, M. Cymborowski, Z. Otwinowski, and M. Chruszcz, "HKL-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes," *Acta Cryst.*, vol. D62, pp. 859–866, Aug 2006.

[51] G. M. Sheldrick *University of Gottingen, Germany*, 1996.

[52] R. M. F. Leal, S. C. M. Teixeira, V. Rey, V. T. Forsyth, and E. P. Mitchell, "Absorption correction based on a three-dimensional model reconstruction from visual images," *J. Appl. Cryst.*, vol. 41, pp. 729–737, Aug 2008.

[53] T. Strutz, "3d shape reconstruction of loop objects in x-ray protein crystallography," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, pp. 797–807, 2011.

[54] D. R. Merrifield, V. Ramachandran, K. J. Roberts, W. Armour, D. Axford, M. Basham, T. Connolley, G. Evans, K. E. McAuley, R. L. Owen, *et al.*, "A novel technique combining high-resolution synchrotron x-ray microtomography and x-ray diffraction for characterization of micro particulates," *Meas. Sci. Technol.*, vol. 22, p. 115703, 2011.

[55] A. J. Warren, W. Armour, D. Axford, M. Basham, T. Connolley, D. R. Hall, S. Horrell, K. E. McAuley, V. Mykhaylyk, A. Wagner, *et al.*, "Visualization of membrane protein crystals in lipid cubic phase using x-ray imaging," *Acta Cryst.*, vol. D69, pp. 1252–1259, 2013.

[56] S. Brockhauser, M. Di Michiel, J. E. McGeehan, A. A. McCarthy, and R. B. G. Ravelli, "X-ray tomographic reconstruction of macromolecular samples," *J. Appl. Cryst.*, vol. A41, pp. 1057–1066, Dec 2008.

[57] Wikipedia contributors, "Projection-slice theorem," 2024. Accessed: 2024-07-29.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[59] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[60] S. Ali, S. Mayo, A. K. Gostar, R. Tennakoon, A. Bab-Hadiashar, T. MCann, H. Tuhumury, and J. Favaro, "Automatic segmentation for synchrotron-based imaging of porous bread dough using deep learning approach," *J. Synchrotron Radiat.*, vol. 28, pp. 566–575, Mar 2021.

[61] "High performance savu software for fast 3d model-based iterative reconstruction of large data at diamond light source," *SoftwareX*, vol. 19, p. 101157, 2022.

[62] D. Kazantsev, R. Duman, A. Wagner, V. Mykhaylyk, K. Wanelik, M. Basham, and N. Wadeson, "X-ray tomographic reconstruction and segmentation pipeline for the long-wavelength macromolecular crystallography beamline at Diamond Light Source," *J. Synchrotron Radiat.*, vol. 28, pp. 889–901, May 2021.

[63] N. T. Vo, R. C. Atwood, and M. Drakopoulos, "Superior techniques for eliminating ring artifacts in x-ray micro-tomography," *Opt. Express*, vol. 26, pp. 28396–28412, Oct 2018.

[64] P. Benedetti, M. Femminella, and G. Reali, "Mixed-sized biomedical image segmentation based on u-net architectures," *Applied Sciences*, vol. 13, no. 1, 2023.

[65] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, Feb. 2017.

[66] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), (Cham), pp. 149–157, Springer International Publishing, 2016.

[67] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1822–1834, Aug 2018.

[68] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[69] S. Liu, D. Xu, S. K. Zhou, O. Pauly, S. Grbic, T. Mertelmeier, J. Wicklein, A. Jerebko, W. Cai, and D. Comaniciu, "3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), (Cham), pp. 851–858, Springer International Publishing, 2018.

[70] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[71] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, "Hierarchical 3d fully convolutional networks for multi-organ segmentation," *arXiv preprint arXiv:1704.06382*, 2017.

[72] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[74] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[75] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[76] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, January 2022.

[77] H. L. Monaco and G. Artioli, "Data reduction," in *Fundamentals of Crystallography* (H. Giacovazzo, ed.), ch. 5, pp. 376–388, Oxford: Oxford University Press, 2nd ed., 2002.

[78] U. W. Arndt, "Optimum X-ray wavelength for protein crystallography," *J. Appl. Cryst.*, vol. 17, pp. 118–119, Apr 1984.

[79] W. R. Busing and H. A. Levy, "High-speed computation of the absorption correction for single-crystal diffraction measurements," *Acta Cryst.*, vol. 10, pp. 180–182, Mar 1957.

[80] G. T. DeTitta, "ABSORB: An absorption correction program for crystals enclosed in capillaries with trapped mother liquor," *J. Appl. Cryst.*, vol. 18, pp. 75–79, Apr 1985.

[81] R. C. Clark and J. S. Reid, "The analytical calculation of absorption in multifaceted crystals," *Acta Cryst.*, vol. A51, pp. 887–897, Nov 1995.

[82] A. X. S. Bruker, "APEX3 package, APEX3, SAINT and SADABS," 2016.

[83] A. C. T. North, D. C. Phillips, and F. S. Mathews, "A semi-empirical method of absorption correction," *Acta Cryst.*, vol. A24, pp. 351–359, May 1968.

[84] G. Kopfmann and R. Huber, "A method of absorption correction by X-ray intensity measurements," *Acta Cryst.*, vol. A24, pp. 348–351, May 1968.

[85] C. Katayama, N. Sakabe, and K. Sakabe, "A statistical evaluation of absorption," *Acta Cryst.*, vol. 28, pp. 293–295, May 1972.

[86] N. Walker and D. Stuart, "An empirical method for correcting diffractometer data for absorption effects," *Acta Cryst.*, vol. A39, pp. 158–166, Jan 1983.

[87] J. L. Wong, M. Romano, L. E. Kerry, H.-S. Kwong, W.-W. Low, S. J. Brett, A. Clements, K. Beis, and G. Frankel, "Ompk36-mediated carbapenem resistance attenuates st258 klebsiella pneumoniae in vivo," *Nat. Commun.*, vol. 10, pp. 1–10, 2019.

[88] I. Schaffner, G. Mlynek, N. Flego, D. Pühringer, J. Libiseller-Egger, L. Coates, S. Hofbauer, M. Bellei, P. G. Furtmüller, G. Battistuzzi, G. Smulevich, K. Djinović-Carugo, and C. Obinger, "Molecular mechanism of enzymatic chlorite detoxification: Insights from structural and kinetic studies," *ACS Catalysis*, vol. 7, pp. 7962–7976, 2017.

[89] R. Duman, C. M. Orr, V. Mykhaylyk, K. El Omari, R. Pocock, V. Grama, and A. Wagner, "Sample preparation and transfer protocol for in-vacuum long-wavelength crystallography on beamline i23 at diamond light source," *J. Vis. Exp.*, vol. 170, p. e62364, 2021.

[90] O. B. Zeldin, M. Gerstel, and E. F. Garman, "RADDOSE-3D: time- and space-resolved modelling of dose in macromolecular crystallography," *J. Appl. Cryst.*, vol. 46, pp. 1225–1230, Aug 2013.

[91] D. Kazantsev, N. Wadeson, and M. Basham, "High performance savu software for fast 3D model-based iterative reconstruction of large data at diamond light source," *SoftwareX*, vol. 19, p. 101157, July 2022.

[92] D. Gürsoy, F. De Carlo, X. Xiao, and C. Jacobsen, "TomoPy: a framework for the analysis of synchrotron tomographic data," *Journal of Synchrotron Radiation*, vol. 21, pp. 1188–1193, Sep 2014.

[93] A. Thorn and G. M. Sheldrick, "ANODE: anomalous and heavy-atom density calculation," *J. Appl. Cryst.*, vol. 44, pp. 1285–1287, Dec 2011.

[94] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, "Refinement of Macromolecular Structures by the Maximum-Likelihood Method," *Acta Cryst.*, vol. D53, pp. 240–255, May 1997.

[95] P. Skubák and N. S. Pannu, "Automatic protein structure solution from weak x-ray data," *Nature Communications*, vol. 4, no. 1, p. 2777, 2013.

[96] R. J. Angel, "Absorption corrections for diamond-anvil pressure cells implemented in the software package Absorb6.0," *J. Appl. Cryst.*, vol. 37, pp. 486–492, Jun 2004.

[97] J. Amanatides and A. Woo, "A Fast Voxel Traversal Algorithm for Ray Tracing," in *EG 1987-Technical Papers*, Eurographics Association, 1987.

[98] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency," *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 741–753, 1977.

[99] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 55, pp. 532–537, 1993.

[100] P.-S. Liao, T.-S. Chen, and P.-C. Chung, "A fast algorithm for multilevel thresholding," *Journal of Information Science and Engineering*, vol. 17, no. 5, pp. 713–727, 2001. Available at: `https://ftp.iis.sinica.edu.tw/JISE/2001/200109_01.pdf`.

[101] C. H. Li and C. K. Lee, "Minimum cross entropy thresholding," *Pattern Recognition*, vol. 26, no. 4, pp. 617–625, 1993.

[102] J. C. Yen, F. J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, 1995.

[103] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, pp. 630–632, 1978.

[104] R. C. Gonzalez and R. E. Wood, *Digital Image Processing (2nd Edition)*. Prentice-Hall Inc., 2002.

[105] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sept. 2020.

[106] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015.

[107] G. Winter, J. Beilsten-Edmands, N. Devenish, M. Gerstel, R. J. Gildea, D. McDonagh, E. Pascal, D. G. Waterman, B. H. Williams, and G. Evans, "Dials as a toolkit," *Protein Science*, vol. 31, no. 1, pp. 232–250, 2022.

[108] P. A. Karplus and K. Diederichs, "Linking crystallographic model and data quality," *Science*, vol. 336, pp. 1030–1033, 2012.

[109] C. Yang, J. W. Pflugrath, D. A. Courville, C. N. Stence, and J. D. Ferrara, "Away from the edge: SAD phasing from the sulfur anomalous signal measured in-house with chromium radiation," *Acta Cryst.*, vol. D59, pp. 1943–1957, Nov 2003.

[110] P. Evans, "Scaling and assessment of data quality," *Acta Cryst.*, vol. D62, pp. 72–82, Jan 2006.

[111] A. Richards, "University of oxford advanced research computing," 2015.

[112] J. Beilsten-Edmands, G. Winter, R. Gildea, J. Parkhurst, D. Waterman, and G. Evans, "Scaling diffraction data in the DIALS software package: algorithms and new approaches for multi-crystal scaling," *Acta Cryst.*, vol. D76, pp. 385–399, Apr 2020.

[113] A. Burkhardt, M. Warmer, S. Panneerselvam, A. Wagner, A. Zouni, C. Glöckner, R. Reimer, H. Hohenberg, and A. Meents, "Fast high-pressure freezing of protein crystals in their mother liquor," *Acta Cryst.*, vol. F68, no. 4, pp. 495–500, 2012.

[114] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[115] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, "Gnu scientific library," *No. Release*, vol. 2, 1996.

[116] N. Corporation, *CUDA C Programming Guide*, 2024. Accessed: 2024-10-07.

[117] A. Alvarenga de Moura Meneses, A. Giusti, A. P. de Almeida, L. Parreira Nogueira, D. Braz, R. Cely Barroso, and C. E. deAlmeida, "Automated segmentation of synchrotron radiation micro-computed tomography biomedical images using graph cuts and neural networks," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 660, no. 1, pp. 121–129, 2011.

[118] D. Kazantsev, R. Duman, A. Wagner, V. Mykhaylyk, K. Wanelik, M. Basham, and N. Wadeson, "X-ray tomographic reconstruction and segmentation pipeline for the long-wavelength macromolecular crystallography beamline at diamond light source," *Journal of Synchrotron Radiation*, vol. 28, no. 3, 2021.

[119] A. Kornilov, I. Safonov, and I. Yakimchuk, "A review of watershed implementations for segmentation of volumetric images," *Journal of Imaging*, vol. 8, no. 5, 2022.

[120] T. Strohmann, K. Bugelnig, E. Breitbarth, *et al.*, "Semantic segmentation of synchrotron tomography of multiphase al-si alloys using a convolutional neural network with a pixel-wise weighted loss function," *Sci Rep*, vol. 9, p. 19611, 2019.

[121] S. Ali, S. Mayo, A. K. Gostar, R. Tennakoon, A. Bab-Hadiashar, T. MCann, H. Tuhumury, and J. Favaro, "Automatic segmentation for synchrotron-based imaging of porous bread dough using deep learning approach," *Journal of Synchrotron Radiation*, vol. 28, pp. 566–575, Mar 2021.

[122] A. Tsamos, S. Evsevleev, R. Fioresi, F. Faglioni, and G. Bruno, "A novel iterative algorithm to improve segmentations with deep convolutional neural networks trained with synthetic x-ray computed tomography data," *Computational Materials Science*, vol. 223, p. 112112, 2023.

[123] S. Gaudez, M. Ben Haj Slama, A. Kaestner, and M. V. Upadhyay, "3D deep convolutional neural network segmentation model for precipitate and porosity identification in synchrotron X-ray tomograms," *Journal of Synchrotron Radiation*, vol. 29, pp. 1232–1240, Sep 2022.

[124] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[125] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 87–110, jan 2023.

[126] R. Azad, L. Niggemeier, M. Hüttemann, A. Kazerouni, E. K. Aghdam, Y. Velichko, U. Bagci, and D. Merhof, "Beyond self-attention: Deformable large kernel attention for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1287–1297, 2024.

[127] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6202–6212, 2023.

[128] T. Faragó, P. Mikulík, A. Ershov, M. Vogelgesang, D. Hänschke, and T. Baumbach, "Syris: a flexible and efficient framework for x-ray imaging experiments simulation," *Journal of Synchrotron Radiation*, vol. 24, no. 6, pp. 1283–1295, 2017.

[129] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[130] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company publishers, 2005.

[131] M. Dera, "Crystals generator," 2024. GitHub repository.

[132] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[133] B. Henke, E. Gullikson, and J. Davis, "X-ray interactions: photoabsorption, scattering, transmission, and reflection at e=50-30000 ev, z=1-92," *Atomic Data and Nuclear Data Tables*, vol. 54, pp. 181–342, July 1993.

[134] B. Henke, E. Gullikson, and J. Davis, "X-ray interactions: Photoabsorption, scattering, transmission, and reflection at e = 50-30,000 ev, z = 1-92," *Atomic Data and Nuclear Data Tables*, vol. 54, no. 2, pp. 181–342, 1993.

[135] H. Fischer, I. Polikarpov, and A. F. Craievich, "Average protein density is a molecular-weight-dependent function," *Protein Science*, vol. 13, no. 10, pp. 2825–2828, 2004.

[136] R. Okuta, Y. Unno, D. Nishino, S. Hido, and C. Loomis, "Cupy: A numpy-compatible library for nvidia gpu calculations," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[137] "Integration of tomopy and the astra toolbox for advanced processing and reconstruction of tomographic synchrotron data," *Journal of synchrotron radiation*, vol. 23, no. 3, pp. 842–849, 2016.

[138] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[139] "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, pp. 733–752, 2023.

[140] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.

[141] P. Wang, W. Zheng, T. Chen, and Z. Wang, "Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice," *arXiv preprint arXiv:2203.05962*, 2022.

[142] B. Xu, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[143] S. Ioffe, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[144] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[145] D. J. Pearce, "An improved algorithm for finding the strongly connected components of a directed graph," *Victoria University, Wellington, NZ, Tech. Rep*, 2005.