Indexing algorithms for current and upcoming challenges in serial crystallography

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

von

Yaroslav Gevorkov

aus

Kiev, Ukraine

2020

Gutachter:
1) Prof. Dr.-Ing. Rolf-Rainer Grigat
2) Prof. Dr. Dr. h. c. Henry N. Chapman

Tag der mündlichen Prüfung: 22.07.2020

DOI: https://doi.org/10.15480/882.2853
ORCID: https://orcid.org/0000-0001-9273-7648

Acknowledgments

I wish to express my gratitude to Prof. Dr.-Ing. Rolf-Rainer Grigat for guiding and supporting me throughout my doctoral studies. Without your exploratory nature and remarkable mediation skills I wouldn't be where I am today. I want to thank you for teaching me to focus on the most scientifically valuable topics in my studies.

I would like to thank Prof. Dr. Dr. h. c. Henry N. Chapman for giving me the opportunity to conduct my doctoral studies in his group. I am very grateful for you giving me the freedom to choose my scientific topics and the way how to work on them. Your in-depth knowledge in a wide range of scientific fields was very valuable for my studies.

Special thanks go to Dr. Oleksandr Yefanov for an endless number of fruitful discussions and for sharing his in-depth knowledge in various fields of physics with me. I am grateful that you saw the potential in the methods I developed long before results could be obtained and encouraged me to conduct full-time research in this direction.

Many thanks go to Dr. Thomas A. White for his very valuable help in crystallography related topics, for his patience and skills in reviewing publications, and for the chance to participate in the development of the CrystFEL software suite. I appreciate your scientific advice and the thorough review of this thesis.

I want to thank Aleksandra Tolstikova for very interesting and fruitful discussions on data processing related topics and for being an awesome partner at many experiments.

It has been a true pleasure to work with Wolfgang Brehm. I would like to thank you for the great number of rewarding discussions and for helping me understand many physics-related relationships.

Appreciation is due to Dr. Valerio Mariani for showing me the importance of user-friendliness of software that is intended to be used by a large community.

I want to show my deep gratitude to Dr. Anton Barty for very valuable and always brought to the point advice of technical as well as non-technical nature.

Very special thanks go to my family - in particular to my parents Robert Gevorkov and Jana Gevorkov, my wife Elena Gevorkov, and my daughter Alyssa Gevorkov for their continuous and unconditional support. Without your patience and support none of this would have been possible.

Contents

List of Figures **v**

List of Tables **ix**

List of Symbols **xi**

1 Motivation **2**

2 Introduction **5**
2.1 Basic physical concepts 5
2.1.1 Scattering 5
2.1.2 Limitations 7
2.2 Conventional crystallography 9
2.2.1 Data processing pipeline 12
2.2.2 Practical limitations of conventional crystallography . . . 17
2.3 Serial crystallography 18
2.3.1 Sample delivery 18
2.3.2 Partiality estimation 19
2.3.3 Indexing still diffraction patterns 20

3 XGANDALF - Extended Gradient Descent Algorithm for Lattice Finding **22**
3.1 Introduction 22
3.2 Algorithm description 26
3.2.1 Overview 26
3.2.2 Relation between nodes and the indexing solution 27
3.2.3 Continuous proximity function for noise tolerance 28
3.2.4 Gradient descent extension 34
3.2.5 Heuristic algorithm for locating maxima in the score function 34
3.2.6 Selection of lattice bases 37
3.2.7 Refinement 39

3.3 Evaluation of the algorithm 39
3.3.1 Indexing rate 39
3.3.2 Execution time 43
3.3.3 Final merged data quality 43
3.4 Code availability 45
3.5 Conclusion 45

4 Pink beam serial crystallography 47
4.1 Synchrotron light sources 47
4.2 Thick Ewald sphere 50
4.3 Pink beam crystallography 53
4.3.1 Overlapping Brag spots 54
4.3.2 Streaky Bragg spots 56
4.3.3 Background noise 58
4.3.4 Partiality estimation 59
4.3.5 Intensity scaling 61
4.3.6 Time-resolved crystallography 61
4.3.7 Conclusion 62

5 PinkIndexer - A universal indexer for pink-beam X-ray and electron diffraction snapshots 63
5.1 Introduction 63
5.2 The pinkIndexer Algorithm 65
5.2.1 Diffraction geometry 65
5.2.2 Determining the crystal orientation 68
5.2.3 Algorithm details 71
5.2.4 Implementation details 74
5.3 Evaluation of algorithm performance 75
5.3.1 Monochromatic X-ray beam Crystallography 75
5.3.2 Pink-Beam Serial Crystallography 78
5.3.3 Serial Electron Crystallography 78
5.4 Conclusion 79
5.5 Code availability 80

6 Indexing in convergent beam serial crystallography 82
6.1 Introduction 82
6.2 Convergent beam diffraction geometry 83
6.3 Indexing convergent beam diffraction patterns 86
6.4 Conclusion 92

7 Conclusion **94**
7.1 Contribution of this thesis to the current research 94
7.2 Future perspectives 97
7.3 Closing words 99

List of Figures

2.1 Visualization of a scattering event in the employed scattering model. 6
2.2 Relation between the real space and the reciprocal space. 8
2.3 One-dimensional illustration of the mathematical model of a crystal and its Fourier transform. 11
2.4 Relation between the reciprocal space and the real space for a crystal. 12
2.5 Visualization of peak detection for indexing and prediction after indexing. 14
2.6 One-dimensional illustration of the derivation of the Fourier transform of a partial lattice, as used by the Fourier methods. 16
2.7 A typical X-ray serial crystallography setup. 19

3.1 A lattice and its unit-cell 24
3.2 Overall structure of the XGANDALF algorithm. 25
3.3 Relation between nodes and the line series' they form. 29
3.4 Sample score function for two nodes. 30
3.5 Sample score function for 13 nodes. 31
3.6 Implemented proximity functions. 32
3.7 Score functions with different proximity functions. 36
3.8 Different bases formed from lattice points of the same lattice. The reduced basis is the one with the shortest possible basis vectors. 38
3.9 Comparison of the success rates of algorithms in indexing patterns as a function of the numbers of Bragg spots N in those patterns. XGANDALF outperforms the other indexers over the whole range of Bragg spot counts. 41
3.10 Comparison of the achieved figure of merit R_{split} (White et al., 2013) (lower is better) for XGANDALF and current state-of-the-art indexers. XGANDALF outperforms the other indexers in all significant resolution shells. 44

3.11 Comparison of the achieved figure of merit $CC*$ (Karplus et al., 2012) (higher is better) for XGANDALF and current state of the art indexers. XGANDALF outperforms the other indexers in all significant resolution shells. 45
3.12 Comparison of the estimated profile radii of MOSFLM and XGANDALF. The estimated radii for patterns indexed by XGANDALF are generally smaller than the ones of MOSFLM, which means that the indexing solution is more precise. 46

4.1 Illustration of a wiggler or undulator. 49
4.2 Spectra of a bending magnet, a wiggler, and an undulator. . . . 50
4.3 Sample spectrum of an undulator where all but one spike have been filtered out. 51
4.4 Illustration of the diffracted intensity calculation for monochromatic and pink beam diffraction patterns. 52
4.5 Simulated diffraction patterns showing the same crystal with (a) monochromatic beam and (a) pink beam. 53
4.6 Directions of Bragg spots generated by (a) a monochromatic beam and (b) a pink beam. 55
4.7 Identification of the wavelength that excited an RLP. 56
4.8 Systematically overlapping Bragg spots in a pink beam crystallography experiment. 57
4.9 Mosaic crystal and its Fourier transform. 58
4.10 Stretched Bragg spot from a mosaic crystal. 59
4.11 Simulated pink beam diffraction pattern of a mosaic crystal. . . 60

5.1 A two-dimensional representation of reciprocal space, showing RLPs of a crystal and the Ewald sphere for monochromatic diffraction. 66
5.2 A two-dimensional representation of reciprocal space depicting Ewald spheres for polychromatic diffraction. 67
5.3 Method to compute all possible rotations of a candidate RLP on a ULS. 69
5.4 Schematic sketch of a rotogram in a 2D vector space. 70
5.5 Plot of the curves $\boldsymbol{v} = \arctan(\theta/4)\,\hat{\boldsymbol{e}}$ for a particular Bragg point direction $\hat{\boldsymbol{q}}$ and a set of 56 candidate RLPs in a cubic lattice. $\phi \in [-2\pi, 0)$ 73
5.6 Comparison of pinkIndexer to other indexing algorithms. 77
5.7 A snapshot pink beam diffraction pattern of two Proteinase K crystals indexed by the pinkIndexer algorithm. 79

5.8 Electron diffraction rotation series data of Proteinase K indexed by the pinkIndexer algorithm. 81

6.1 Two views of Ewald spheres' *centers* of a rotationally symmetric monochromatic convergent beam. 83
6.2 Center slice through the rotationally symmetric thick Ewald sphere of a convergent monochromatic beam. 84
6.3 Calculation of the shape of a Bragg spot generated by an RLP in the thick Ewald sphere of a convergent beam. 85
6.4 Extension of the circular arc of Ewald sphere centers, that excite the RLP, to a whole circle. 87
6.5 Real space representation of the rays in figure 6.4. 88
6.6 Form of the Bragg curve. 89
6.7 Derivation of the uncertainty circular arc (UCA) in the reciprocal space of a convergent beam experiment. 91
6.8 Simulated diffraction pattern with 1.15° convergence radius. . . 92

List of Tables

3.1 Numbers of crystals of CXIDB ID 83 indexed without prior unit cell knowledge. "Indexed with correct unit cell" means here that the detected unit cell is close to the real unit cell and not a supercell or completely different. 42
3.2 Numbers of crystals of CXIDB ID 83 indexed with prior unit cell knowledge. 42
3.3 Comparison of mean execution times (per pattern) and indexing results for a dataset consisting of 1000 patterns. 43

5.1 Indexing results and execution times for monochromatic indexers and pinkIndexer. 76

List of Symbols

$\boldsymbol{\Lambda}$	Real lattice
$\boldsymbol{\Lambda}^*$	Reciprocal lattice
$\widetilde{\boldsymbol{b}}$	Real space basis vector
$\widetilde{\boldsymbol{m}}$	Miller indices
$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$	Real lattice basis vectors
$\boldsymbol{a}^*$, $\boldsymbol{b}^*$, $\boldsymbol{c}^*$	Reciprocal lattice basis vectors
$\boldsymbol{B}$	Real lattice basis
$\boldsymbol{B}^*$	Reciprocal lattice basis
$\boldsymbol{h}$	Vector pointing from zero to an RLP
$\boldsymbol{k}$	Wave vector
$\boldsymbol{k}_0$	Incident wave vector
$\boldsymbol{k}_1$	Outgoing wave vector
$\boldsymbol{M}$	Matrix of Miller indices
$\boldsymbol{m}$	Miller indices
$\boldsymbol{N}$	Matrix of nodes in the reciprocal space
$\boldsymbol{n}$	Node in the reciprocal space
$\boldsymbol{q}$	A point in the reciprocal space. Depending on the point of view, it can be seen as a (negative) momentum transfer vector or as a spatial frequency
$\boldsymbol{R}$	A point in the 3D real space

$\boldsymbol{r}$	A point in the 3D real space
$\boldsymbol{v}$	Point in the rotogram
$\delta()$	Dirac delta function
$\hat{\boldsymbol{e}}$	Rotation axis of the composite rotation
$\hat{\boldsymbol{h}}$	Unit vector in the direction of $\boldsymbol{h}$
$\hat{\boldsymbol{k}}$	Unit vector in the direction of $\boldsymbol{k}$
$\hat{\boldsymbol{k}}_0$	Unit vector in the direction of $\boldsymbol{k}_0$
$\hat{\boldsymbol{k}}_1$	Unit vector in the direction of $\boldsymbol{k}_1$
$\hat{\boldsymbol{m}}$	Rotation axis
$\hat{\boldsymbol{q}}$	Unit vector in the direction of $\boldsymbol{q}$
$\hat{\boldsymbol{v}}$	Unit vector in the direction of $\boldsymbol{v}$
λ	Wavelength
$\mathcal{F}()$	Fourier transform
μ	Mean
ν	Frequency
$\overline{u^2}$	Mean square displacement of a scatterer
ϕ	Angle
ψ	Rotation angle
$\rho(\boldsymbol{r})$	Scatterer density
σ	Variance
θ	Rotation angle of the composite rotation
$\widetilde{m}$	Miller index
A	Amplitude of a wave
A'	Amplitude of the scattered wave

B	Debye-Waller factor, also called temperature factor or B-factor
$b_{i,j}$	Element of the real lattice basis
$c()$	Proximity function
$E_1(\boldsymbol{r}, t)$	Outgoing wave
$f()$	Score function
$F(\boldsymbol{q})$	Structure factor
I	Intensity
n	One coordinate of a node
$R_{\hat{\boldsymbol{q}}}(\phi)$	Rotation around the axis $\hat{\boldsymbol{q}}$ with an angle of ϕ
T	Period duration
t	Time
v_x, v_y, v_z	Coordinates of the rotogram
x, y, z	Coordinates of the real space
x^*, y^*, z^*	Coordinates of the reciprocal space
2D	Two dimensional
3D	Three dimensional
CMOS	Complementary metal-oxide-semiconductor
DNA	Deoxyribonucleic acid
FEL	Free-electron laser
K	Number of nodes
RLP	Reciprocal lattice point
SNR	Signal to noise ratio
SX	Serial crystallography
TEM	Transmission electron microscope
UCA	Uncertainty circular arc
ULS	Uncertainty line segment

Abstract

Crystallography is the most prolific method in structure determination of biological molecules. Serial crystallography advances this field with the cost of significantly higher computational complexity. This thesis presents solutions for one of the most affected stages in the data processing pipeline: The indexing stage. It improves the current data processing and enables very promising upcoming methods.

Zusammenfassung

Kristallographie ist die z.Zt. produktivste Methode zur Ermittlung der Struktur biologischer Moleküle. Die serielle Kristallographie entwickelt dieses Feld auf Kosten signifikant höheren Rechenaufwands weiter. Diese Dissertation stellt Lösungen für die am meisten betroffenen Algorithmen vor: Die Indizierungsalgorithmen. Dies verbessert die aktuellen- und ermöglicht die Anwendung neuer Methoden.

Chapter 1

Motivation

A fundamental desire of biology and the primary purpose of chemistry is to understand the structure of molecules on an atomic level. The arrangement of atomic nuclei and electrons in a molecule essentially determines its behavior and interaction with other molecules and particles. Structural knowledge of a molecule allows its technical exploitation. Examples include the design of new materials, modification of DNA, and drug design. A prominent application of the principle of inferring the underlying mechanisms by measuring molecule structures is structural biology. It is concerned with providing information about the atomic architecture of biological molecules - proteins in particular.

A very successful and widely used method to experimentally gain information about the structure of molecules is diffraction: The examined sample is irradiated, and the diffracted radiation is measured on a detector. The diffraction pattern is analyzed by the use of computational methods to extract from it information about the atomic structure of the sample. To recover the full three dimensional structure of the examined object, diffraction patterns from different orientations of the object need to be recorded.

The resolution of the reconstructed structure is theoretically limited by half of the wavelength of the employed radiation. To achieve atomic resolution, radiation with a wavelength in the order of 1 Å or less is required. Popular types of radiation are photons, electrons, and neutrons. All three kinds of radiation destructively interact with matter when used in the required range of wavelengths. Valuable signal from a sample can only be measured as long as the degree of damage is relatively small. In particular for biological samples, the tolerable radiation dose is very small. As a result, the signal measured from one single molecule before it is destroyed is not enough to reconstruct the molecule structure. To tackle this problem, the method of crystallography has been developed. Instead of a single molecule, a large crystal consisting of many such molecules is irradiated. The large number of molecules in the crystal

creates significantly higher diffracted intensities. Computational methods allow reconstructing the molecule structure from the diffraction images of the crystal. Often additional prior knowledge of the molecule is utilized to facilitate the reconstruction. Same as with a single molecule, the crystal needs to be measured in different orientations to collect enough information for the reconstruction. In conventional crystallography, this is done by rotating the crystal in the radiation beam. The most prolific method in protein structure determination is X-ray crystallography (Berman et al., 2000). Since the first application of crystallography over a century ago, it has furthered our knowledge of cellular processes and significantly facilitated the evolution of modern medicine.

Unfortunately, a large amount of highly interesting molecules currently cannot be crystallized to crystal sizes that allow structure determination by conventional crystallography. This and the desire to do time-resolved studies of chemical reactions have led to the development of *serial crystallography* in the recent decade. Instead of rotating a crystal in the beam to measure diffraction patterns from different orientations, multiple crystals of the same molecules are used. Each crystal is exposed to the radiation in only one orientation. By spending the total tolerable radiation dose to this orientation instead of distributing it over all required orientations, the diffracted signal is enhanced, thus increasing the signal to noise ratio. The diffraction patterns from the different crystals afterward can be merged using computational methods.

Serial crystallography allows using large numbers of small crystals to compensate for the lack of diffracting volume. Apart from the larger amount of crystals required, the price for the relaxed requirement on crystal size is a new experiment setup and significantly increased requirements on the data processing pipeline. One such requirement is the determination of the orientation of the crystal in the radiation beam from a single diffraction pattern - in place of a whole rotation series, which is available in conventional crystallography. This task is known as “indexing”, and the state-of-the-art algorithms that attempt to solve it are clearly in need of improvement. All diffraction patterns that cannot be indexed cannot be utilized by further data processing stages, therefore wasting the respective precious crystals and experiment time. This thesis presents an algorithm that significantly improves the indexing performance without adding restrictions on the experiment setup or the data processing pipeline. As a result, typically, significantly fewer crystals are wasted, which leads to an improved quality of the extracted molecule structures or a shorter experiment duration.

Serial crystallography experiments today are typically performed at synchrotrons or free-electron lasers and require many thousands of crystals. This limits the practicality of the method, since on the one hand, the large facilities are very costly, and beam time is limited. On the other hand, crystals of interest-

ing biological molecules often are very precious and cannot be produced in large quantities[1]. Two very promising upcoming methods for serial crystallography tackle one of these problems each. The first, *serial electron crystallography*, performs the diffraction experiment in transmission electron microscopes that are relatively inexpensive and widespread. This gives many groups quick and easy access to powerful serial crystallography experiments in contrast to the need of writing proposals to the large facilities and waiting for months to perform an experiment in the assigned beamtime. The second method, *pink beam serial crystallography*, allows more efficient use of the X-ray beam of synchrotrons or free-electron lasers and can drastically reduce the number of crystals required for a complete dataset. This enables a larger range of molecules to exploit the benefits of serial crystallography. Both methods have in common that several groups around the world have performed several experiments, and collected vast amounts of data without the availability of a working automatic data processing pipeline that is necessary for feasible serial crystallography. The bottleneck is the indexing stage. The changes in the experiment setup relaxed the requirements on the beam and some parts of the data processing pipeline but tightened the requirements on the indexing algorithm. As a result, conventional indexing algorithms fail in determining the orientation of crystals in the beam, by this prohibiting further processing of the diffraction patterns. This thesis presents an algorithm that can index both serial electron and pink beam crystallography data as well as normal serial crystallography data with unprecedented accuracy, enforcing only insignificant additional requirements on the experiment setup. Removing the bottleneck, the new algorithm enables automatic data processing for pink beam and electron crystallography, by this enabling a wide exploitation of these very promising techniques.

Crystallography suffers from the so-called phase problem that arises because current detectors cannot measure the phase but only the intensity of the incoming radiation. One method that potentially can tackle this problem is the *convergent beam serial crystallography*. This technique is very new, and only little data has been collected since the availability of X-ray lenses with high numerical aperture (Morgan et al., 2015). Same as with the two above techniques, convergent beam serial crystallography data cannot be indexed by conventional indexing algorithms. This thesis presents a slight modification of the algorithm for pink beam and electron beam indexing that can index diffraction patterns of convergent beam serial crystallography experiments. A working indexing step is essential for future exploitation of this upcoming and very promising method.

[1]It is not uncommon, that scientists would need to spend months or years to produce enough sample material to grow the number of crystals required for a serial crystallography experiment.

Chapter 2

Introduction

2.1 Basic physical concepts

The concepts in this thesis can be applied to different types of radiation, e.g. photons, electrons, and neutrons. The models presented here base on a common property among them all: Their propagation can be modeled by a wave.

2.1.1 Scattering

A wave in one dimension can be described either by its real form $E(r,t) = A\sin(2\pi(kr - \nu t) + \phi)$ or the complex form $E(r,t) = Ae^{i(2\pi(kr-\nu t)+\phi)}$ with the positive amplitude A, the wavenumber $k = \frac{1}{\lambda}$, the frequency $\nu = \frac{1}{T}$ (corresponding to the period T) and the phase shift ϕ. In three dimensions the description of a planar wave then results in

$$E(\boldsymbol{r},t) = Ae^{i(2\pi(\boldsymbol{k}\cdot\boldsymbol{r}-\nu t)+\phi)} \tag{2.1}$$

with the wave vector $\boldsymbol{k}$.

For the computation of scattering effects, a very simple model of elastic scattering that operates on point scatterers is employed: The Huygens-Fresnel principle states that if a scattering object is affected by a wave, it becomes the source of a spherical wave with the same wavelength as the stimulating wave and with a relative phase shift of π. Moreover, we are only interested in the far-field, i.e. the detector distance is assumed to be large with respect to distances within the ensemble of scatterers. This allows approximating the spherical scattered waves as planar waves when computing the interference of multiple scattered waves in one specific direction, i.e. approximating the detector distance to be infinitely large. The amplitude of the scattered wave in this simple model is only

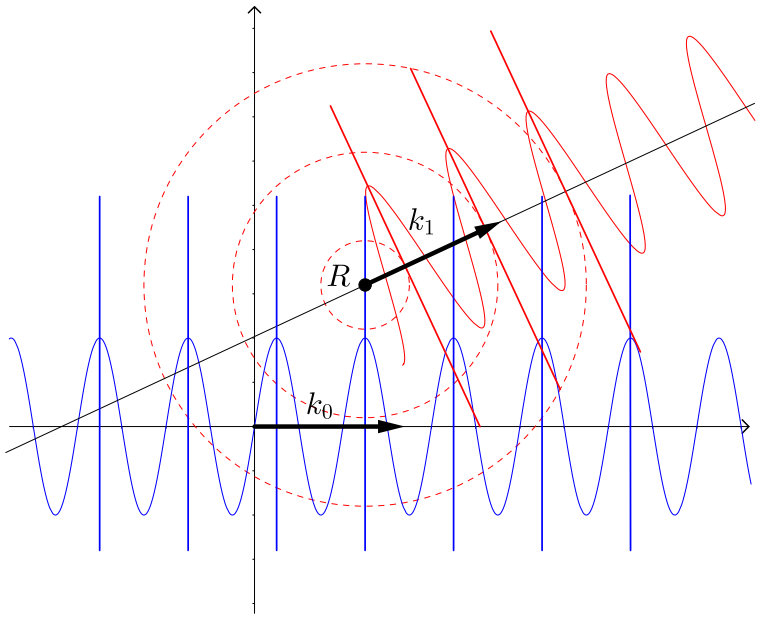


Figure 2.1: *Visualization of a scattering event at position* $\boldsymbol{R}$*. The incoming wave (blue) has the wave vector* $\boldsymbol{k}_0$*. The scatterer at position* $\boldsymbol{R}$ *is marked as a black dot. The scattered wave (red) is spherical but can be treated as a planar wave with wave vector* $\boldsymbol{k}_1$ *when computing the interference at an infinitely far located detector in the direction of* $\boldsymbol{k}_1$*. The blue/red lines and the dashed circles mark the locations of maximum amplitude.*

dependent on the scattering object and is proportional to the amplitude of the stimulating wave, i.e. $A' \sim A$. The proportionality factor depends on the type of the scatterer and the type of radiation. With an incident wave in direction $\hat{\boldsymbol{k}}_0$, the scattered wave of a scatterer in position $\boldsymbol{R}$ approximated as a planar wave in the direction $\hat{\boldsymbol{k}}_1$ can be described as

$$E_1(\boldsymbol{r}, t) = A' e^{i(2\pi(\boldsymbol{k}_1 \cdot \boldsymbol{r} - \nu t + \boldsymbol{k}_0 \cdot \boldsymbol{R} - \boldsymbol{k}_1 \cdot \boldsymbol{R}) + \phi - \pi)} \ , \ \|\boldsymbol{k}_1\| = \|\boldsymbol{k}_0\| \tag{2.2}$$

The scattering is elastic so $\|\boldsymbol{k}_1\| = \|\boldsymbol{k}_0\|$ and $\Delta\phi = -\pi$. The phase shift $\boldsymbol{k}_0 \cdot \boldsymbol{R}$ is due to the position of the scatterer relative to the incoming wave (blue in figure 2.1). The scattered wave (red in figure 2.1) has its origin at the position of the scatterer, but needs to be described in the global coordinate system, hence the phase shift $-\boldsymbol{k}_1 \cdot \boldsymbol{R}$.

In an ensemble of similar scatterers, each scatterer is a source of a wave. The interference on an infinitely distant detector in direction of $\boldsymbol{k}_1$ can now be computed by summing the contributions of all scatterers:

$$E_1(\boldsymbol{r},t) = \sum_n A' e^{i(2\pi(\boldsymbol{k}_1\cdot\boldsymbol{r}-\nu t+\boldsymbol{k}_0\cdot\boldsymbol{R}_n-\boldsymbol{k}_1\cdot\boldsymbol{R}_n)+\phi-\pi)} = \\ = A' e^{i(2\pi(\boldsymbol{k}_1\cdot\boldsymbol{r}-\nu t)+\phi-\pi)} \sum_n e^{-i2\pi(\boldsymbol{k}_1-\boldsymbol{k}_0)\cdot\boldsymbol{R}_n} \tag{2.3}$$

Using a scatterer density $\rho(\boldsymbol{r})$ instead of distinct scatterers and introducing the scattering vector $\boldsymbol{q} = \boldsymbol{k}_1 - \boldsymbol{k}_0$, above formula can be rewritten to

$$E_1(\boldsymbol{r},t) = A' e^{i(2\pi(\boldsymbol{k}_1\cdot\boldsymbol{r}-\nu t)+\phi-\pi)} \int \rho(\boldsymbol{r}) e^{-i2\pi\boldsymbol{q}\cdot\boldsymbol{r}} d\boldsymbol{r} = F(\boldsymbol{q}) A' e^{i(2\pi(\boldsymbol{k}_1\cdot\boldsymbol{r}-\nu t)+\phi-\pi)} \tag{2.4}$$

Apart from constants, the amplitude and phase of the scattered wave are defined by the integral in the above function. This integral is usually called the structure factor $F(\boldsymbol{q}) = \int \rho(\boldsymbol{r}) e^{-i2\pi\boldsymbol{q}\cdot\boldsymbol{r}} d\boldsymbol{r}$, it is the 3D Fourier transform of the scatterer density $\rho(\boldsymbol{r})$ with respect to the scattering vector $\boldsymbol{q}$. The complex valued $F(\boldsymbol{q})$ ranges over the so called reciprocal space - as opposed to the real space where the real valued $\rho(\boldsymbol{r})$ is defined. Measuring the amplitude and phase of a scattered wave in direction $\hat{\boldsymbol{k}}_1$ is then equivalent to measuring the reciprocal space (i.e. the Fourier transform of the real space) at position $\boldsymbol{q}$. The requirement $\|\boldsymbol{k}_1\| = \|\boldsymbol{k}_0\|$ constrains $\boldsymbol{q}$ to be on a sphere in the reciprocal space with radius $\frac{1}{\lambda}$ and center at $-\boldsymbol{k}_0$. This sphere is called Ewald sphere, see figure 2.2.

The Ewald sphere construction shows that with a fixed wavelength and distribution of scatterers (also called "sample"), only a miniscule part of the reciprocal space can be measured - the part on the Ewald sphere. This effect is similar to gathering only a part of the 3D information in imaging by recording an image of an object from only one direction. A typical way to overcome this restriction is to measure diffraction from different views by rotating the sample and merge it to a 3D model. Another restriction shown by the Ewald sphere is that $\|\boldsymbol{q}\| \leq \frac{2}{\lambda}$. This is similar to multiplying the Fourier transform with a top-hat function, which is equivalent to low-pass filtering the scatterer density. This shows that the minimal wavelength of the employed radiation source limits the maximum measurable resolution of $\rho(\boldsymbol{r})$.

2.1.2 Limitations

Apart from the wavelength, several effects limit the measurable resolution and the signal to noise ratio of the measured signal. These effects and their significance vary for different types of radiation (e.g. photons, electrons, or neutrons) and sample materials (e.g. metal, perovskites, or proteins). These effects are not modeled by the simple scattering model above.

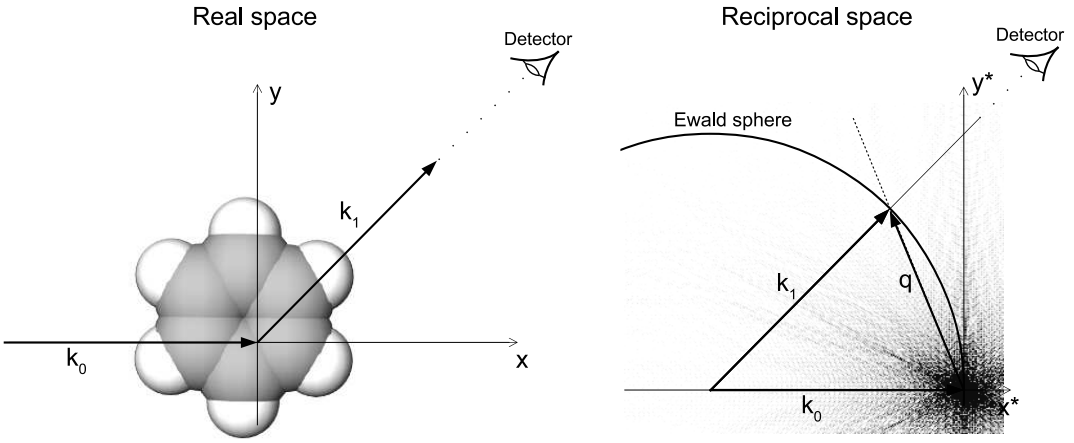


Figure 2.2: *Relation between the real space and the reciprocal space. In real space a plane wave in the direction of $\boldsymbol{k}_0$ excites the scatterers described by the scatterer density $\rho(\boldsymbol{r})$ (illustrated as a molecule, image generated with MolView (Bergwerf, 2014)). An infinitely distant detector measures the diffracted wave in the direction of $\boldsymbol{k}_1$. The amplitude and phase of the diffracted wave (encoded in $F(\boldsymbol{q})$) can be computed by evaluating the Fourier transform of the real space in the reciprocal space at the point $\boldsymbol{q} = \boldsymbol{k}_1 - \boldsymbol{k}_0$, i.e. $F(\boldsymbol{q}) = \mathcal{F}(\rho(\boldsymbol{r}))(\boldsymbol{q})$. The scattering is elastic, so $\|\boldsymbol{k}_1\| = \|\boldsymbol{k}_0\|$. This has the effect that all possible $\boldsymbol{q}$ are located on a sphere - the so called Ewald sphere. Only points on the Ewald sphere can by measured by an infinitely distant detector.*

2.1.2.1 Shot noise

The radiation employed in diffraction experiments is quantized, i.e. it can only be detected in packets of defined energy. The number of packets arriving at a given detector in a certain time interval can be modeled by a Poisson distributed stochastic process where the expected value μ is proportional to the intensity (i.e. the squared amplitude, $I = A^2$) of the respective wave: $\mu \sim I$. The variance in a Poisson distribution is coupled to the mean by $\sigma^2 = \mu$. The noise that is inherent to the stochastic process is called shot noise σ. Knowing the signal and the noise allows the computation of the maximum achievable signal to noise ratio (SNR) as $\frac{\mu}{\sigma} = \frac{I}{\sqrt{I}} = \sqrt{I} = A$, i.e. for a fixed detector and measurement time low intensities have a low SNR. Additional noise added by the detector and other factors involved in the experiment setup decrease the SNR further. In conclusion, high diffracted intensities I are required for precise measurements.

2.1.2.2 Temperature factor

The scatterers for diffraction experiments considered in this thesis are electrons and atomic nuclei in large molecules. Both vibrate due to temperature. If the time scale of measuring the diffracted signal is in the same order of magnitude as the thermal motion or longer, this movement smears out the scatterer density. This is equivalent to a low pass filter, which lowers the norm $\|F(\boldsymbol{q})\|$ at high $\|\boldsymbol{q}\|$. This effect can be described by the Debye-Waller factor, also called temperature factor or B-factor $B = 8\pi^2\overline{u^2}$ where $\overline{u^2}$ is the mean square 3D displacement of a scatterer under the assumption of isotropic temperature movement. The magnitude of the structure factor $\|F(\boldsymbol{q})\|$ (i.e. the amplitude of scattered waves) is scaled by a factor of $e^{-\frac{B\|\boldsymbol{q}\|^2}{4}}$, i.e. the high-resolution signal is exponentially attenuated.

2.1.2.3 Radiation damage

Not all interaction of waves with matter is elastic, i.e. without a loss of energy. A large part is inelastic, thus transferring energy into the matter. E.g. in a typical conventional macromolecule crystallography experiment, 98% of the 12 keV X-ray photons pass through the sample without interaction. Of the remaining 2%, 84% of the photons are absorbed damaging the molecule, 8% undergo Compton scattering (i.e. the scattered wavelength is smaller than the original one), and only the remaining 8% undergo useful elastic scattering (Spence et al., 2012).

The amount of energy deposited in a medium of unit mass is called dose. With conventional approaches, the dose is limited to values below which the degradation is considered to be too severe, as too severely damaged molecules produce more noise than valuable signal. This limits the scattered energy per unit mass. Using the "diffraction before destruction" principle (Chapman et al., 2014) with extremely short radiation pulses, this limitation can be partially overcome. Nevertheless, in practice, it does not entirely remove the limitation on the scattered energy per mass - it is still limited by the achievable intensity of the radiation source and possibly other high-intensity effects.

2.2 Conventional crystallography

The achievable resolution and signal to noise ratio in measurements of real molecules are limited by the temperature factor and the radiation damage. One way to overcome this limit is to increase the mass of the measured object by measuring N molecules at the same time, thus increasing the scattered amplitude

by a factor of N. For many molecules, this can be achieved by growing crystals out of them. The method of using crystals for diffraction experiments to solve (i.e. determine) the molecular structure is called crystallography.

A crystal is a translationally symmetric arrangement of molecules or clusters of molecules. The translational symmetry can be described by a lattice. A lattice $\boldsymbol{\Lambda}$ in $\mathbb{R}^3$ is a set of points that can be described by

$$\boldsymbol{\Lambda} = u\boldsymbol{a} + v\boldsymbol{b} + w\boldsymbol{c},\ u, v, w \in \mathbb{Z} \tag{2.5}$$

Here $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are basis vectors that form the lattice basis $\boldsymbol{B} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. An alternative way to describe a lattice is to define the lattice points as Dirac delta functions, i.e $\boldsymbol{\Lambda}(\boldsymbol{r}) = \sum_{u,v,w\in\mathbb{Z}} \delta^3(\boldsymbol{r} - \boldsymbol{B}\begin{pmatrix} u \\ v \\ w \end{pmatrix})$. Using the scatterer density in the molecule $\rho(\boldsymbol{r})$, a crystal can be described by the convolution $\rho(\boldsymbol{r}) * \boldsymbol{\Lambda}(\boldsymbol{r})$.

The Fourier transform of a lattice $\boldsymbol{\Lambda}$ is a lattice as well - the reciprocal lattice $\boldsymbol{\Lambda}^*$ (as opposed to the direct lattice in the real space):

$$\boldsymbol{\Lambda}^* = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*,\ h, k, l \in \mathbb{Z} \tag{2.6}$$

A convenient derivation has been published by Thompson, 1996. The integers h, k, and l that define a reciprocal lattice point (RLP) are called "Miller indices". The reciprocal lattice $\boldsymbol{\Lambda}^*$ has the basis vectors

$$\begin{aligned} \boldsymbol{a}^* &= \frac{\boldsymbol{b} \times \boldsymbol{c}}{\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c})} \\ \boldsymbol{b}^* &= \frac{\boldsymbol{c} \times \boldsymbol{a}}{\boldsymbol{b} \cdot (\boldsymbol{c} \times \boldsymbol{a})} \\ \boldsymbol{c}^* &= \frac{\boldsymbol{a} \times \boldsymbol{b}}{\boldsymbol{c} \cdot (\boldsymbol{a} \times \boldsymbol{b})} \end{aligned} \tag{2.7}$$

The basis vectors of the reciprocal lattice are orthogonal to planes spanned by respectively two basis vectors of the real lattice. Their length is reciprocal to the spacing of these lattice planes. The calculation of the basis $\boldsymbol{B}^* = (\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)$ can be written in matrix representation as

$$\boldsymbol{B}^* = (\boldsymbol{B}^{-1})^T \tag{2.8}$$

Using the reciprocal lattice, the Fourier transform of a crystal can be derived as $\mathcal{F}(\rho(\boldsymbol{r}) * \boldsymbol{\Lambda}(\boldsymbol{r})) = F(\boldsymbol{q})\boldsymbol{\Lambda}^*(\boldsymbol{q})$, i.e. the reciprocal space is sampled by the reciprocal lattice and thus becomes discrete. A one-dimensional illustration of the relationship between real and reciprocal space is given in figure 2.3. A

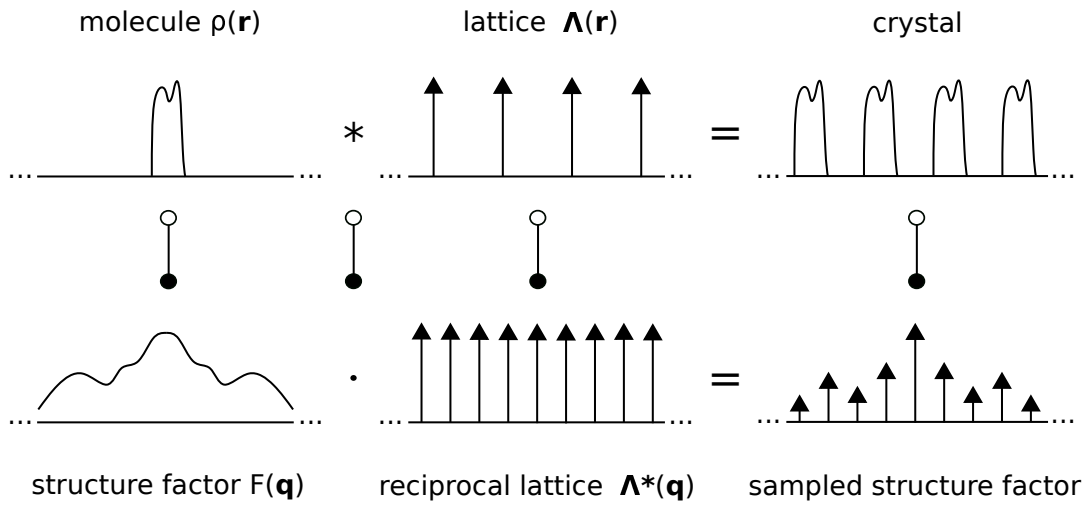


Figure 2.3: *One-dimensional illustration of the mathematical model of a crystal and its Fourier transform. A crystal can be described by the convolution of a molecule with a lattice:* $\rho(\boldsymbol{r}) * \boldsymbol{\Lambda}(\boldsymbol{r})$*. The Fourier transform of a crystal can be derived by multiplying the Fourier transforms of the molecule and the lattice:* $F(\boldsymbol{q})\boldsymbol{\Lambda}^*(\boldsymbol{q})$

two dimensional illustration of the scattering effects for a crystal is given in figure 2.4.

As $F(\boldsymbol{q})\boldsymbol{\Lambda}^*(\boldsymbol{q})$ is discrete, the directions in which nonzero amplitudes can be measured are discrete as well. Diffraction is only measurable if a lattice point is located on the Ewald sphere. Such a lattice point that satisfies the diffraction condition is called "excited". If two-dimensional detectors are utilized, excited lattice points produce spots on the detector. These spots are called Bragg spots or simply peaks.

In an ideal, infinitely large crystal, a reciprocal lattice point must be located precisely on the Ewald sphere to cause diffraction. The Bragg spot in this model is infinitely sharp. In practice, the crystal does not extend to infinity, which can be modeled by the multiplication of the direct lattice with a top-hat function in the shape of the crystal. The effect in the reciprocal space is the convolution of the reciprocal lattice with the Fourier transform of the top-hat function, i.e. low pass filtering or smoothing. As a result, the Fourier transform of the crystal is not discrete in practice, but the points in the reciprocal space have a nonzero extension - the smaller the crystal, the larger the extent of the RLPs. Other properties that cause similar effects are strain in the crystal or imperfectly grown crystals that consist of mosaic blocks (mosaicity) (Nave, 2014). This way, even lattice points that are close to the Ewald sphere, but not exactly on it, are excited and thus create Bragg spots. In practice, usually several reciprocal

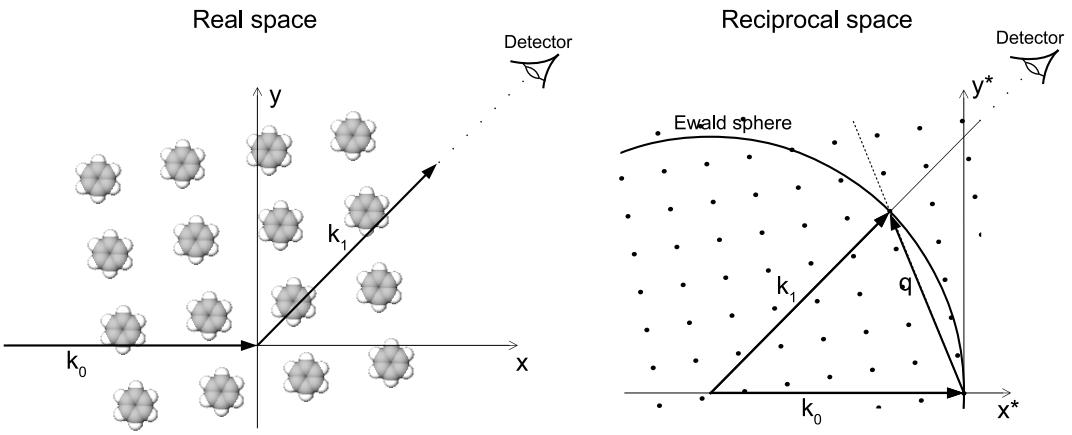


Figure 2.4: *Relation between the reciprocal space and the real space for a crystal. In the real space, a crystal is illustrated as molecules (image generated with MolView (Bergwerf, 2014)) on a lattice. The Fourier transform of a crystal $F(\boldsymbol{q})\boldsymbol{\Lambda}^*(\boldsymbol{q})$ is discrete. This means that the directions in which nonzero amplitude can be measured are discrete as well. Diffraction is only measurable in the directions where the Ewald sphere intersects a lattice point.*

lattice points are excited simultaneously.

The structure factor at all relevant reciprocal lattice points needs to be measured to be able to reconstruct the structure of the examined sample. In conventional crystallography, this is done by rotating the crystal with a fixed angular velocity in the beam. As rotation in the real space corresponds to the same rotation in the reciprocal space, reciprocal lattice points are swept through the Ewald sphere, creating measurable Bragg spots.

2.2.1 Data processing pipeline

The diffracted radiation typically is recorded on a two-dimensional detector. The detector integrates the diffraction for a short fixed time interval, saves it as an image, and starts the integration for the next time interval. One image is called a diffraction pattern. The total collected data from a rotation of a crystal is called a rotation series.

2.2.1.1 Detector correction

As a first step in the data processing pipeline, the collected diffraction patterns are corrected for known detector and experiment artifacts. Today, often 2D

CMOS[1] detectors are employed. They often consist of different panels with gaps in between. The data is saved per panel and needs to be assembled according to the detector geometry (Barty et al., 2014). Typically not all available pixels are fully functional. Malfunctioning pixels need to be identified and masked. Detectors usually need frequent re-calibration. Calibration constants that allow computing the absorbed energy from the measured quantities must be estimated and applied to the raw data. The probably most prominent calibration constant is the dark offset, i.e. the average noise signal that is generated when the beam is off. This value can change significantly due to temperature changes and other environmental effects. Periodic measurements without signal can be used to correct even changing dark offsets. Other calibration constants like the gain or optional switching thresholds are more challenging to determine. Some even need a separate calibration setup or special calibration-circuitry (Mezza et al., 2016).

The experimental setup can induce unwanted artifacts in the diffraction pattern. The sources of such artifacts are numerous: Shadows of installations can be recorded, magnetic lenses in electron microscopes can cause aberrations, artefacts from the sample delivery device can cause massive amounts of noise in specific regions of the diffraction image. All these effects need to be corrected for or masked if uncorrectable.

2.2.1.2 Identification of excited reciprocal lattice points

In the second step, the locations of the recorded Bragg spots in the patterns are identified. The respective algorithms are called peak finders. The typical approach is to estimate the background upon which to decide whether a peak is significantly above the noise level. The background estimation can be local to the examined area or take advantage of knowledge about the experimental setup. E.g., many diffraction experiments have a radially symmetric background that can be used to either speed up or improve the quality of the background estimation. Figure 2.5 (b) shows the peaks detected on a real diffraction pattern. Peak finders typically detect a subset of the Bragg spots plus some spurious peaks, i.e. false positives.

Using the geometry of the experimental setup, the detected peaks, and the known wavelength of the utilized radiation, the wave vectors $\boldsymbol{k}_1$ of peak intensity can be identified in the real space. This allows mapping the location of the found Bragg spots to locations on the Ewald sphere in the reciprocal space. These locations are good approximations of the excited reciprocal lattice points, given that the found peaks are not spurious ones. Due to the known rotation

[1]See List of Symbols for descriptions of acronyms.

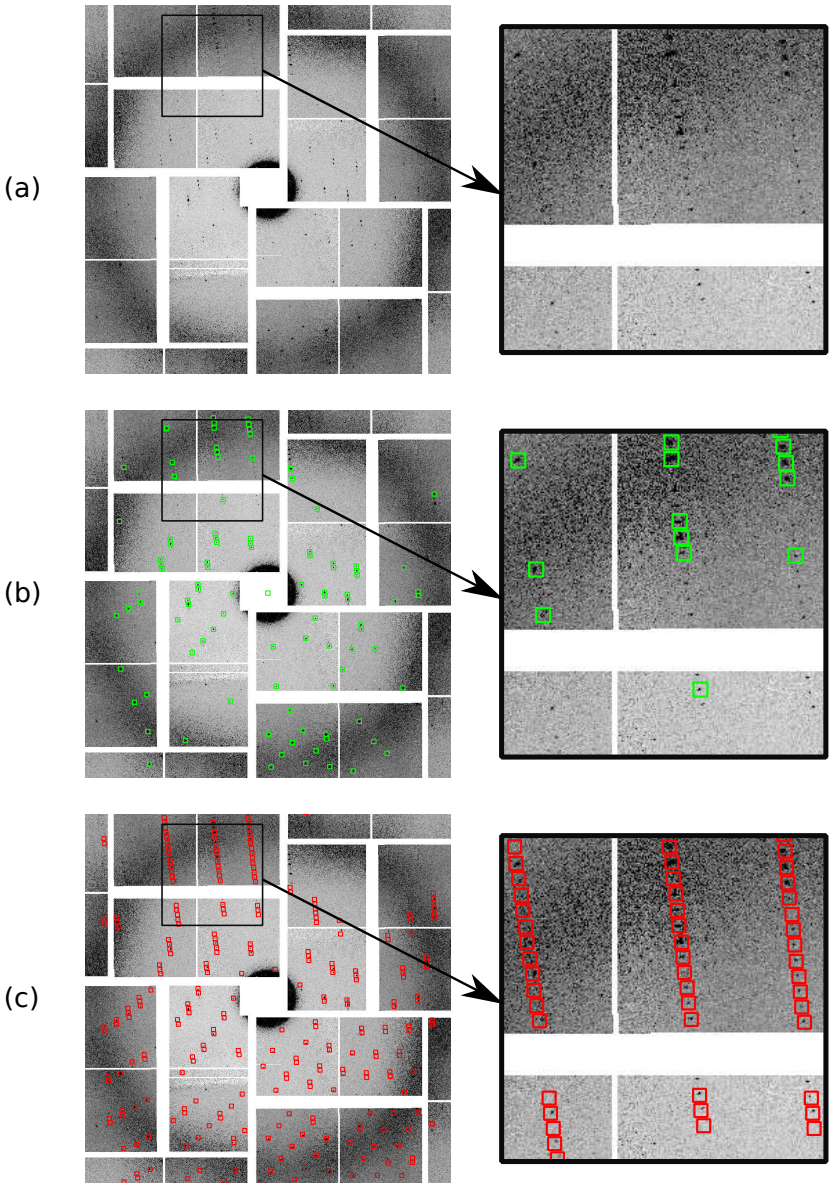


Figure 2.5: *Visualization of peak detection for indexing and prediction after indexing. (a) shows a diffraction pattern of a crystal. In (b), the peaks detected by the peak finding algorithm are marked green. Several weak peaks have not been detected - probably due to conservative peak finding parameters. The indexing step was successful, (c) shows the predicted peaks marked red. Bragg spots previously missed by the peak finder are now correctly predicted and can be integrated. The diffraction data is of crystals from serotonin receptor* $5-\mathrm{HT}_{2B}$ *bound to ergotamine. The dataset is publicly available from the CXIDB (Maia, 2012) entry 21 (Liu et al., 2013).*

angle between consecutive recordings, these approximate RLPs from different diffraction patterns can be merged to cover a significant part of the reciprocal lattice. The merge significantly simplifies the next step.

2.2.1.3 Indexing

In the third step, the reciprocal lattice is determined based on the identified approximate RLPs. Due to the known difference in rotation angle between consecutive diffraction patterns, the RLPs identified in different diffraction patterns can be easily merged into a combined dataset. Common approaches to identify the crystal orientation are the Fourier methods. The main idea is to detect maxima in the Fourier transform of the RLPs (Steller et al., 1997). The RLPs are modeled as Dirac delta functions. Such maxima mark the directions with maximum periodic repetition and the respective period, i.e. directions orthogonal to the lattice planes of the reciprocal lattice and their inter-plane distances. This allows directly recovering the reciprocal lattice. Another perspective on this method is, that the measured RLPs are a part of the reciprocal lattice, which can be modeled by multiplication of the reciprocal lattice with a binary mask function[2]. The Fourier transform of the partial lattice can, therefore, be expressed as the convolution of the real-space lattice with the Fourier transform of the mask function. The result is a smeared version of the real space lattice. The main maxima are the lattice points of the real space lattice. In the real space, all lattice points can be detected as main maxima, although not all RLPs are available in the partial reciprocal lattice. The relationship is visualized in figure 2.6. Having determined the real space lattice, the reciprocal lattice can be recovered by the relation in equation 2.8. Due to this relation, the Fourier methods sometimes are called the real-space indexing methods. More detailed discussions about various indexing algorithms are given in the chapters 3 and 5.

The determination of the reciprocal lattice is required to calculate all locations where Bragg spots are supposed to be measured. This calculation is typically referred to as “prediction”. In the process of prediction, Miller indices are assigned to the predicted Bragg spot locations - which is the final result of this step. Consequently, this step is called indexing. Figure 2.5 shows a typical diffraction pattern that has been successfully indexed.

2.2.1.4 Integration and scaling

The prediction not only assigns Miller indices to already detected peaks but also identifies the locations of Bragg spots that have been missed by the peak

[2]The mask is a function that has the value 1 at the position of an observed peak and the value 0 at the position of not observed peaks.

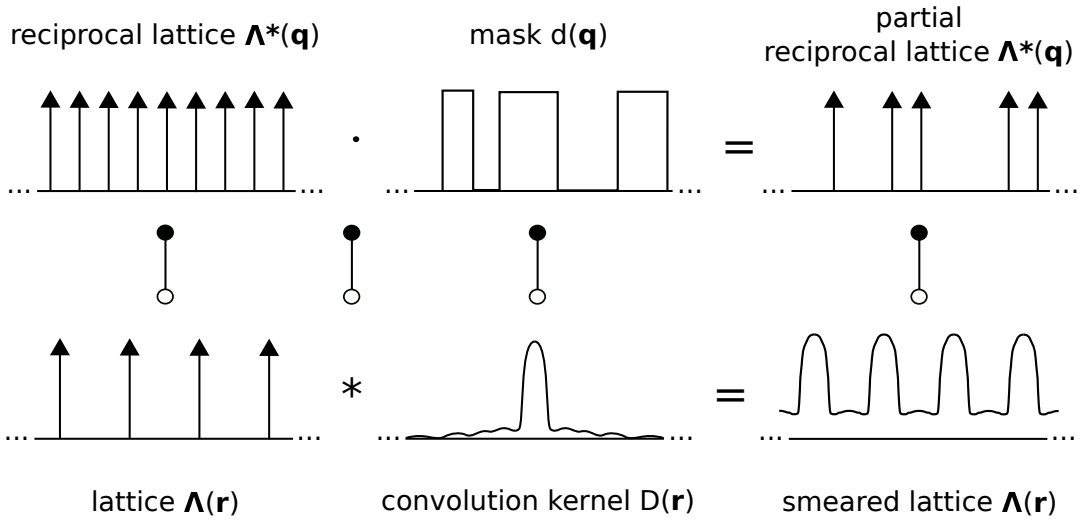


Figure 2.6: *One-dimensional illustration of the derivation of the Fourier transform of a partial lattice, as used by the Fourier methods. The partial reciprocal lattice consists of the measured RLPs and can be modeled by multiplication of the reciprocal lattice with a binary mask function. The Fourier transform of the partial lattice can, therefore, be expressed as the convolution of the real-space lattice with the Fourier transform of the mask function. The result is a smeared version of the real space lattice.*

finder. Thus prediction discloses the false positives of the peak finding stage, as these peaks are not predicted. The predicted locations are used to measure the intensity of the Bragg spots by integrating across their shape on the detector. This step is called integration. Subsequently, in the scaling step, the integrated intensity is scaled to account for known effects that arise from the radiation and experimental setup. This step e.g. corrects for the different times that RLPs at different $\|\boldsymbol{q}\|$ are in diffraction condition (Buerger, 1940), different crystal sizes and varying beam intensities. Unknown quantities like the crystal size need to be estimated from the diffraction patterns.

2.2.1.5 Phase retrieval

The last step in principle should be the inverse Fourier transform of $F(\boldsymbol{q})$ and thus the identification of the scatterer density in the molecule. Unfortunately, this typically is not possible, since current detectors for X-rays, electrons, and neutrons cannot measure the phase, but only the intensity of the radiation, i.e. only the squared magnitude of the structure factor.

There are efforts to measure the phase with electrons (Dorset, 1992, 1996; Unwin et al., 1975; Wan et al., 2012), but this is not yet established for general

applications.

For large molecules, often the knowledge of a similar molecule structure can be used as a starting point for iterative phase determination (Crowther, 1972; Rossmann et al., 1962). Another possibility is to exploit physical effects such as anomalous scattering (Hendrickson, 1991). Despite large success for many types of molecules, the phase problem remains a field of ongoing research.

Once the phases of the structure factors are determined, the scatterer density can be computed, and based on that, chemical models of the molecule can be fitted. The quality of the chemical models depends on the resolution and the quality of the scatterer density, which itself depends on up to which $\|\boldsymbol{q}\|$ the structure factor $F(\boldsymbol{q})$ has been measured with sufficiently high signal to noise ratio.

2.2.2 Practical limitations of conventional crystallography

Due to the tolerable dose, measurements to high resolution in conventional crystallography require large crystals. The maximum affordable size of crystals for an experiment in practice is limited by several factors from different fields: The crystallization capabilities of the employed molecules, the availability of computational methods that account for complex effects, and the experimental setups.

First, it is not always even feasible to have large crystals. Many molecules are complicated to crystallize. Examples include membrane proteins that compose 30% of all proteins and play an important role in drug delivery. It has been estimated that 70% of drug molecules interact with a membrane protein (Spence et al., 2004). These molecules often grow to very small or very imperfect crystals, both limiting the scattered intensity.

Second, in the simple model for scattering presented at the beginning of this chapter, scatterers are affected only by the incident wave. By only slightly improving the model, it becomes clear that the scattered wave of one scatterer also affects all the other scatterers in the molecule. A theory that covers multiple scattering is called “dynamical theory”. The thicker the crystal in beam direction, the more significant are the dynamical effects. For crystals of large molecules sufficiently good methods to process data with a significant ratio of dynamic scattering are not available yet, thus limiting the maximum usable crystal size to one where dynamical effects are negligible. This is particularly important for electron beams due to their strong interaction with matter.

Last but not least, new experimental setups also limit the crystal size. Examples are time-resolved measurements where a chemical reaction is initiated

before measuring the diffraction with a fixed delay. The initiation can e.g. be done by illuminating the crystal with visible light[3] or diffusing chemicals into it. The former is limited by the penetration depth of the visible light and the latter by the diffusion rate. Both require small crystals. Time-resolved measurements with conventional crystallography intrinsically have a time resolution that is longer than the time to rotate the crystal, i.e., in the range of seconds. In contrast, desired time resolutions often are orders of magnitude shorter (Valafar et al., 2012).

2.3 Serial crystallography

With only small crystals available, collecting a rotation series is often unfeasible, since the allowed dose has to be spread across all diffraction patterns. Thus, it is reasonable to spend the total dose on one single orientation, increasing the scattered intensity and, therefore, the achievable resolution and SNR. This approach allows only a small part of the reciprocal space to be measured using one single crystal. The solution to this problem is to use measurements from many thousands of crystals to build up a complete data set. This approach is called serial crystallography (SX) (Boutet et al., 2012; Chapman et al., 2011; Schlichting, 2015). Eliminating the need for a time-consuming rotation, SX enables time-resolved measurements of several orders of magnitude better time-resolution than what is possible with a rotation series.

2.3.1 Sample delivery

While promising to solve several old problems, SX brings with it a set of new problems that do not arise with conventional crystallography. Probably the most tangible one is the sample delivery. In serial crystallography, typically many thousands of diffraction patterns need to be recorded. To do this in a tractable time, an automatic way of replacement of the crystals in the beam needed to be developed. Today a variety of sample delivery techniques exist, examples include gas-focused liquid jets (DePonte et al., 2008), extrusion devices for viscous media (Weierstall, 2014) and crystals on a moving silicon chip (fixed target) (Roedig et al., 2017). Electron beams can be moved by magnetic fields, leaving the crystals on a silicon chip completely still (Bücker et al., 2020). A typical experimental setup for X-ray serial crystallography with a gas-focused liquid jet is illustrated in figure 2.7.

[3]Infrared or ultraviolet light is also common.

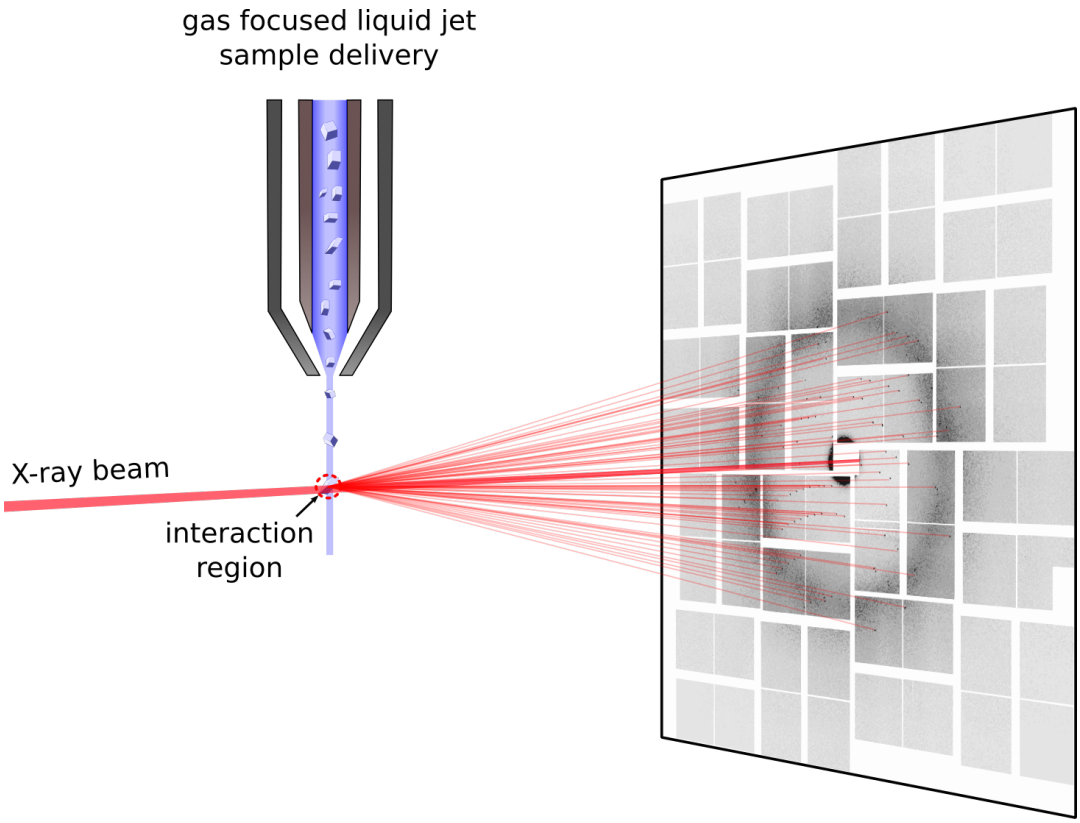


Figure 2.7: *A typical X-ray serial crystallography setup. The crystals are delivered to the beam by a gas focused liquid jet. In the interaction region, the crystals are irradiated by X-rays and cause diffraction into Bragg spots. The water of the liquid jet and in the crystals causes a radially symmetric background noise. The diffraction is measured by a two-dimensional segmented pixel array detector. The diffraction image is of crystals from serotonin receptor* $5-\mathrm{HT}_{2B}$ *bound to ergotamine, publicly available from the CXIDB (Maia, 2012) entry 21 (Liu et al., 2013).*

2.3.2 Partiality estimation

Another problem arising with SX is the so-called reflection partiality correction. While in a rotation series, typically, whole RLPs are integrated in a diffraction pattern by being swept through the Ewald sphere, in serial crystallography, the RLPs are only partially excited since the crystal orientation does not change during the exposure. The Ewald sphere intersects an RLP and only excites the 2D shape of the intersection, not the whole 3D shape[4] of the RLP. This partial

[4]As mentioned above, due to the finite size of the crystal the RLPs are low pass filtered and thus smoothed

excitation has to be corrected for. The estimation of the partiality is a difficult problem that requires careful physical modeling and subtle computational refinement of the parameters. While there has been progress in recent years in this field (Ginn et al., 2015; Kroon-Batenburg et al., 2015; White et al., 2013), the partiality correction remains a field of ongoing research. The problem of not having sufficiently precise partialities is currently addressed by averaging the corrected intensities from a large number of measurements (Kirian et al., 2010). This method is known as Monte Carlo integration. It has proven to be functional since the early SX at free-electron lasers (FELs) (Kirian et al., 2011). The noise induced by wrong partiality estimation, and as a result the large number of measurements required for Monte Carlo integration, is one of the reasons why serial crystallography has a considerably high sample consumption of many thousands of crystals per complete dataset.

2.3.3 Indexing still diffraction patterns

The current sample delivery techniques for serial crystallography cannot control the orientation of the crystals in the beam. Partial orientation alignment occurs in fixed target sample delivery techniques that sometimes show a preferred orientation when crystal faces stick to the silicon chips (Cohen et al., 2014). Similarly, the liquid jets sometimes show preferred orientation when elongated crystals are oriented along the jet direction. Both effects are typically parasitic, have at least one degree of freedom, and depend on the form of the crystal. Unexpected preferred orientation alignment can even be counterproductive, e.g. when it causes only partial measurement of the reciprocal space because orientations that would excite the missing parts simply are not available. In practice, the orientation of crystals in the beam is considered to be random, and preferred orientation is avoided e.g. by having different inclinations of the silicon chips in fixed target sample delivery. The random orientation and the limit of only one diffraction image per crystal make indexing significantly more difficult in the SX case than it is for rotation series data. While in conventional crystallography due to the crystal rotation a significant part of the reciprocal lattice is available for the indexing step, in serial crystallography only the RLPs close to the Ewald sphere in one snapshot are measured. Moreover, with the use of small crystals, the intensity of the Bragg spots is small as well. The result is that only few peaks can be detected in the peak finding stage, and the probability of false positives is higher. In total, the indexing algorithms have significantly less input, which is, to make matters worse, more noisy. Nevertheless, the indexing result, i.e. essentially the orientation of the crystal in the beam, must be very precise to ensure a good prediction of Bragg spots. The prediction is critical to ensure proper integration of the Bragg spots, which allows extending

the measured resolution through statistical averaging of spots from thousands of diffraction patterns.

The indexing algorithms from conventional crystallography have been adapted to be used with still images (Kirian et al., 2010), and new ones have been developed specifically for use in serial crystallography (e.g. Ginn et al., 2016; Li et al., 2019). They all fail in indexing a significant amount of diffraction patterns, and they all are only usable with radiation of limited wavelength[5] due to deficient 3D information of the reciprocal lattice, as very large Ewald spheres are almost flat in the region of measurable RLPs.

New and very promising applications for serial crystallography have been developed (Meents et al., 2017; Smeets et al., 2018), but the respective data processing pipelines have not been capable of dealing with thousands of images since indexing had to be done in a semi-automatic way and with poor success rates, imposing a significant bottleneck for the application of these methods.

This thesis presents two new indexing techniques: The first, named XGANDALF, significantly improves the indexing results for high throughput serial crystallography experiments with monochromatic X-ray radiation, thus improving the experimental results of past and future experiments. The other, named pinkIndexer, enables automatic indexing for emerging techniques such as serial electron crystallography (Bücker et al., 2020; Smeets et al., 2018) and pink beam serial crystallography (Meents et al., 2017), which have been lacking functional automatic indexers before. Moreover, a variant of pinkIndexer allows indexing convergent beam X-ray crystallography data, which has the potential to be used to experimentally gain information that helps to solve the phase problem.

[5]The limit is rather on a minimal diffraction angle, but due to the lack of very high frequencies in the Fourier transform of real crystals both metrics are correlated.

Chapter 3

XGANDALF - Extended Gradient Descent Algorithm for Lattice Finding

Large parts of this chapter are taken from the XGANDALF publication (Gevorkov et al., 2019).

3.1 Introduction

Serial crystallography experiments record many thousands of diffraction patterns, each from a different crystal in a random and unknown orientation. The measurements are merged to build up a complete data set. Experiments usually aim to measure not more than one crystal per diffraction pattern, although the contribution of multiple crystals in a single diffraction measurement is not uncommon. The difference in the measurement approach compared to conventional rotation crystallography has necessitated the development of new software for processing SX data, with several software packages now available including CrystFEL (White et al., 2012), DIALS (Winter et al., 2018) and nXDS (Kabsch, 2014). A key step is indexing the Bragg spots observed in a pattern, which is required to integrate and scale Bragg intensities into a common lattice and to predict the locations of other Bragg spots to be included in this merging process. Several automatic indexing algorithms have been developed and implemented in widely-used software like MOSFLM (Powell, 1999), XDS (Kabsch, 1993) (Kabsch, 2010), DirAx (Duisenberg, 1992) and LABELIT (Sauter et al., 2009). Although originally devised for rotation-series data, these algorithms are also capable of indexing snapshot diffraction patterns.

Other algorithms have been devised specifically for snapshot data (Gildea et al., 2014; Ginn et al., 2016).

In serial crystallography experiments, often crystal diffraction patterns can be observed that appear to correspond to crystal lattices but nonetheless cannot be indexed by the existing approaches. Even when several different indexing algorithms are applied to each pattern, only a fraction of the frames can be indexed. Patterns with small numbers of Bragg spots, large amounts of background noise that lead to spurious peaks in the Bragg peak detection stage, or with multiple overlapping crystal diffraction patterns ("multiple hits"), are particularly problematic and often cannot be indexed by current algorithms. In principle, it should be possible to index every diffraction pattern provided that Bragg spot locations are consistent with a true crystal diffraction pattern rather than spurious noise. It thus appears advantageous to deviate from previous approaches adapted from indexing rotation data and instead develop an algorithm for the explicit purpose of indexing SX crystal diffraction patterns. The ambition is to develop such a new and computationally-efficient algorithm for indexing SX crystal diffraction patterns with the aim of maximizing the indexing rate while being robust to outliers.

Indexing involves identifying the diffraction order of all Bragg spots measured in a diffraction pattern, equivalent to determining the crystal orientation. In most indexing algorithms, the process begins by mapping the positions of Bragg spots found on the detector to radiation scattering momentum transfer vectors $\boldsymbol{n}$ in the three-dimensional (3D) reciprocal space using prior information about the detector geometry (including sample to detector distance) and the wavelength of the incident beam. The resulting points in 3D reciprocal space approximate the points of the reciprocal lattice, which is initially unknown. We call these points "nodes" to abstract the problem from crystallographic indexing to the more general problem of fitting a lattice to noisy locations. One possible approach to indexing is to detect maxima in the Fourier transform of the pattern of nodes (Steller et al., 1997). Such maxima mark the directions with maximum periodic repetition, which can form the basis vectors of the wanted lattice. This approach is taken in DIALS (Gildea et al., 2014) and MOSFLM (Powell, 1999). The related DirAx algorithm finds principal repeat directions by searching for frequently-occurring repeats perpendicular to triplets of nodes (Duisenberg, 1992). Another popular indexing approach is to search for frequently-occurring difference vectors between the nodes, as is done in XDS (Kabsch, 2010) and TakeTwo (Ginn et al., 2016). The FELIX indexer (Beyerlein et al., 2017) uses a different approach which is to map the set of possible crystal orientations that are consistent with particular Bragg spots to lines in Rodrigues-Frank space to find a consensus orientation for all peaks.

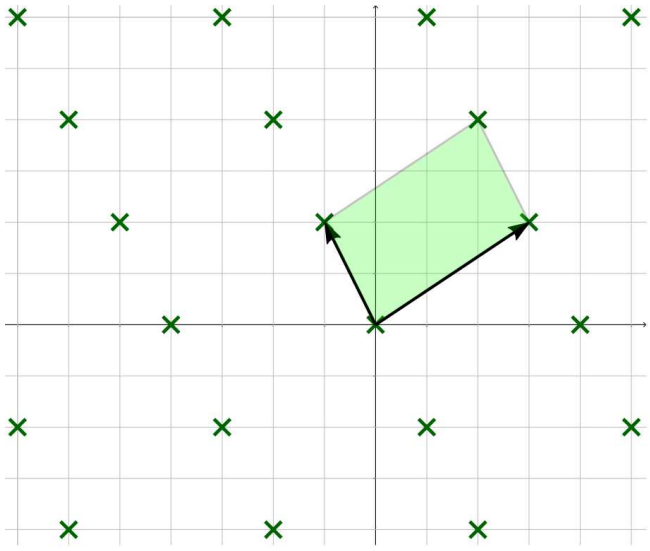


Figure 3.1: *A crystal lattice is defined by its basis. The lattice basis partitions the space into identical parts. The shape of a part equals the shape of the parallelepiped spanned by the basis vectors - the unit-cell. This figure shows a two dimensional lattice with basis vectors sketched as black arrows and the unit-cell spanned by the basis vectors sketched as a light-green area.*

The here developed algorithm can be considered as a modified version of the Fourier methods. To improve noise tolerance, the Fourier transform is replaced by a similar transform that uses periodic basis functions combined with a non-linear weighting scheme. To achieve fast execution, the algorithm employs a multi-step heuristic, i.e. an approximate but efficient method, based on an extended gradient approach to identify maxima in the transformed pattern of nodes corresponding to points on the real-space crystal lattice. The real-space lattice basis is then formed from these maxima while maximizing the number of the observed nodes that are consistent with that basis choice and minimizing the distances between those nodes and their closest lattice point. An overview of the main steps of the proposed algorithm is provided in figure 3.2.

If knowledge about the crystal lattice is available, it can be be exploited by passing unit cell parameters to the algorithm. A unit cell is the volume spanned by the lattice basis vectors. Unit cell parameters include the lengths of the basis vectors and their pairwise angles. A unit cell is sketched in figure 3.1.

We call the new algorithm XGANDALF - Extended Gradient Descent Algorithm for Lattice Finding.

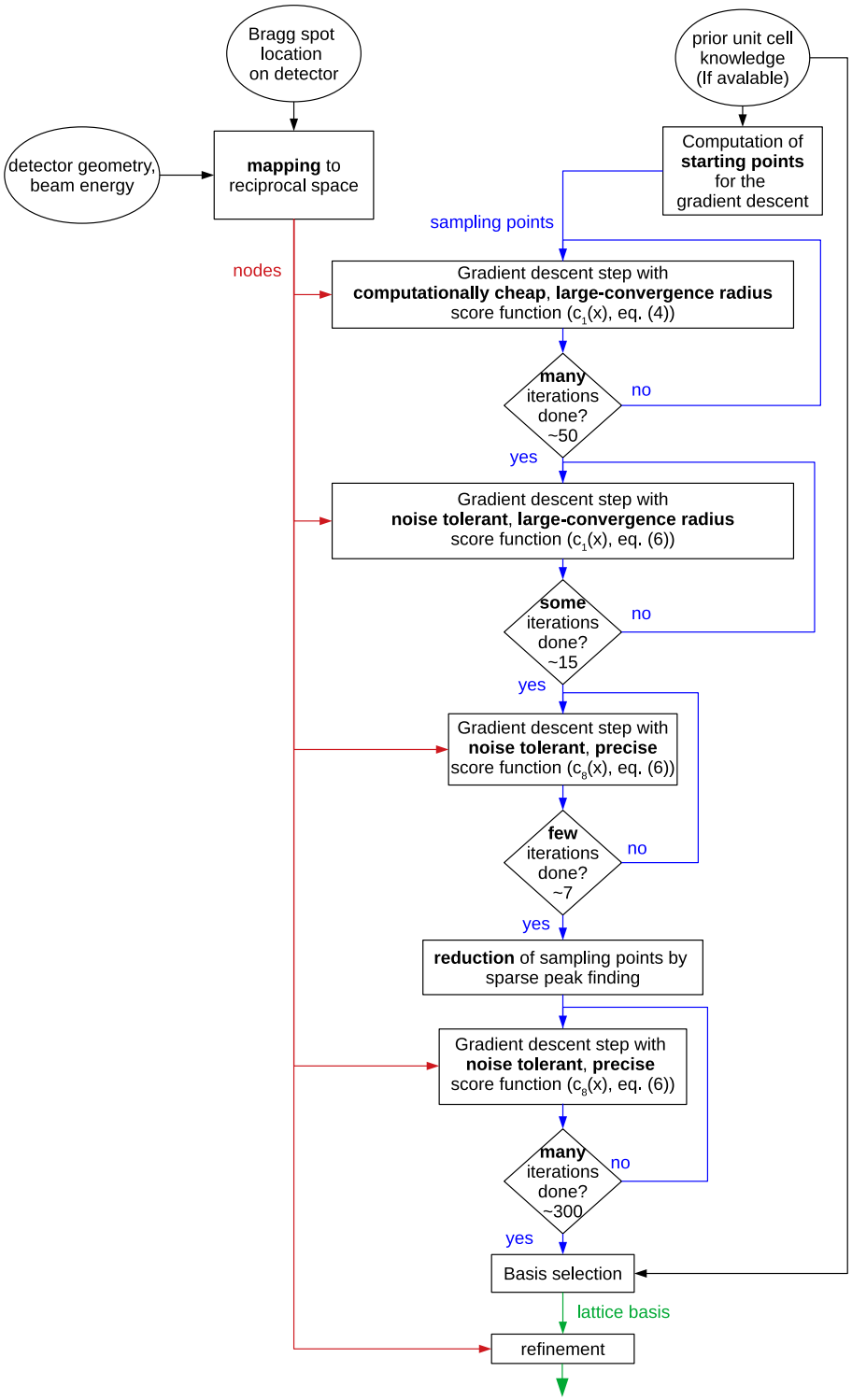


Figure 3.2: *Overall structure of the XGANDALF algorithm.*

3.2 Algorithm description

3.2.1 Overview

The indexing algorithm determines the Miller indices of the observed Bragg peaks in a still snapshot diffraction pattern, given knowledge of the experiment geometry and optionally the unit cell parameters of the crystal. It consists of the following key steps:

1. Bragg spot locations on the two-dimensional detector are mapped to the surface of the Ewald sphere in three-dimensional reciprocal space. We call these locations "nodes" to distinguish them from the exact reciprocal lattice points, the locations of which are initially unknown.

2. Each node in the three-dimensional reciprocal space is used to define a set of equidistant parallel planes in the three-dimensional real space. A sketch that simplifies this relation to two dimensions is shown in figure 3.3. Intersections of parallel planes generated from different nodes are solutions to the indexing problem. The rest of the algorithm is devoted to finding these intersections in the presence of noise, spurious peaks, and multiple lattices.

3. Continuous "proximity functions" based on the distance to each node's parallel planes are defined and summed to create a score function to find the points of intersection. Intersections of planes become maxima of the score function, with the continuous score function serving to suppress the effect of experiment noise and inaccuracies. A series of progressively sharper and steeper proximity functions are used with the result that spurious nodes corresponding to falsely-identified Bragg spots belonging to competing lattices are removed from the set of Bragg spots that are used to generate the nodes.

4. A heuristic (a fast technique for finding approximate solutions) is used to find maxima of the score function. Sharper proximity functions require more computations to find maxima, hence the more computationally-efficient proximity functions are chosen to reduce the search space early in the computation. An extended gradient descent method is applied to migrate the starting points to the maxima of the score function and avoid otherwise slow convergence due to zigzag optimization trajectories.

5. The bases of the found lattices provide the indexing solutions once the maxima of the score function have been found. A refinement step is then

performed to minimize the mean Euclidean distance between the observed and predicted nodes using a gradient descent approach.

3.2.2 Relation between nodes and the indexing solution

For a diffraction pattern, with K measured Bragg spots, the relation between the K nodes $\boldsymbol{n}_k \in \mathbb{R}^3$, the reciprocal lattice basis $\boldsymbol{B}^* \in \mathbb{R}^{3\times3}$ and the K sets of Miller indices $\boldsymbol{m}_k$, can be described by:

$$\begin{aligned}&(\boldsymbol{n}_1, \boldsymbol{n}_2, ..., \boldsymbol{n}_K) = \boldsymbol{B}^*(\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_K),\ \forall k : \boldsymbol{m}_k \in \mathbb{Z}^3 \\ \Leftrightarrow &\boldsymbol{N} = \boldsymbol{B}^*\boldsymbol{M},\ \boldsymbol{M} \in \mathbb{Z}^{3\times K}\end{aligned} \tag{3.1}$$

The nodes are the known observables. The reciprocal basis, as well as the Miller indices, need to be identified. Multiplying from the left with $\boldsymbol{B}^{*-1}$ and applying the relation between real and reciprocal basis from equation 2.8 leads to:

$$\begin{aligned}&\boldsymbol{B}^{*-1}(\boldsymbol{n}_1, \boldsymbol{n}_2, ..., \boldsymbol{n}_K) = (\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_K),\ \forall k : \boldsymbol{m}_k \in \mathbb{Z}^3 \\ \Leftrightarrow &\boldsymbol{B}^{*-1}\boldsymbol{N} = \boldsymbol{M},\ \boldsymbol{M} \in \mathbb{Z}^{3\times K} \\ \Leftrightarrow &\boldsymbol{B}^T\boldsymbol{N} = \boldsymbol{M},\ \boldsymbol{M} \in \mathbb{Z}^{3\times K}\end{aligned} \tag{3.2}$$

Using the following nomenclature for matrix indices:

$$\boldsymbol{B} = \begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{pmatrix} \tag{3.3}$$

And applying the rules for matrix multiplication, equation (3.2) can be rewritten to a system of linear equations:

$$\begin{aligned}b_{1,j} \cdot n_{1,k} + b_{2,j} \cdot n_{2,k} + b_{3,j} \cdot n_{3,k} = m_{j,k}, \\ \forall (j,k) : k \in [1, 2, ..., K], j \in \{1, 2, 3\}, m_{j,k} \in \mathbb{Z}\end{aligned} \tag{3.4}$$

The problem of finding a suitable lattice basis can be greatly simplified. Instead of solving the system of equations (3.4) for all three real space basis vectors, it is sufficient to find three linearly independent solutions to the simplified system of equations for just one real space basis vector:

$$\begin{aligned}\widetilde{b}_1 \cdot n_{1,k} + \widetilde{b}_2 \cdot n_{2,k} + \widetilde{b}_3 \cdot n_{3,k} = \widetilde{m}_k,\ \forall k : k \in [1, 2, ..., K], \widetilde{m}_k \in \mathbb{Z} \\ \Leftrightarrow \boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}} = \widetilde{m}_k,\ \forall k : k \in [1, 2, ..., K], \widetilde{m}_k \in \mathbb{Z} \\ \Leftrightarrow \boldsymbol{N}^T\widetilde{\boldsymbol{b}} = \widetilde{\boldsymbol{m}},\ \widetilde{\boldsymbol{m}} \in \mathbb{Z}^K,\ \boldsymbol{N} \in \mathbb{R}^{3\times K}\ \widetilde{\boldsymbol{b}} \in \mathbb{R}^3,\end{aligned} \tag{3.5}$$

Note that for more than three nodes this is an over-determined system of linear equations. Nevertheless, since the Miller indices $\widetilde{\boldsymbol{m}} \in \mathbb{Z}^K$ are not known, this problem cannot be solved by linear regression.

Each node $\boldsymbol{n}_k$ forms through equation (3.5) a series of equidistant parallel planes in the real space distinguished and enumerated by the Miller indices m_k. These are the planes of the real-space lattice of the crystal associated with the node. Any point on any of the planes is a solution to the equation formed by this node. The planes are orthogonal to $\boldsymbol{n}_k$ and their spacing is given by $1/\|\boldsymbol{n}_k\|$. Different nodes form different series' of real-space planes; their intersections correspond to the real-space lattice, which are thus the points that solve the equations 3.5. In an equivalent two-dimensional (2D) model, every node would form a series of equidistant parallel lines, as depicted in figure 3.3.

To solve equation (3.5), three linearly independent vectors must be found, where each vector points to one of the planes of each and every node (that is, to their intersections). Under real conditions, there are usually more than three nodes, making the problem over-determined. However, due to noise, the planes corresponding to these nodes will generally not all intersect at common points in real-space (see red node in figure 3.3), so an exact solution usually will not exist. The optimal solution $\widetilde{\boldsymbol{b}}$ for equation (3.5) is, therefore, one that minimizes the average distance to one of the planes of each node. To find this solution a score function is introduced, it is defined as a sum of proximity functions which themselves encapsulate the distance of the assumed solution from these geometrical planes.

3.2.3 Continuous proximity function for noise tolerance

Every node defines a series of real-space parallel planes according to equation (3.5) and as sketched in figure 3.3 for the 2D case. Since the nodes are assumed to be noisy, the locations of the parallel planes also must be assumed to be noisy. This implies that the best estimate of the lattice basis vectors $\widetilde{\boldsymbol{b}}$ (the optimal solution) might not exactly lie on the planes, but may instead lie close to the planes. To account for this effect, a real-space proximity function c is defined, which indicates how close a real-space vector is to a plane. This function is chosen to equal its maximum value at points on the planes, and is equal to its minimum value at points equidistant between two planes. A score function is then constructed for the entire arrangement of nodes as a normalized weighted sum of proximity functions given by

$$\mathrm{f}(\widetilde{\boldsymbol{b}}) = \frac{1}{K} \sum_{k=1}^{K} w_k \, \mathrm{c}(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}}). \tag{3.6}$$

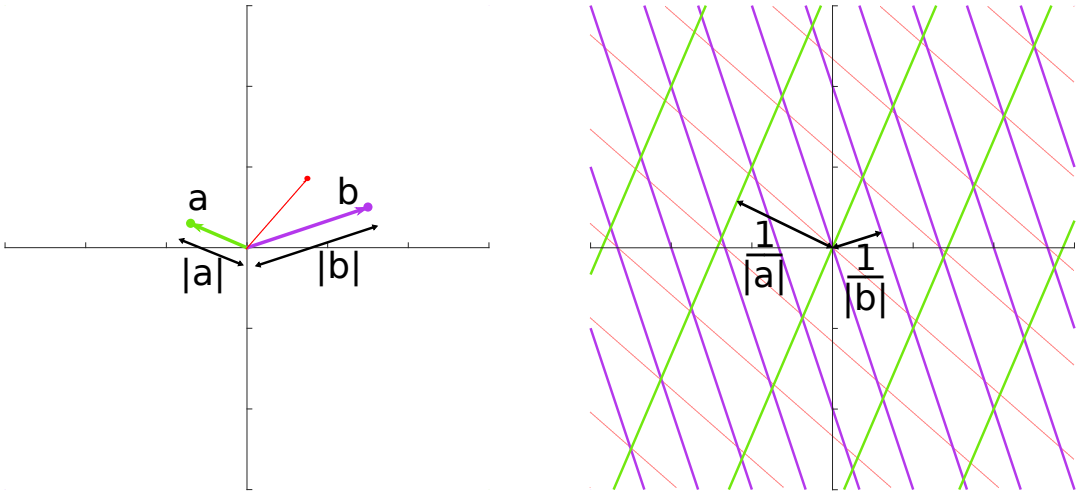


Figure 3.3: *Line series in real space (green and purple, right panel) generated by two nodes a and b in the 2D reciprocal space (left). The distance between adjacent parallel lines is given by the reciprocal magnitude of the respective node. A third node ($a + b +$ noise, red) is shown, along with a corresponding set of lines in the right panel, to show that in the presence of noise there usually are no points where all sets of lines intersect.*

The weighting w_k can depend on the intensities of the nodes, their norms, or other properties. The maxima of this score function are the feasible solutions of equation (3.5), corresponding to real-space lattice points. From these, three linearly-independent vectors are obtained that describe that lattice.

An example for the 2D case can be seen in figures 3.4 and 3.5. Figure 3.4 shows the interpolation of the lines in figure 3.3 using proximity functions that vary linearly from their minimum to maximum values. Figure 3.5 shows the score function of a sample arrangement of 13 nodes with this same choice of linear proximity functions.

As mentioned above, the proximity function indicates the distance from sets of parallel planes of equal spacing. While it is defined in 3D real space, it is a function only of distance along lines orthogonal to those planes. It is reasonable to define the proximity function to equal 1 on the planes, -1 midway between two planes, and to vary monotonically between these values. Combining these considerations, a proximity function of the following form is reasonable

$$\mathrm{c}(\boldsymbol{n} \cdot \widetilde{\boldsymbol{b}}) = \begin{cases} 1, & \text{if } \boldsymbol{n} \cdot \widetilde{\boldsymbol{b}} = m, \quad m \in \mathbb{Z} \\ -1, & \text{if } \boldsymbol{n} \cdot \widetilde{\boldsymbol{b}} = m - 0.5, \; m \in \mathbb{Z} \\ \text{monotonic} & \text{in between.} \end{cases} \tag{3.7}$$

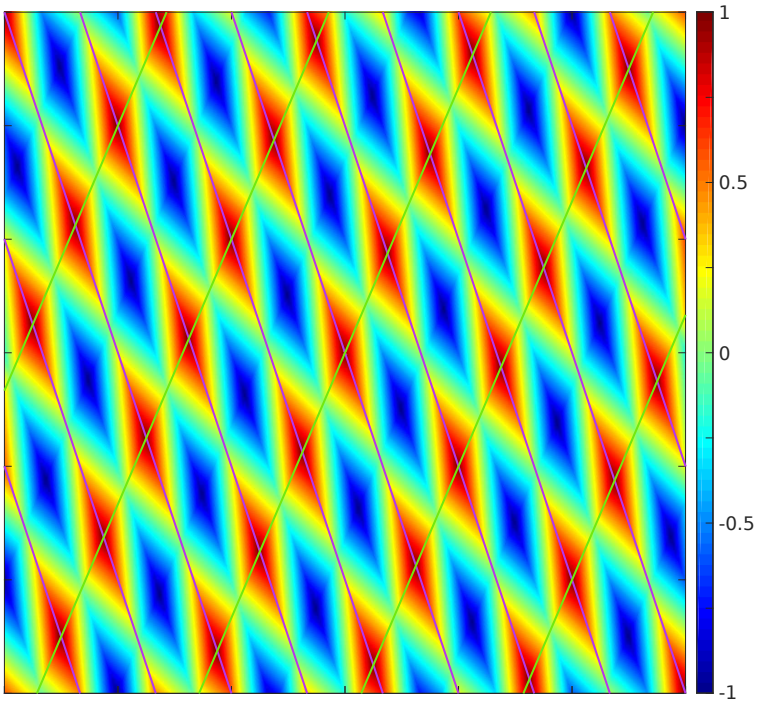


Figure 3.4: *Score function for the lines from vector a and b in figure 3.3. For each node, positions on the lines are assigned a proximity of 1 and positions in the middle between two lines are assigned the proximity −1. The rest of the proximity function for a single node is a linear interpolation of these values (as sketched in the top left corner of figure 3.3). The score function is formed by the normalized sum of the proximity functions of each node.*

Many different functions are suitable for use as proximity functions. The execution time and thus the complexity of the function evaluation must be considered in its selection. The proximity function is periodic with a period of 1, so it can be defined in the interval $[-0.5,\ 0.5]$ with $\mathrm{c}(x) \coloneqq \mathrm{c}(x - \mathrm{round}(x))$. The following proximity functions are available in a tool-kit for further exploration and development of the program.

- $\mathrm{c}_1(x) = \cos^s(2\pi x),\ s \in \{2n-1 \mid n \in \mathbb{N}^+\}$
- $\mathrm{c}_2(x) = 1 - 4\,|x|$
- $\mathrm{c}_3(x) = 32x^4 - 16x^2 + 1$
- $\mathrm{c}_4(x) = -8x^2 + 1$
- $\mathrm{c}_5(x) = -32x^4 + 1$

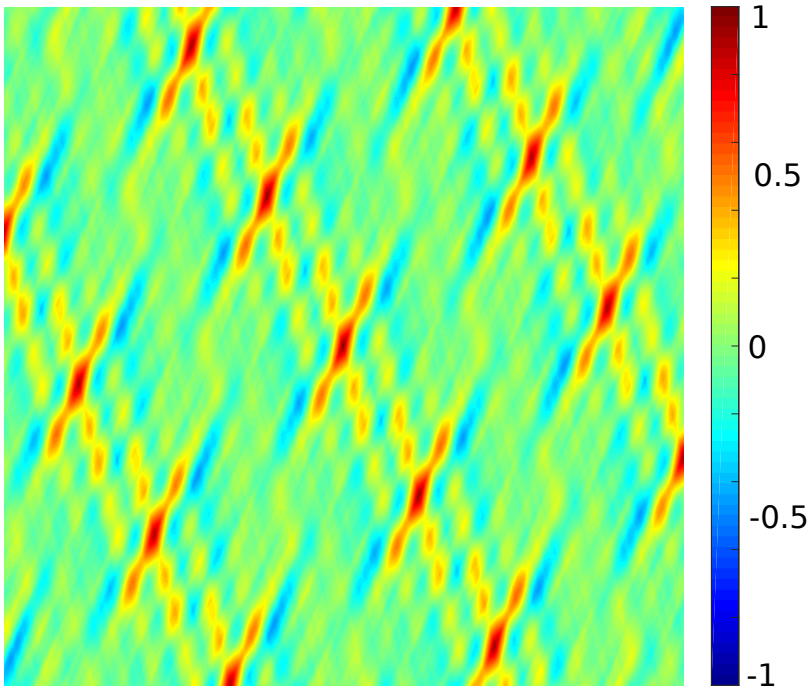


Figure 3.5: *Score function for a set of 13 nodes that were generated by adding noise to the position of randomly chosen points on a lattice grid. For each node, positions on the lines are assigned a proximity of 1 and positions in the middle between two lines are assigned the proximity -1. The rest of the proximity function for a single node is a linear interpolation of these values. The score function is formed by the sum of the proximity functions of each node.*

- $c_6(x) = -\mathrm{sgn}(|x| - w/2),\ w \in (0,\ 1)$
- $c_7(x) = 8(|x| - 0.5)^2 - 1$
- $c_8(x) = 2(1 - 2\,|x|)^s - 1,\ s \geq 1$

The proximity functions are visualized in figure 3.6. In the implemented heuristic only c_1 and c_8 are used, as they allow a good trade-off between execution speed and fitting accuracy.

It can be noted that using $\cos(2\pi x)$ as a proximity function turns the score function (equation (3.6)) into the real part of the Fourier transform. Assuming that the geometry of the experiment is accurately known, the locations of the nodes in reciprocal space are centrosymmetric and so, if symmetrized, the Fourier transform of the arrangement of nodes would indeed be real. That is, the score function using $c_1(x)$ with $s = 1$ produces the Fourier transform of the given arrangement of nodes. Such a score function is used in the Fourier indexing

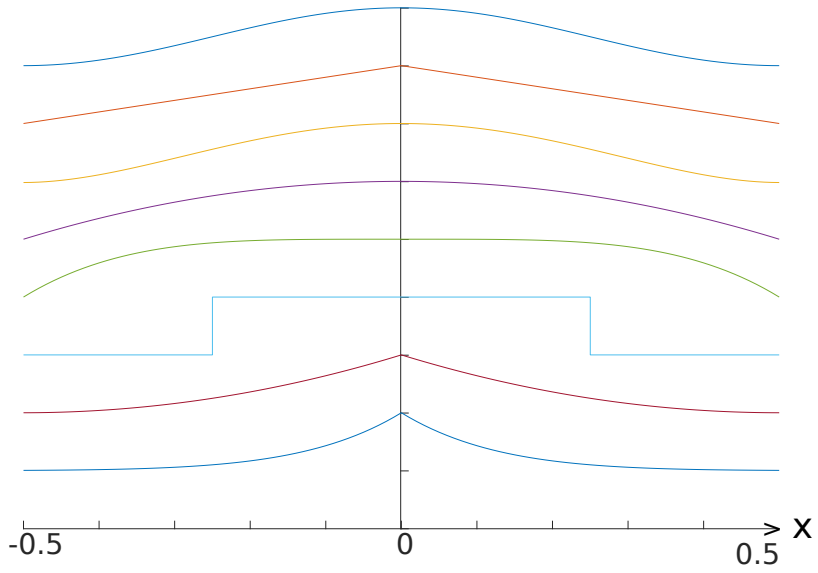


Figure 3.6: *Plotted proximity functions c_1 (top) to c_8 (bottom).*

methods, where lattice vectors are found by searching for maxima in the Fourier transform of a given arrangement of nodes. The new approach generalizes the use of a Fourier transform to that of an arbitrary proximity function. This extension provides a means to tune the proximity function to either achieve a greater noise tolerance (with a narrowly-peaked function) or a larger convergence radius for the search (with a broad function).

Not every Bragg spot found belongs to the same lattice. There may be false positives in the peak finding algorithm or peaks from different crystals in the same diffraction pattern. Such spurious peaks should ideally have as little impact as possible on the maxima of the score function. Their contribution can be removed by introducing a tolerance parameter $\varepsilon > 0$. Nodes that generate planes that are too far away from the inspected vector are excluded from the computation of the score function. This distance of inclusion is given by ε, so the smaller ε, the more resistant the score function is to spurious peaks. The drawback of this method is that the score function can be discontinuous. The resulting score function is given by

$$\mathrm{f}(\widetilde{\boldsymbol{b}}) = \frac{\sum_{k=1}^{N} \delta(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}})\, w_k\, \mathrm{c}(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}})}{\sum_{k=1}^{N} \delta(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}})}, \tag{3.8}$$

with

$$\delta(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}}) = \begin{cases} 1 \text{ if } \left| \boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}} - \mathrm{round}(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}}) \right| \leq \varepsilon \\ 0 \text{ if } \left| \boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}} - \mathrm{round}(\boldsymbol{n}_k \cdot \widetilde{\boldsymbol{b}}) \right| > \varepsilon. \end{cases} \tag{3.9}$$

The solution to the indexing problem requires finding maxima of the score function. This is done by a local search in the 3D real space of the $\widetilde{\boldsymbol{b}}$ vectors, which is supposed to be carried out efficiently to reduce computational time. The search must be started from a diverse number of starting points to ensure that more than one maximum is found. However, the search needs only be conducted within a volume of the real space which can feasibly contain the real-space lattice vectors of the crystal. If the lattice parameters are not known in advance, then this volume can be restricted to a shell centered on the origin ranging in radius from the minimum to maximum possible lattice vector magnitudes, given reasonable assumptions. If the lattice parameters are known, then this search volume can be restricted considerably further, to spherical shells, each with a mean radius given by each of the real-space lattice parameters, as done by Gildea *et al.* 2014. The width of the shells is set to a tolerance that is dependent on the uncertainty of the lattice parameters.

The search is started simultaneously from a large number of evenly-spaced points within the search volume that later migrate to the maxima of $\mathrm{f}(\widetilde{\boldsymbol{b}})$ by a gradient descent approach. A typical number of starting points is 50 000. A set of starting points that are approximately uniformly separated and distributed throughout the volume of the spherical shell is formed by first obtaining positions of points on a spherical surface that are approximately equally spaced from each other. This is done by minimizing a generalized electrostatic potential energy of a system of charged particles (Semechko, 2015). Since such computations can take a very long time, a set of pre-computed distributions of points on the unit sphere is used. This distribution is then scaled to several spherical surfaces that span the desired search shell. The radial increment of neighboring surfaces is chosen to equal the average distance of neighboring points on the sphere. To avoid systematic alignment of the points on each sphere, each point set is randomly rotated about the origin.

While in theory it is sufficient to find the primary lattice vectors (i.e. the vectors of the reduced real-space lattice basis with Miller indices 100, 010, and 001), in the presence of spurious Bragg spots or multiple lattices it is often beneficial to also search for the lattice vectors with Miller indices 110, 011 or 101. This is because spurious Bragg spots or spots from other lattices can significantly diminish some peaks in the score function. The use of additional lattice vectors adds redundancy and allows handling cases where the peaks in the score function belonging to the primary lattice basis vectors of a lattice are not detected. This procedure increases the execution time but improves the success rate of the algorithm, and is therefore provided as an option in the implementation of the algorithm.

Using even proximity functions $\mathrm{c}(\boldsymbol{n} \cdot \widetilde{\boldsymbol{b}})$, the score function $\mathrm{f}(\widetilde{\boldsymbol{b}})$ is centrosym-

metric about the origin. The algorithm exploits this symmetry which allows the use of only half of the starting points.

3.2.4 Gradient descent extension

The algorithm employs an extended gradient descent method to let the starting points migrate to the maxima of $f(\widetilde{\boldsymbol{b}})$[1]. Empirical analysis shows that, for typical score functions $f(\widetilde{\boldsymbol{b}})$, the gradient is often large at locations close to the maximum (see figure 3.5). The ordinary gradient descent method uses large step sizes for large gradients, which here is counterproductive. Instead, a step size is generated using a combination of the previous step length, the change in step direction, the value of $f(\widetilde{\boldsymbol{b}})$, the number of well fitted nodes, a parameter γ (as it is used in the ordinary gradient descent to regulate the relative step length), and clipping to a minimum and maximum step size. The parameters for the choice of the step size are empirically optimized and are not visible to the user.

As with the ordinary gradient descent algorithm, the problem occurs that convergence is often severely slowed down by zigzag migration trajectories (Wang, 2008). A common approach to overcome this problem is to use the conjugate gradient method (Hestenes et al., 1952). Given the known composition of the score function, a different method is used instead. For every step, it is checked whether the direction of the current step is nearly opposite that of the previous step. If this is the case, a zigzag path is probable and the current step direction is replaced by the sum of the unit length vector pointing in the current direction and the unit length vector pointing in the previous direction. This takes the search in a direction almost orthogonal to the previous ones, helping to overcome zigzag paths while being computationally very cheap.

3.2.5 Heuristic algorithm for locating maxima in the score function

The goal of the heuristic is to find peaks in the score function, and hence probable lattice vectors, quickly and precisely. A large radius of convergence is required, but at the same time, a very precise detection of the maxima is important. We, therefore, use a custom, empirically-tuned algorithm with a multi-step design to home in on the maxima in stages. In this method, the earlier stages use smoother proximity functions, whereas in the later stages one with a sharper peak is used to achieve a more precise determination of the maxima:

[1] Strictly, the negative of the score function is minimized by gradient *descent*.

1. Gradient descent: proximity function $c_1(x)$ with $s = 1$, score function from equation (3.6), inverse radial weighting. (See figure 3.7 (d) for visualization of the score function.)

   The first stage is responsible for bringing the sampling positions close to the peak maximum without getting stuck in the local maxima. This is accomplished by using the Fourier-transform proximity function $c_1(x)$ in conjunction with a weighting of the nodes proportional to $1/\left\|\boldsymbol{q_k}\right\|^2$. The radius dependent weighting ensures a smooth score function by reducing weights of short period proximity functions from high-resolution Bragg peaks, that otherwise would cause many local maxima. This stage is the most computationally expensive one since it contains many gradient descent steps and operates on a large number of sampling points to ensure capturing the peak within the radius of convergence.

2. Gradient descent: proximity function $c_1(x)$, score function from equation (3.8). (See figure 3.7 (e) for visualization of the score function.)

   This and all subsequent stages bring the sampling points closer to their corresponding local maximum. These stages use the noise-tolerant and computationally expensive score function from equation (3.8), and unity weighting.

3. Gradient descent: proximity function $c_8(x)$ with $s = 4$, score function from equation (3.8), few steps. (See figure 3.7 (f) for visualization of the score function.)

   The third stage uses the very local and computationally-expensive proximity function $c_8(x)$ with $s = 4$. Using finer gradient descent steps, it is responsible to bring the sampling points close enough to the maxima to be able to identify even very sharp maxima by the score function evaluation at these sampling points.

4. Sparse peak finding on the sampling points.

   Only the 50 sampling points with the highest score function evaluation in their respective local environments are kept. This drastically reduces the number of sampling points.

5. Gradient descent: proximity function $c_8(x)$ with $s = 4$, score function from equation (3.8), many steps. (See figure 3.7 (f) for visualization of the score function.)

   This last stage uses many fine gradient descent steps with the local and computationally-expensive proximity function $c_8(x)$ with $s = 4$, and with

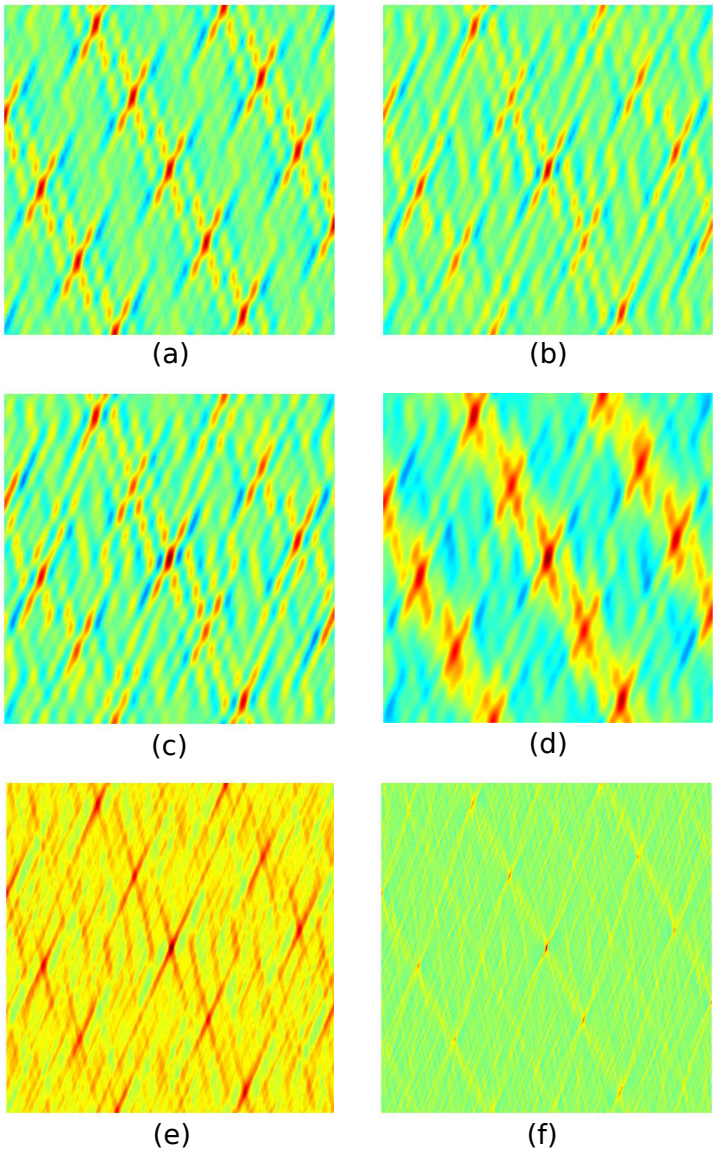


Figure 3.7: *Score functions for a set of 13 simulated nodes that were generated by adding noise to the position of randomly chosen points on a lattice grid. For images (b) - (f) additional seven spurious nodes were added, i.e. nodes not lying on the lattice. (a) Proximity function* c_2*, no spurious nodes. (b) Proximity function* c_2*. (c) Proximity function* c_1*. (d) Proximity function* c_1 *with inverse radial weighting. (e) Proximity function* c_1 *with score function from equation* (3.8)*. (f) Proximity function* c_8 *with* $s = 4$ *with score function from equation* (3.8)*.*

the score function of equation (3.8). This ensures that the sampling points migrate extremely close to the maxima, yet maintains an affordable computational effort due to the small number of sampling points used in this stage.

The number of steps for each stage can be chosen by a flag to the program.

A visualization of the employed score functions can be seen in Fig. 3.7, which shows score functions for a set of 13 simulated nodes that were generated by adding noise to the position of randomly chosen points on a lattice grid. For images (b)–(f) additional seven spurious nodes were added, i.e. nodes not lying on the lattice. Despite the noise in the positions of the nodes, image (a) shows a high degree of periodicity. The additional seven spurious peaks significantly diminish some of the maxima in image (b). Image (c) shows a slightly better contrast than image (b) at the expense of a more computationally-expensive proximity function. Image (d) has a high radius of convergence for the gradient descent approach but does not provide exact peak locations. Case (e) provides more accurate peak locations but has a small convergence radius for the gradient descent approach. Case (f) uses a computationally-expensive proximity function and suffers from a small convergence radius, but provides better noise suppression capabilities and accurate peak locations.

3.2.6 Selection of lattice bases

Once the maxima of the score function have been found, the bases of the found lattices can be formed. As a first step, all possible lattice bases are selected that each correctly predicts at least five nodes. In theory, five nodes (given that they span the $\mathbb{R}^3$) are more than what is minimally required to define a single lattice, but this increases the noise tolerance. The selection of candidate bases is computationally expensive since there are $\binom{K}{3}$ basis choices for K found peaks in the score function. To reduce computation time, a multi-step filter is employed that in steps with increasing complexity filters out non-sensible solutions. As a first step all those vectors that predict less than five nodes are excluded. The next steps check for a reasonably high determinant[2] of the basis and the number of correctly predicted nodes using two vectors and afterwards using three vectors. If the lattice parameters are known, the candidate lattices not fitting to these parameters are excluded as well. Finally, the basis vectors are sorted by the sum of each vector's score function and the best 500 are kept.

Each kept candidate basis is reduced (to find the shortest possible basis vectors, see figure 3.8) using an efficient algorithm described by Semaev 2001. Afterwards, for each reduced basis, the absolute defect (mean distance between

[2]The determinant of the basis is the volume of the unit cell.

the nodes and their positions predicted by the basis) and the relative defect (mean difference between the Miller indices of the nodes and the fractional Miller indices of the predicted nodes) is computed. From the 500 candidate lattices, 15 with the largest score function evaluation and 50 with the smallest relative defects are kept for the final stage.

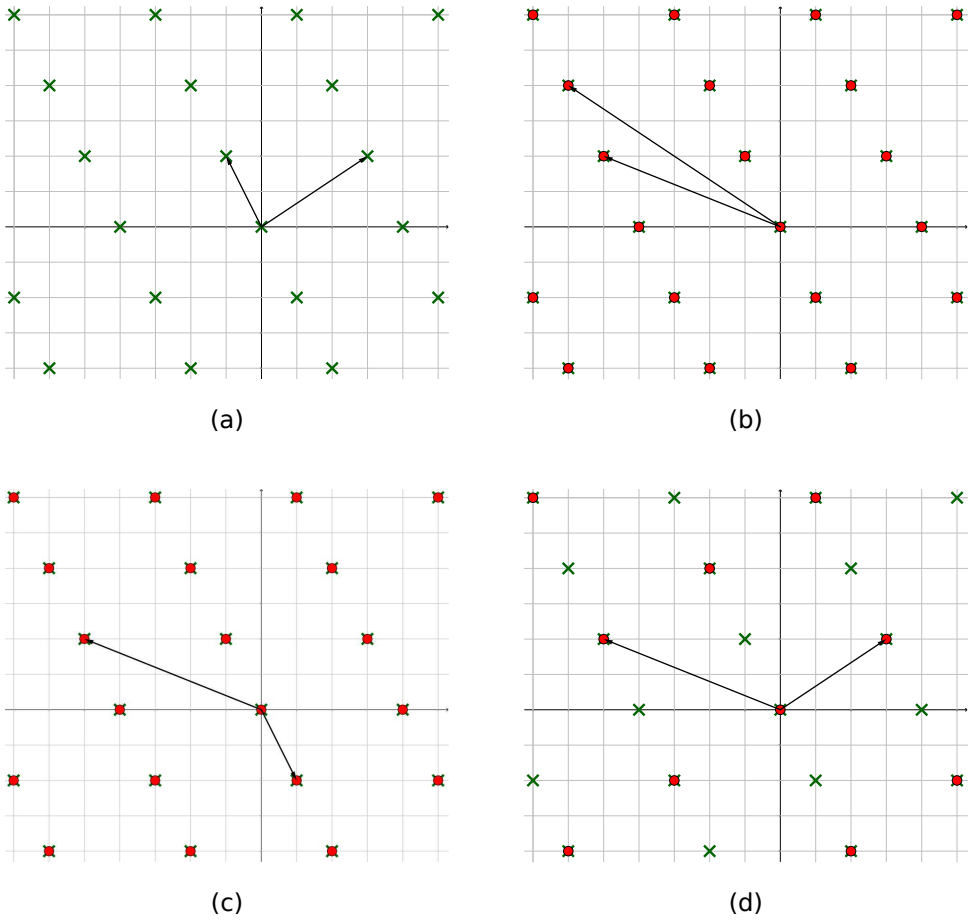


Figure 3.8: *(a) shows a lattice marked by its basis vectors and lattice points (green crosses). The basis vectors are chosen to be the shortest possible of all linear possible basis vectors, i.e. the basis is reduced. (b) and (c) show the same lattice (red dots), but with a different (not reduced) basis. The determinant of the basis (i.e. the volume of the unit-cell) is equal for (a),(b), and(c). (d) shows a lattice that is a subset of the lattices in (a)-(c). The determinant of the basis is a multiple of the one in (a)-(c), the unit-cell is a super-cell with respect to the cell in (a)-(c).*

In the final stage, the bases which best predict the nodes are selected. For this, the candidate bases are sorted in descending order by the number of nodes

they correctly predict. Starting with the basis predicting the most nodes, it is considered as generating a true lattice if it either predicts at least five points that were not predicted by any other basis or if it has significantly smaller defects than a previously accepted basis. In the latter case, the newly found basis replaces the previous one. To avoid super-cells in cases with unknown lattice parameters, bases with smaller determinants are preferred.

XGANDALF thus can detect several lattices in a diffraction pattern in one pass. This allows fast data processing despite the presence of several lattices in the pattern. If processing time is not of concern, employing the delete-and-retry technique (i.e. detect the strongest lattice in a pattern, delete the corresponding peaks and retry the indexing) can lead to better results. However, only this latter method is implemented in the interface to CrystFEL 0.8.0 (White et al., 2012).

3.2.7 Refinement

After the identification of the bases, a refinement step is performed. The lattice bases are refined to minimize the mean Euclidean distance between the nodes and the predicted nodes using a gradient descent approach. Only the nodes close to the predicted nodes are used for refinement to improve noise tolerance.

3.3 Evaluation of the algorithm

3.3.1 Indexing rate

Indexing solutions of measured diffraction patterns are often tested for correctness by comparing the locations of Bragg spots predicted by the lattice basis to those of the observed spots. If the pattern contains a large number of Bragg spots (say, 50) then this test usually yields a reliable estimate of correctness. If on the other hand, the number of found spots is small then there can be several incorrect orientations of a crystal that predict the found spots, often giving a false indication of correctness. A reliable evaluation of the algorithm to index patterns as a function of the number of Bragg spots, therefore, requires ground truths, but ones that are as close as possible to real data. The ground truths are generated from a set of diffraction patterns which all had large numbers of Bragg spots, and as such were reliably indexed using MOSFLM (giving more than 50 correctly-predicted peaks). The patterns were chosen from serial femtosecond crystallography data of crystals from serotonin receptor $5-\mathrm{HT}_{2\mathrm{B}}$ bound to ergotamine, publicly available from the CXIDB (Maia, 2012) entry 21 (Liu et al., 2013). Patterns with fewer spots were created by simply deleting spots from

these previously-indexed patterns. This way realistic patterns with five to 50 spots are obtained, all with known crystal orientation. Two sets of patterns are created: One with the spots randomly distributed throughout the pattern, and the other with low-resolution spots only, generated by removing Bragg spots from the original patterns at high scattering angles.

To compare the indexing algorithm with others, the patterns from the two datasets were indexed using the indexamajig program from CrystFEL (White et al., 2012). The use of CrystFEL allows a fair comparison of several (although not all) indexing algorithms with limited effort. The employed indexers are MOSFLM (Powell, 1999), XDS (Kabsch, 1993; Kabsch, 2010), DirAx (Duisenberg, 1992), TakeTwo (Ginn et al., 2016), and two different modes of XGANDALF. One of these modes implemented many starting points and many gradient descent steps while the other mode used fewer starting points and fewer gradient descent steps. These are labelled "XGANDALF_precise" and "XGANDALF_fast" in figure 3.9. In all cases, the lattice parameters were specified to the indexing algorithm. No additional tuning of the indexing algorithms was performed. The indexing results were compared with the ground truths obtained from the original indexing of the patterns with MOSFLM. This comparison was accomplished by applying the Kabsch algorithm (Kabsch, 1976) to compute the angle needed to rotate one lattice basis onto another. Indexing solutions that required rotations of no more than 3° to bring them into coincidence with the ground-truth solution of MOSFLM (prior to removing spots) were counted as correct. For this test, all CrystFEL optimizations were turned off by using the options `-no-retry -no-refine -no-check-cell`. Only one indexing solution per pattern was accepted (using the option `-no-multi`). Although 3° is a significant deviation, this value is usually good enough for the subsequent refinement. For patterns with few peaks and a significant amount of noise, large deviations are anyway unavoidable. The results of the comparison are displayed in figure 3.9.

The most practical test of indexing is the quality of the final merged data, as detailed in Sec. 3.3.3. Before that data can be merged, the full dataset must be indexed. For comparison, diffraction of beta-lactamase crystals from CXIDB ID 83 (Wiedorn et al., 2018) is used. This dataset consists of a total of 14 445 patterns identified as containing crystal diffraction, which were indexed by a variety of algorithms—the results are summarised in Tables 3.1 and 3.2. No additional tuning of the indexing algorithms was performed. Most patterns contained multiple hits, resulting in a total number of indexed crystals that for many indexers was higher than the number of patterns. Based on the experiment setup, the quality of the prediction and the quality of the merge results, it is most likely that these patterns really arose from multiple crystals. Although

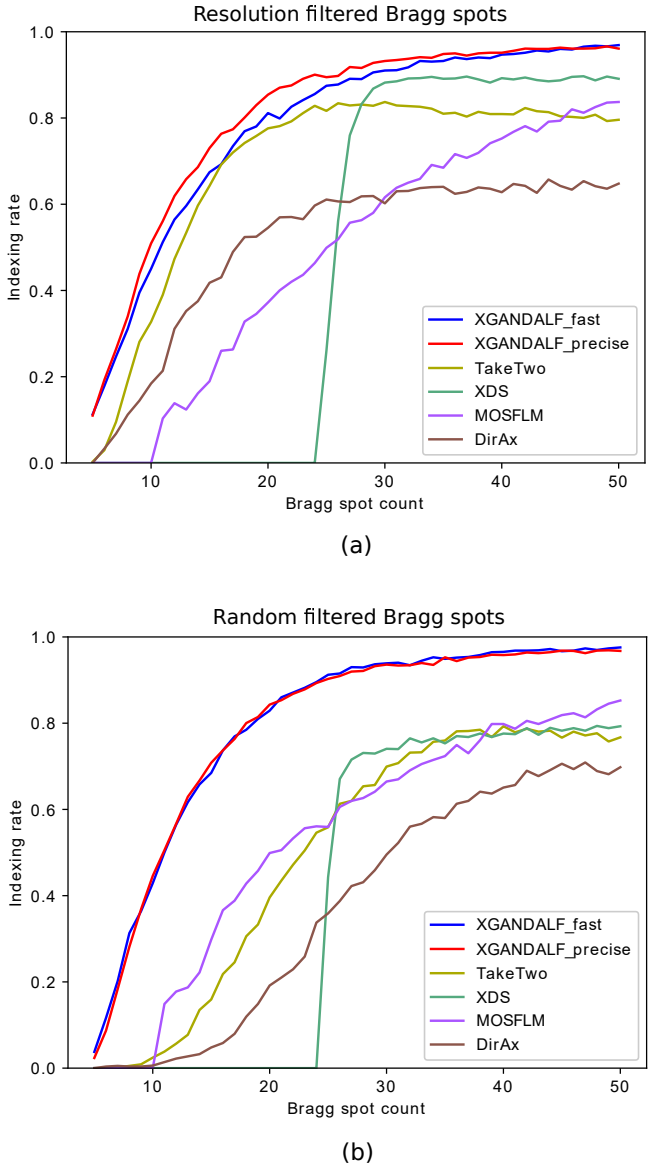


Figure 3.9: *Comparison of the success rates of algorithms in indexing patterns as a function of the numbers of Bragg spots N in those patterns. The patterns were generated by selecting N Bragg spots from real diffraction patterns: (a) the N low-resolution Bragg spots were selected, (b) random N Bragg spots were selected. XGANDALF was used with "precise" and with "fast" settings. XGANDALF outperforms the other indexers over the whole range of Bragg spot counts in both (a) and (b).*

Table 3.1: *Numbers of crystals of CXIDB ID 83 indexed without prior unit cell knowledge. "Indexed with correct unit cell" means here that the detected unit cell is close to the real unit cell and not a supercell or completely different.*

Indexer (no prior cell information)	Total indexed	Indexed with correct unit cell
DirAx	28 832	3553
MOSFLM	18 346	11 742
XGANDALF fast mode	26 040	14 631
XGANDALF precise mode	24 748	10 899

Table 3.2: *Numbers of crystals of CXIDB ID 83 indexed with prior unit cell knowledge.*

Indexer	Indexed
XDS	13 922
MOSFLM	16 120
MOSFLM DirAx XDS	17 433
TakeTwo	18 808
XGANDALF "fast" mode	19 914

the unit-cell parameters were known, the indexing was processed in one case without providing that knowledge (only MOSFLM, DirAx, and XGANDALF provided reasonably high indexing rates) and another with these parameters provided. Table 3.1 shows results for the case where the unit-cell parameters were not provided to the algorithms. In this case, the indexers often report unit cells that differ from the known ones. For a fair comparison, the numbers of correctly-identified unit cells are also listed. This indexing mode is typically used as a first step to estimate the unit-cell size. After this estimation, the data is re-indexed using the estimated unit-cell as prior knowledge, usually leading to better indexing results. Table 3.2 presents the case using known unit-cell parameters.

As seen from Tables 3.1 and 3.2, XGANDALF outperforms all the other state-of-the-art indexers with and without prior cell information. Surprisingly, without prior cell information, XGANDALF performs better in fast mode than in precise mode for this dataset. It is likely that in the "precise" mode more local maxima are found, making the choice for the basis selection algorithm more difficult. DirAx in total indexes the most patterns, but the number of useful, i.e. correctly identified, unit cells is very small. This is the case because DirAx significantly overestimates the unit-cell size and thus predicts too many Bragg spots by which it fools the check in CrystFEL for the correctness of an

Table 3.3: *Comparison of mean execution times (per pattern) and indexing results for a dataset consisting of 1000 patterns.*

Indexer name	Indexed patterns	Mean execution time (ms)
MOSFLM	452	17
XDS	400	22
DirAx	394	12
TakeTwo	545	662
XGANDALF fast mode	724	19
XGANDALF precise mode	725	106

indexing solution.

3.3.2 Execution time

For comparison of execution time, 1000 random patterns from the same dataset of CXIDB ID 21 as described above in Sec. 3.3.1 are indexed in the same fashion using the CrystFEL software suite on an INTEL E5-2698 v4 CPU. Here, however, no spots were removed from any of the patterns, nor were patterns only with a high number of Bragg peaks selected to create ground truths. The average number of Bragg peaks per pattern was 49. As before, the test was carried out for the two modes of XGANDALF—"XGANDALF precise mode" and "XGANDALF fast mode"—where parameters are chosen to either maximize the indexing success or maximize the indexing speed. Settings in between are also possible. The mean times to index the patterns are given in Table 3.3.

As can be seen in Table 3.3, again XGANDALF has the highest indexing rate among all tested indexers (in agreement with figure 3.9), while having an execution time of the same order of magnitude as the fastest-tested indexer. The high execution time for the TakeTwo algorithm reflects its mode of operation: if it does not find a solution, it will keep searching in the hope of eventually finding one, hence maximizing its indexing rate. Most patterns could be indexed by TakeTwo in a very short time, but several resulted in a long search. CrystFEL imposes a maximum running time on the indexing routines, and as a result, the execution time shown for TakeTwo reflects this maximum time rather than the performance of the algorithm.

3.3.3 Final merged data quality

After indexing, the next stage in the processing pipeline is the merge of the measured Bragg spots of all patterns into a set of structure factors. Better

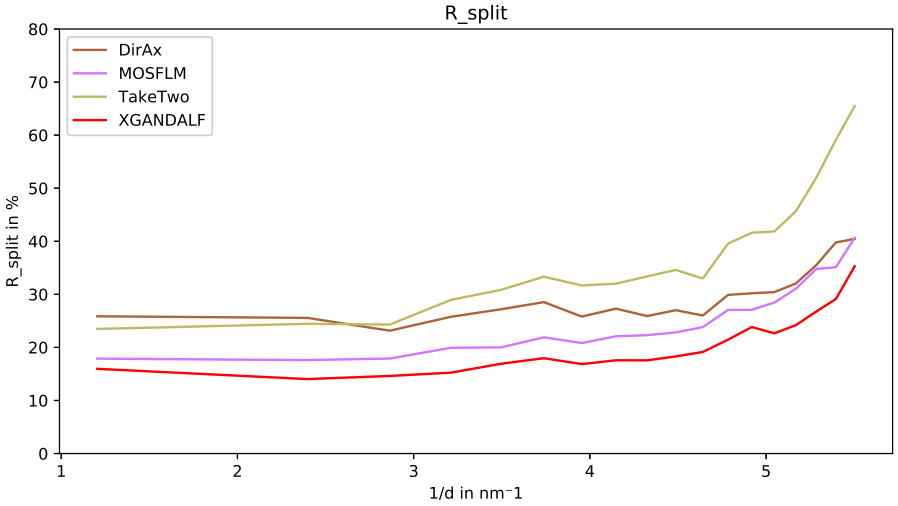


Figure 3.10: *Comparison of the achieved figure of merit R_{split} (White et al., 2013) (lower is better) for XGANDALF and current state-of-the-art indexers. XGANDALF outperforms the other indexers in all significant resolution shells.*

indexing results should presumably lead to better statistics of the merged data, so the quality of the merge can be used as a measure of the quality of the indexing results. Here a merged dataset was created with the indexed data of CXIDB ID 83 that were summarized in Tables 3.2 and 3.1. Apart from the indexing algorithm selection, all parameters to CrystFEL were the same for all tests. For comparison of the merge results the widely used crystallographic figures of merit $CC*$ (Karplus et al., 2012) and R_{split} (White et al., 2013) were employed.

As shown in Figs. 3.10 and 3.11, XGANDALF significantly outperforms the other indexers in both of these figures of merit.

For each indexed pattern, CrystFEL (White et al., 2012) estimates a profile radius of the Bragg spots. This is defined as the maximum distance of a reciprocal lattice point to the Ewald sphere that still gives rise to a Bragg spot, and can be considered as a property of the crystal, influenced by mosaicity for example. CrystFEL estimates this measure from the detected Bragg spots and the reciprocal lattice points that predict them best. A similar metric, called the Ewald proximal volume, was used by (Lyubimov et al., 2016) in their software *IOTA*. Errors in the indexing solution generally increase the estimated profile radius. A comparison of the profile radius estimation for MOSFLM and XGANDALF is depicted in figure 3.12. The estimated profile radii for patterns indexed by XGANDALF are generally smaller than the ones of MOSFLM,

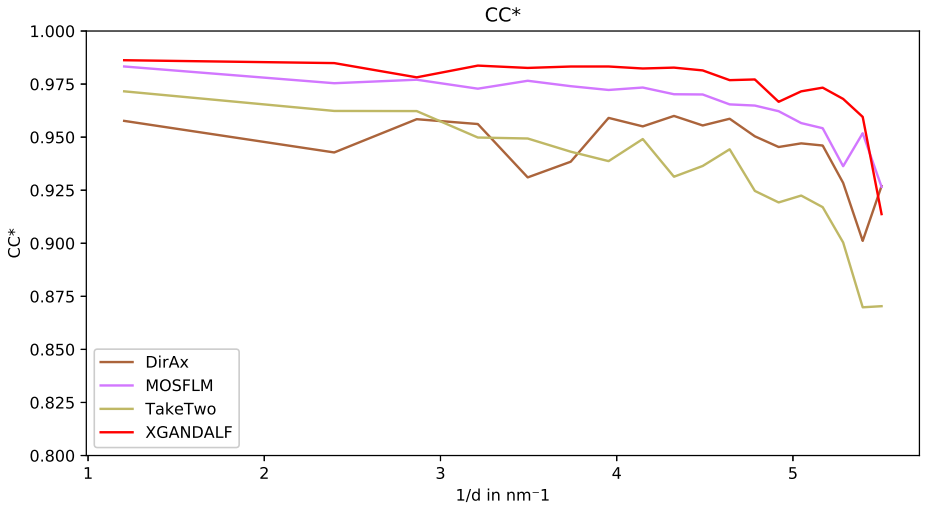


Figure 3.11: *Comparison of the achieved figure of merit CC∗ (Karplus et al., 2012) (higher is better) for XGANDALF and current state of the art indexers. XGANDALF outperforms the other indexers in all significant resolution shells.*

indicating that the indexing solution is more precise.

3.4 Code availability

XGANDALF is implemented as an open-source C++ library, which can be used directly from applications written in C or C++, or from a Python program using a Cython interface. XGANDALF has been implemented in CrystFEL (White et al., 2012) and is available from version 0.8.0 onwards. The XGANDALF implementation provides the tools for programmers to adjust the heuristic by defining their own high-level heuristic stages based upon optimized low-level implementations. The library is distributed under the LGPLv3 license, and the source code can be downloaded from
`https://stash.desy.de/users/gevorkov/repos/xgandalf`.

3.5 Conclusion

A new indexing algorithm, XGANDALF, has been presented which was designed specifically for indexing still diffraction patterns for snapshot serial crystallography experiments. As such, it outperforms the current state-of-the-art indexers that, although commonly used in serial crystallography, were mostly created

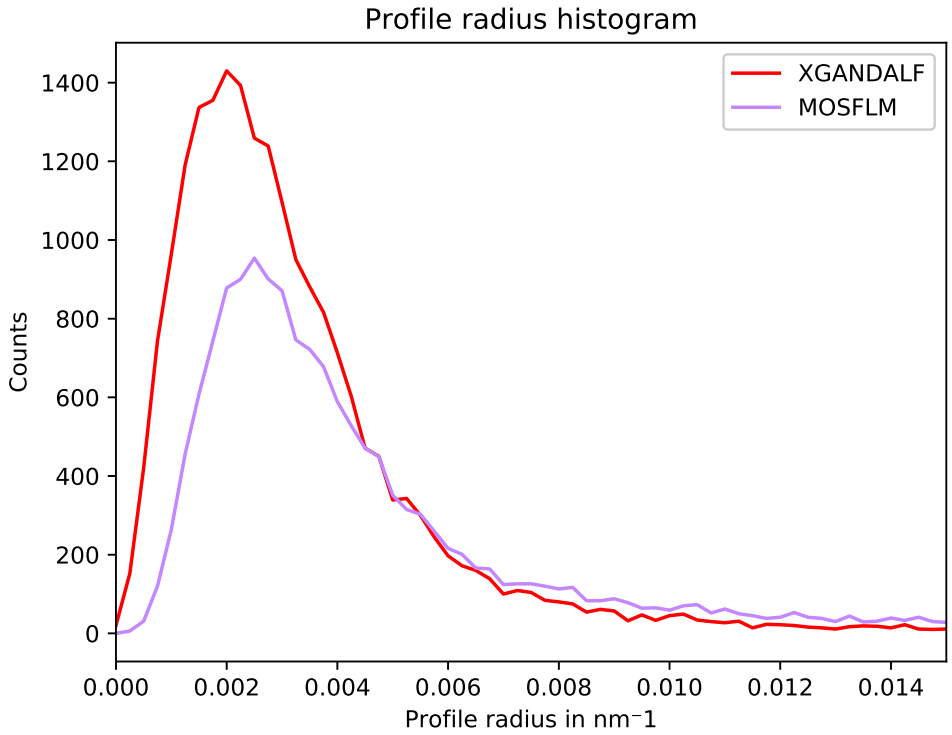


Figure 3.12: *Comparison of the estimated profile radii of MOSFLM and XGANDALF. The estimated radii for patterns indexed by XGANDALF are generally smaller than the ones of MOSFLM, which means that the indexing solution is more precise.*

for the indexing and analysis of rotation crystal data. Compared with those programs, XGANDALF gives higher indexing rates and higher indexing precision and can be used both with and without prior unit-cell parameters. The execution time of the implementation is of the same order of magnitude as currently-used indexing algorithms and, with mean indexing times of about 20 ms, is fast enough to allow real-time feedback in experiments. Compared with the available indexers, the algorithm successfully indexes more patterns in test serial crystallography data sets and is more robust to multiple lattices in a single image. The program has already been used in serial crystallography experiments by several other groups with very positive results. We, therefore, anticipate that XGANDALF will be a valuable addition to the collection of software tools for serial crystallography.

Chapter 4

Pink beam serial crystallography

So far, we only considered a monochromatic beam for use in crystallography, since it has been by far the most prominent technique for biological samples in recent decades. Nevertheless, polychromatic diffraction has been successfully applied to protein crystals over several decades as well (Clifton et al., 1991; McIntyre et al., 2006; Moffat et al., 1984; Myles et al., 1997; Ren et al., 1999). The motivation usually was a higher achievable total beam intensity. While polychromatic beam crystallography is feasible with X-ray sources and neutron sources, we concentrate on X-ray sources, since the indexing problem - which is the main topic of this manuscript - is greatly simplified for neutrons due to the availability of energy resolving detectors (Raventós et al., 2019).

Very wide bandwidth polychromatic beams are called white beams, because of the broad spectrum, or Laue beams (Helliwell et al., 1989), because they were used by Max von Laue when he discovered the first diffraction of a crystal. Current high-intensity X-ray sources typically produce polychromatic beams with a significantly narrower spectrum than the white beam used by Laue. A polychromatic beam that is narrower than a typical white beam but wider than monochromatic is called a pink beam (Meents et al., 2017).

4.1 Synchrotron light sources

Synchrotron radiation is the most prominent radiation for protein crystallography experiments.

Synchrotrons have been initially invented and built for particle physics experiments, where bunches of charged particles were accelerated to relativistic speed and then were forced to collide. The experiments aimed to analyze the results of the collision. The bunches of charged particles were accelerated by electromagnetic fields. The acceleration devices had to be reused to allow

acceleration to high energies. For this purpose, bending magnets were installed to force the particles to move in a circle - the synchrotron ring. The ring needs to be evacuated to avoid collisions between the accelerated particles and the air, the bending magnets and accelerators have to be synchronized to the current energy of the charged particle bunches - therefore the name "synchrotron".

Since accelerated charged particles lose energy by emission of photons orthogonal to the acceleration direction, the maximum achievable particle energy is defined by the centripetal acceleration of the particles in the bending magnets. To achieve high energies, large synchrotron rings have been built with radii ranging up to several kilometers. The radiation that results from the bending magnet acceleration is called synchrotron radiation. This parasitic radiation of early particle accelerator synchrotrons was used to conduct X-ray diffraction experiments and other experiments that required high energy photons. These synchrotrons are called first-generation synchrotron light sources.

With the success of the use of synchrotron radiation, more and more equipment was specifically designed to optimize its intensity and quality. Synchrotrons were updated and parts of the operation time were used to exclusively generate UV and X-ray photons. New synchrotrons, dedicated to the generation of photons, were built. They were pure storage rings for electrons, without the ability to produce collisions. These storage rings are the second generation of synchrotron light sources.

With increasing quality of synchrotron light, the number of users in various research areas and the number of experimental techniques increased as well. There was a clear need for brighter sources. This led to the development of the third-generation synchrotron light sources, which are used today. The electrons are kept on their orbit by bending magnets, but the radiation is produced by so-called insertion devices (see figure 4.1). Insertion devices generate periodic alternating magnetic fields on the path of the electrons. Contrary to bending magnets, they don't bend the major orbit of the electrons, and therefore, they can employ significantly stronger magnetic fields and increase the brightness by serially connecting many stages of magnets. By manipulating the period of the alternating magnetic field and their field strength, insertion devices such as wigglers and undulators can be designed to generate photons of specific ranges of energies with high intensities. Wigglers use very strong magnetic fields and generate broad spectra (white beam) with high total beam energy, while undulators have weaker magnets but can make use of coherent interference to create very intense narrow spectra, see figure 4.2.

For crystallography experiments, the lower bound of utilizable energy is given by the desired resolution in the order of 1 Å. Significantly lower energy does not add valuable signal but creates noise. Photons with significantly higher

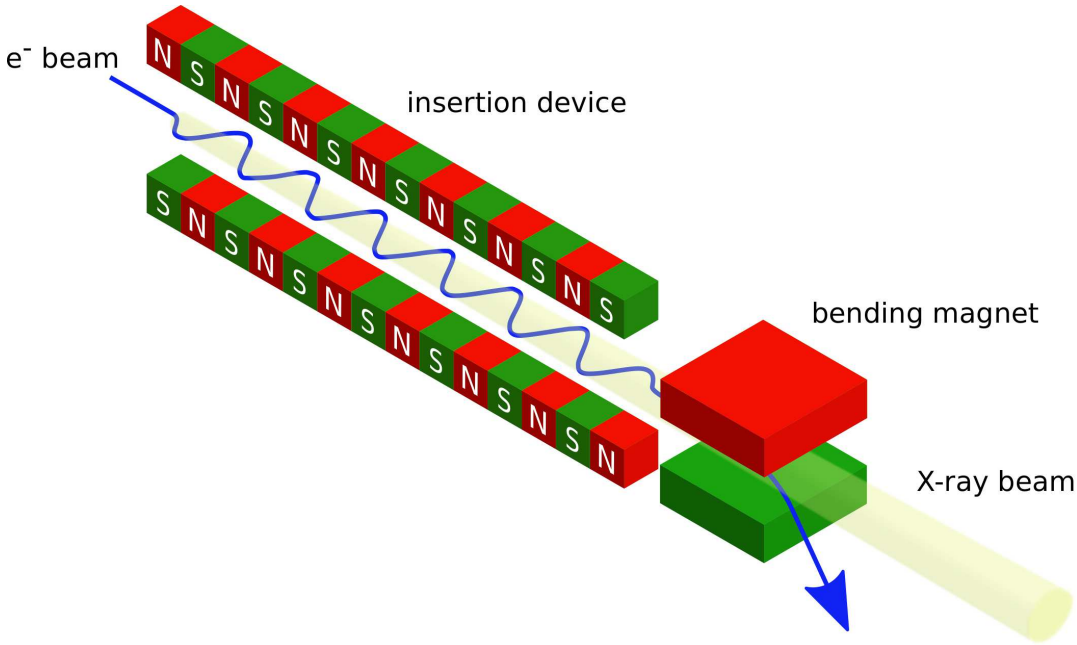


Figure 4.1: *Illustration of a wiggler or undulator. Insertion devices generate strong periodic alternating magnetic fields on the path of the electrons. They can employ significantly stronger magnetic fields than bending magnets since, due to the alternating arrangement, they do not change the major orbit of the electrons. Brightness can be increased by serially connecting many stages of magnets. By manipulating the period of the alternating magnetic field and its field strength, insertion devices such as wigglers and undulators can be designed to generate photons of specific ranges of energies with high intensities. This sketch was created based on figure 3.3 in Pithan, 2017.*

energies have a worse ratio of elastic/inelastic interaction with matter and thus are not desired as well. As a result, one spike of the undulator spectrum in the desired energy range is preferred over the wiggler spectrum. An example of an undulator spectrum where all but one spike have been filtered out can be seen as the green curve in figure 4.3. This is a typical pink beam spectrum. Typical wavelengths for macromolecular crystallography are between 0.75 Å and 3 Å.

Taking the wavelengths at 5% of the maximum intensity as a hard cutoff for the spectrum in figure 4.3, the (relative) bandwidth results in $\frac{|\Delta\text{Energy}|}{\text{Energy}} = \frac{|\Delta\lambda|}{\lambda} = 25\,\%$. Typical crystallography experiments employ a monochromatic beam with a bandwidth in the range of 0.1 %. Such beams can be created by using a monochromator, it requires filtering out over 99 % of the beams' energy (Meents et al., 2017), see red curve in figure 4.3.

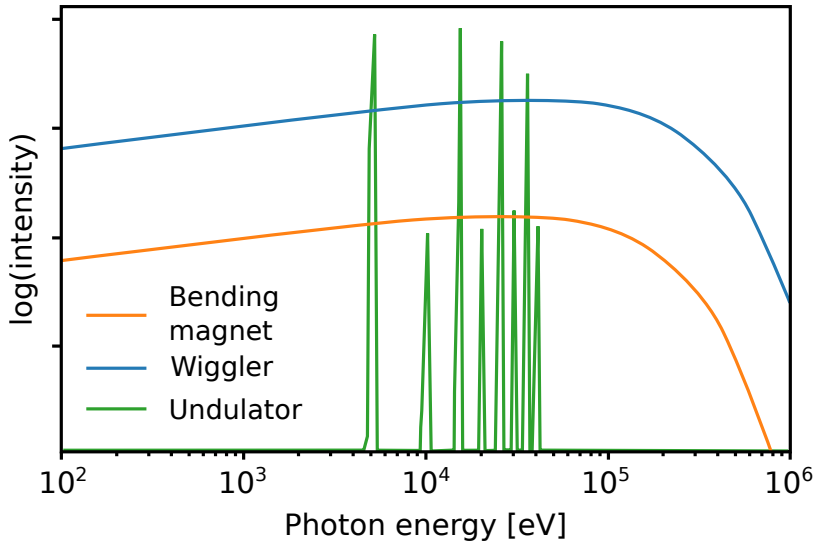


Figure 4.2: *Spectra of a bending magnet, a wiggler, and an undulator. Bending magnets generate a very broad spectrum of synchrotron radiation with a hard cut-off at the maximum energy. The spectrum of a wiggler is very similar to the one of the bending magnet, but with higher maximum energy and total intensity due to the stronger magnetic fields and serial connection of magnets. Undulators employ weak magnetic fields and thus reach lower maximum energies. The small amplitude of the oscillation allows coherent interference, which results in a spiky spectrum with large peak-intensities. This figure was created based on a diagram in a presentation of Clarke, 2007.*

4.2 Thick Ewald sphere

In chapter 2, the scattering for a single wave, i.e. monochromatic radiation, was derived. This model can easily be extended to account for polychromatic radiation by simply treating each wavelength individually: From equation 4.1 can be seen that the sum of two waves of different wavelengths is a beat oscillation. This is fundamentally different from the sum of waves of the same wavelength ($a = c$ in equation 4.1), where the sum is a simple wave and interference effects occur in the sense that the intensity of the resulting wave can be different from the summed intensities of the single waves. For waves with different wavelengths, the differences in instantaneous intensity between the single waves and the beat average out over time, as can be seen by equation 4.2.

$$\sin(ax+b)+\sin(cx+d) = 2\sin(\frac{(a+c)x}{2}+\frac{b+d}{2})\cos(\frac{(a-c)x}{2}+\frac{b-d}{2}) \quad (4.1)$$

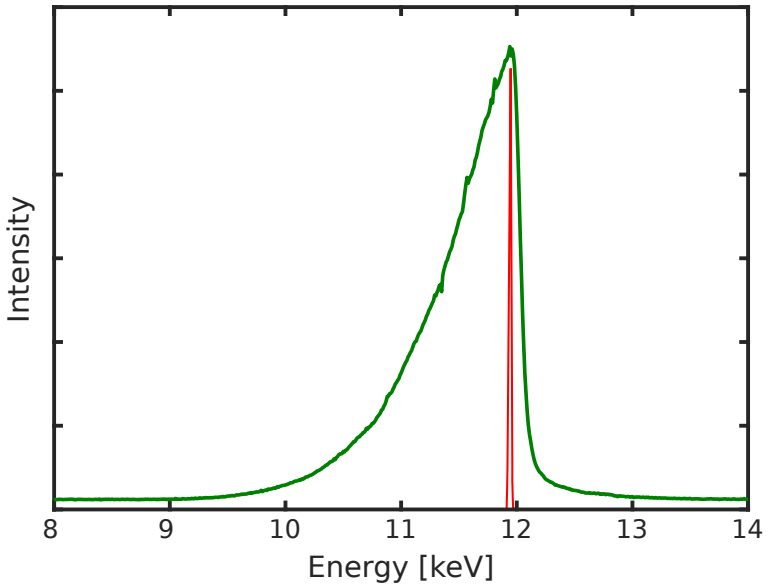


Figure 4.3: *Sample spectrum of an undulator where all but one spike have been filtered out (green). For crystallography experiments with a monochromatic beam, the spectrum is further filtered by a monochromator (red), resulting in a narrow spectrum. The monochromator filters out 99% of the total useful beam energy.*

$$\lim_{T\to\infty}\Big(\frac{\int_0^T \sin^2(ax+b)\mathrm{d}x}{T}+\frac{\int_0^T \sin^2(cx+d)\mathrm{d}x}{T} - \frac{\int_0^T(\sin(ax+b)+\sin(cx+d))^2\mathrm{d}x}{T}\Big)=0 \tag{4.2}$$

As interference effects do not occur with different wavelengths, the diffracted intensity in an experiment with a polychromatic beam can be described by adding up the individual diffracted intensities of all available wavelengths. The individual diffracted intensities can be computed by the aid of an Ewald sphere with radius $\frac{1}{\lambda}$. This results in an Ewald sphere for each nonzero wavelength in the spectrum. If the spectrum is continuous, the individual Ewald spheres combine into one "thick Ewald sphere". All points $\boldsymbol{q} = \frac{1}{\lambda}(\widehat{\boldsymbol{k}}_1 - \widehat{\boldsymbol{k}}_0)$ of the reciprocal space with $\lambda \in [\lambda_{\min},\ \lambda_{\max}]$ contribute to the diffraction intensity in direction $\widehat{\boldsymbol{k}}_1$. Here, $\widehat{\boldsymbol{k}}_1$ and $\widehat{\boldsymbol{k}}_0$ are unit vectors (see List of Symbols). All these points are located on a line segment the extension of which passes through the origin of the reciprocal space (green line segment in figure 4.4(b)).

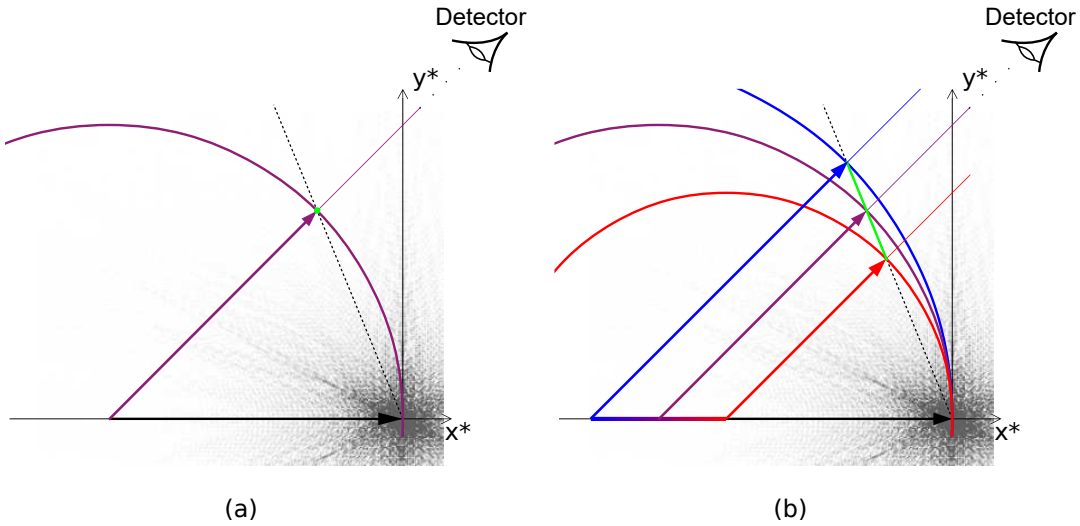
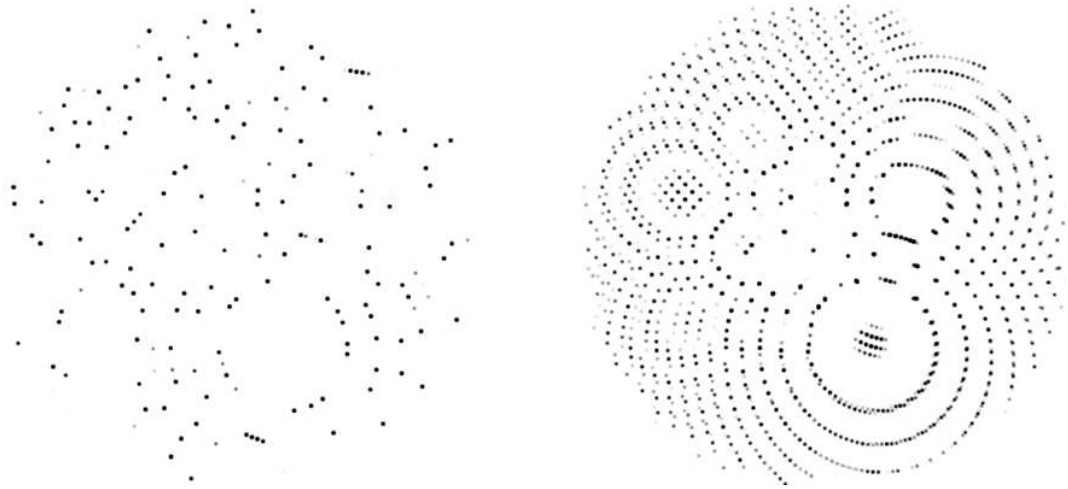


Figure 4.4: *Illustration of the diffracted intensity calculation. (a) Conventional Ewald sphere for a monochromatic beam. The diffracted intensity is proportional to the squared norm of the structure factor at the green dot on the Ewald sphere multiplied by a factor that is proportional to the intensity of the beam. (b) Thick Ewald sphere for a pink beam. The red sphere marks the Ewald sphere for the longest wavelength present in the beam, and the blue sphere marks the Ewald sphere for the respective shortest wavelength. The diffracted intensity is given by the integral of the squared norm of the structure factor along the green line segment and scaled by the respective spectral intensity of the incoming beam. The green line segment is the intersection of the thick Ewald sphere and a line that passes the origin. All outgoing waves marked by their wave vectors (red, purple, and blue arrow) create diffraction in the same direction and thus sum up to a beat oscillation at the detector.*

The total diffracted intensity thus can be computed as

$$\overline{\|E_1\|^2} = \int_{\lambda_{\min}}^{\lambda_{\max}} \|F(\boldsymbol{q}(\lambda))E'(\lambda)\|^2 \,\mathrm{d}\lambda \tag{4.3}$$

Here, $\overline{\|E_1\|^2}$ is the mean intensity of the diffracted wave, $F(\boldsymbol{q}(\lambda))$ is the structure factor for $\boldsymbol{q} = \frac{1}{\lambda}(\widehat{\boldsymbol{k}}_1 - \widehat{\boldsymbol{k}}_0)$, and $E'(\lambda)$ is proportional to the intensity of the incoming wave with wavelength λ.



(a) (b)

Figure 4.5: *Simulated diffraction patterns showing the same crystal with (a) monochromatic beam and (a) pink beam. The Bragg spots visible in (a) are visible in (b) as well. (b) contains many more Bragg spots than (a).*

4.3 Pink beam crystallography

Due to the thickness of the Ewald sphere, in a pink beam crystallography experiment typically significantly more reciprocal lattice points are excited than in an equivalent monochromatic experiment. A comparison of simulated diffraction patterns showing the same crystal with pink and monochromatic beam can be seen in figure 4.5. The direction of a Bragg spot that is generated by a reciprocal lattice point can be computed by first identifying the Ewald sphere that excited the RLP and then constructing $\boldsymbol{k}_1$ in the same way as in the monochromatic model, see figure 4.6.

The Ewald sphere that excites a point $\boldsymbol{q}$ in the reciprocal space can be computed by finding the set intersection between all Ewald spheres that can excite the point and the ones that are available from the experimental setup. An Ewald sphere is characterized by its center, as it, by definition, intersects the origin of the reciprocal space. All centers of Ewald spheres that excite the point $\boldsymbol{q}$ must have the same distance to $\boldsymbol{q}$ as to the origin, i.e. they must be located on the bisector of the line segment between $\boldsymbol{q}$ and the origin (orange line in figure 4.7). For the sketches, it is assumed that the beam direction equals the direction of the x^*-axis, and the bandwidth is finite, i.e. the available centers of Ewald spheres form a line segment on the x^*-axis (color gradient in figure 4.7).

The sphere that excites $\boldsymbol{q}$ thus is defined by the intersection of the bisector and the line segment on the x^*-axis, see figure 4.7.

4.3.1 Overlapping Brag spots

In monochromatic beam experiments, different points on the Ewald sphere are mapped to different positions on the detector. As shown in figure 4.4, this is not true for pink beam experiments - all points that lie on the green line segment are mapped to the same point on the detector. The consequence is that several Bragg spots of one crystal can be mapped to the same position on the detector. In that case, the summed intensity of the Bragg spots is measured instead of their individual intensities. This is problematic insofar as there are specific parts of the reciprocal lattice where RLPs cannot be measured individually in any orientation of the crystal. This is the case when excited RLPs that are located on a line that passes the origin are closer to each other than the length of the intersection of this line with the thick Ewald sphere in that region. One such example is sketched in figure 4.8, it shows RLPs on a line that passes the origin in different orientations. In this example, the violet RLP is always measured in a sum with at least one other (green) RLP. The smaller the spacing between the RLPs on the line that passes the origin, the more pronounced is this effect. i.e. it is most pronounced in the direction of the shortest reciprocal lattice vectors.

With M measurements of the summed intensities Σ_1 to Σ_M of N potentially overlapping RLPs with structure factors F_1 to F_N, a linear system of equations can be formulated as follows:

$$\begin{array}{rcl} p_{1,1}F_1 + p_{1,2}F_2 + ... + p_{1,N}F_N & = & \Sigma_1 \\ ... & = & ... \\ p_{M,1}F_1 + p_{M,2}F_2 + ... + p_{M,N}F_N & = & \Sigma_M \end{array} \tag{4.4}$$

Here, $p_{m,n}$ are the estimated partialities of the measured RLPs, i.e. the portions of the beam that get diffracted due to the RLPs. The partiality of an RLP depends on the shape of the RLP itself, the spectrum of the beam, and the location of the RLP in the Ewald sphere. The above linear system of equations can be equivalently formulated in matrix form as

$$\begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ & ... & & \\ p_{M,1} & p_{M,2} & \cdots & p_{M,N} \end{pmatrix} \begin{pmatrix} F_1 \\ ... \\ F_N \end{pmatrix} = \begin{pmatrix} \Sigma_1 \\ ... \\ \Sigma_M \end{pmatrix} \tag{4.5}$$

For $M > N$ this system of equations is over-determined and can be solved by linear regression to separate the structure factors.

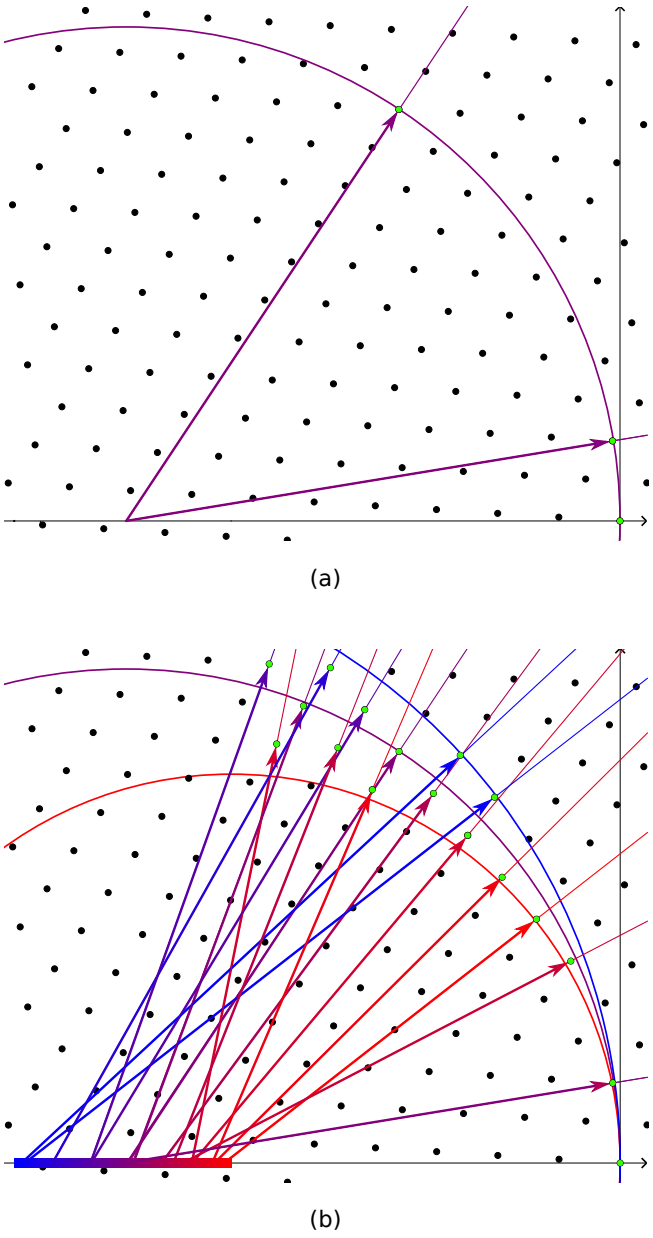


Figure 4.6: *Directions of Bragg spots generated by (a) a monochromatic beam and (b) a pink beam. In the pink beam case, many more RLPs are excited. The direction in which the Bragg spots can be measured can be computed by first identifying the Ewald sphere that excited the RLP and then constructing* $\boldsymbol{k}_1$ *in the same way as in the monochromatic model. Here, the* $\boldsymbol{k}_1$ *are marked by arrows of different colors depending on the wavelength: short wavelengths in blue and long wavelengths in red. Only* $\boldsymbol{k}_1$ *up to a maximum angle are sketched to avoid confusion due to too many arrows.*

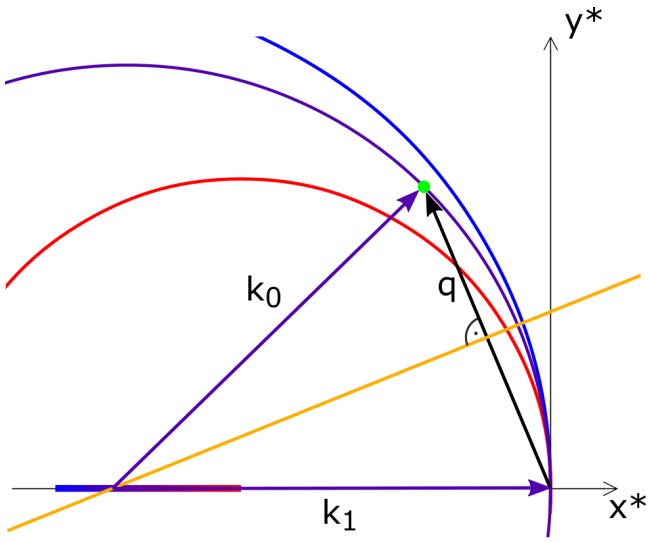


Figure 4.7: *In data processing for pink beam experiments, it is necessary to identify which Ewald sphere (i.e. what color) excited a point in the reciprocal space. This Ewald sphere can be identified by the set intersection between the set of Ewald spheres that can excite the point and the ones that are available in the experimental setup. The centers of Ewald spheres that can excite a point* $\boldsymbol{q}$ *(green) are located on the bisector (orange line) of the line segment between* $\boldsymbol{q}$ *and the origin. The Ewald sphere centers available from the experimental setup are on a line segment on the* x^**-axis, marked as a color gradient. The center of the Ewald sphere that excites* $\boldsymbol{q}$ *thus is on the intersection of the orange line and the color gradient.*

4.3.2 Streaky Bragg spots

With small and imperfect crystals, the RLPs are blurred. A prominent property of a crystal that causes blurry peaks is the so-called mosaicity. Due to defects, crystals usually are not monolithic blocks, but rather consist of many small blocks in slightly different orientations, creating additional RLPs in positions slightly rotated around the origin of the reciprocal space. The slightly rotated RLPs are so close to each other that they form one stretched RLP, see figure 4.9. Another increasingly important property that causes blurring of RLPs is the crystal size. Limited crystal size can be modeled by a multiplication with a top-hat function in the real space. Thus, in reciprocal space, the RLPs are convolved with the Fourier transform of the top-hat function, which becomes wider as the crystals become smaller.

With monochromatic radiation, small crystals can cause slightly enlarged Bragg spots on a two-dimensional detector and mosaic crystals can cause Bragg

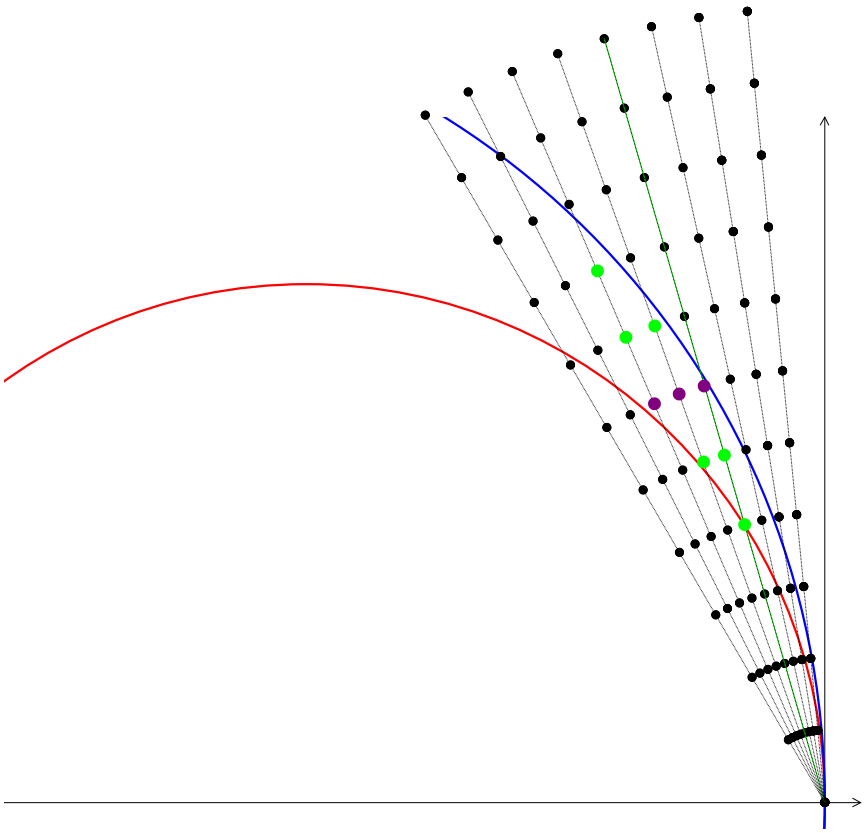


Figure 4.8: *Systematically overlapping Bragg spots in a pink beam crystallography experiment. Bragg spots on a line that passes the origin are sketched in different orientations. All RLPs on the same line produce overlapping Bragg spots, see figure 4.4. In this example the violet RLP is always measured in a sum with at least one other (green) RLP. There is no orientation where it can be measured individually.*

spots that are slightly stretched in the azimuthal direction. With a pink beam, both effects additionally cause a stretch in the radial direction that is typically larger than the stretch in the azimuthal direction. A simple example can be seen in figure 4.10, where a Bragg spot of a mosaic crystal has no extent in the radial direction with a monochromatic beam, but a large radial extent in the pink beam case. The general case that considers all relevant properties is more complex, and different counteracting effects need to be considered. As a rule of thumb, in typical pink beam experiments with significantly imperfect crystals, Bragg spots become streaky and thus can cause even more overlaps than described above. A simulated pink beam diffraction pattern with significant overlaps is given in figure 4.11. These partial overlaps can be corrected for by

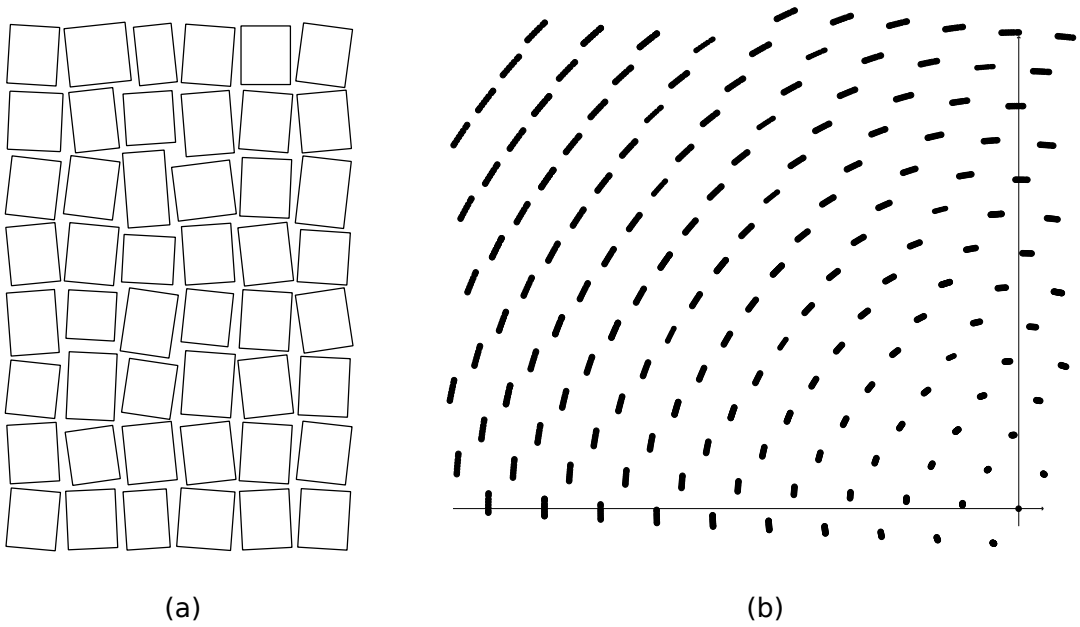


Figure 4.9: *Due to defects, crystals usually do not grow as perfect mono-crystals, but consist of many small mono-crystalline blocks with slightly different orientations. This property of a crystal is called mosaicity. (a) schematically shows a mosaic crystal. (b) shows the Fourier transform of the mosaic crystal. It can be seen as the sum of many slightly rotated versions of the Fourier transform of one of the mono-crystals. The effect is a blurring of the RLPs in the azimuthal direction.*

modeling the Bragg spot shape (Ren et al., 1995; Shrive et al., 1990) and fitting the model parameters for each diffraction pattern.

4.3.3 Background noise

In crystallography experiments, background noise is inevitable. Background noise typically arises from gas in the path of the X-rays or liquids in the sample delivery system, as well as in the crystal itself. The structure factor of liquids and gases is continuous and radially symmetric. In pink beam experiments, the diffracted intensity in one direction is computed by accumulation along a whole line segment in the reciprocal space, see figure 4.4. The useful information (i.e. RLPs) covers only a small part of the line segment, which results in a worse signal to background noise ratio with a pink beam compared to a monochromatic beam. The consequence is that in pink beam experiments, care has to be taken for the reduction of sources of background noise. Meents et al. showed 2017 that a setup optimized for a very low background with a pink beam is practical.

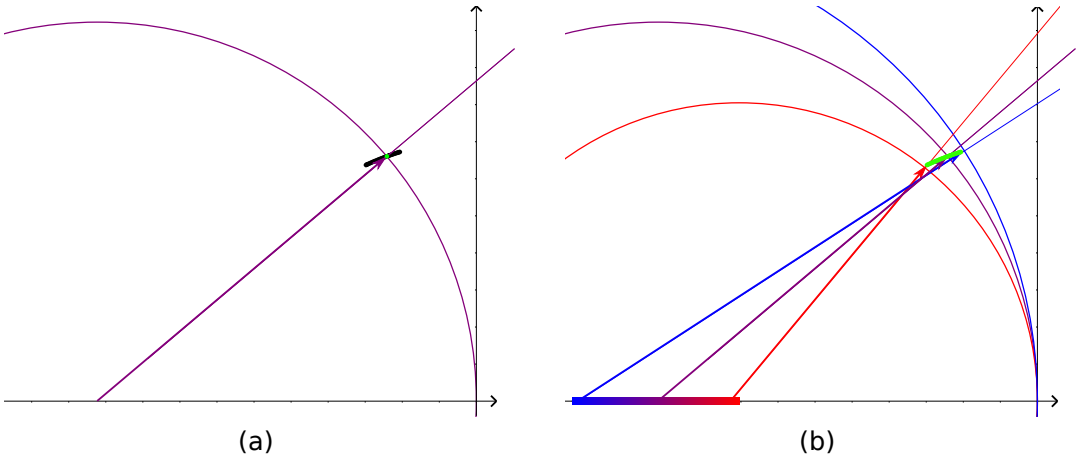


Figure 4.10: *An RLP that is blurred by mosaicity is excited by (a) a monochromatic beam and (b) a pink beam. The monochromatic beam excites only a small part of the blurred RLP. The Bragg spot is sharp in the radial direction. Note that the blurring also extends to the azimuthal direction, which is not sketched in this 2D diagram. This results in blurring of the Bragg spot in the azimuthal direction. In the pink beam case, the whole blurred RLP is excited. Waves with different wavelengths diffract in different directions, causing a blurring of the Bragg spot in the radial direction.*

The significance of the larger background is getting less with smaller and less-perfect crystals, of which the reciprocal lattice points have a significantly large size. Here, the larger bandwidth not only increases the amount of background noise but as well the amount of the valuable signal.

4.3.4 Partiality estimation

Serial crystallography suffers from the partiality problem, see section 2.3.2. With a monochromatic beam, the partiality highly depends on the shape of an RLP and its position in the reciprocal space. Proper estimation of the partiality needs pedantic refinement of the parameters. This is not always possible because the number of Bragg spots is typically low.

With a pink beam, the partiality estimation is simplified in two regards. First, the partialities' dependence on the RLPs position is relaxed. The spectrum is smooth, so slight changes in the position of the RLP cause only slight changes in the partiality. Second, the number of visible Bragg spots is usually high, allowing to precisely fit the model parameters to the diffraction pattern. With small and imperfect crystals, Bragg spots typically cover multiple pixels of the

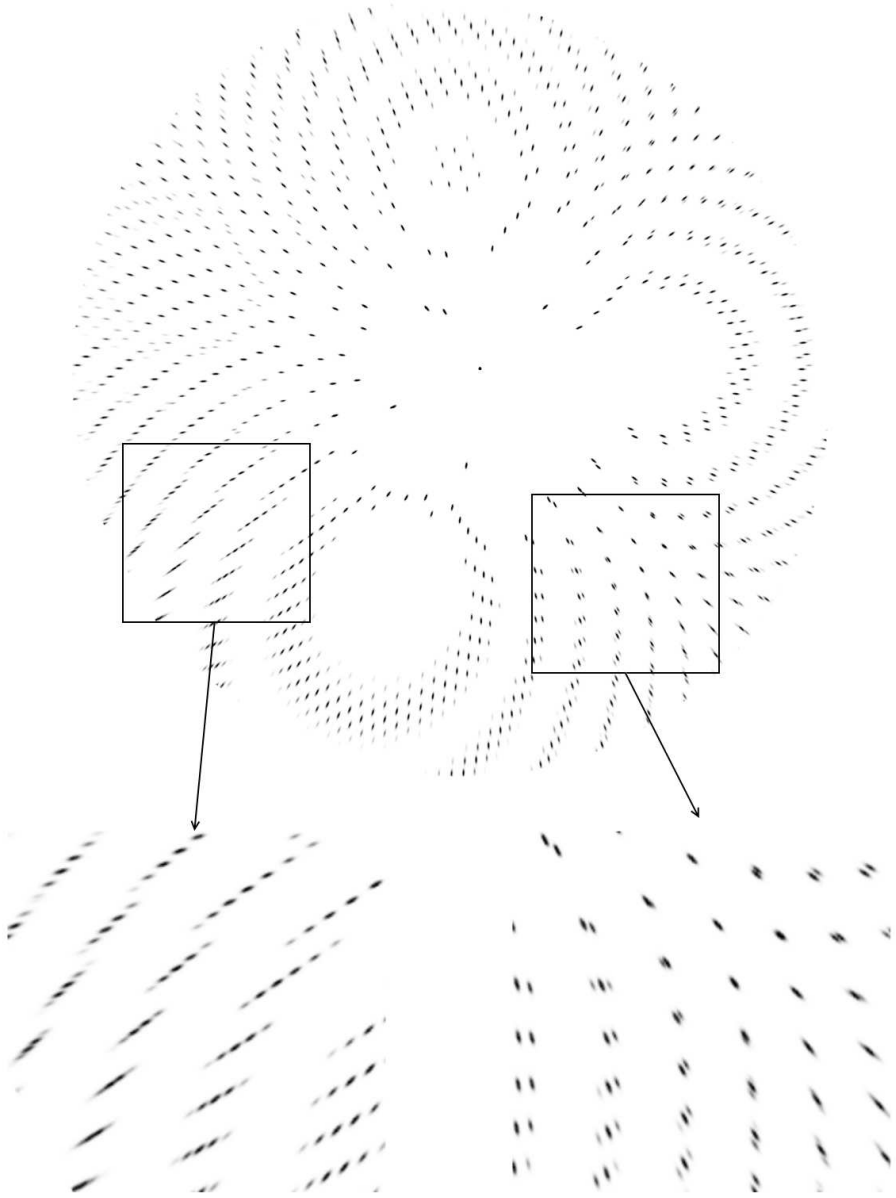


Figure 4.11: *Simulated pink beam diffraction pattern of a mosaic crystal. The zoomed regions have a high number of overlapping Bragg spots. The overlaps need to be corrected for before merging. This can be done by refining the model parameters using the shape of the Bragg spots.*

detector. Inspection of the Bragg spot profiles allows further refinement of the model parameters.

Proper partiality estimation helps drastically reducing the number of required diffraction patterns (Meents et al., 2017).

4.3.5 Intensity scaling

In serial crystallography, data from diffraction patterns of different crystals needs to be merged. Crystals have varying sizes, resulting in varying intensities of the Bragg spots on the detector. Prior to merging measured intensities from different crystals, they need to be scaled for the size of the crystal. Typically the sizes of the crystals are not measured in the experimental setup, so this parameter has to be estimated from the diffraction patterns.

If a Bragg spot from the same RLP appears in two different diffraction patterns, scaling one of the diffraction patterns such that the measured Bragg peak has the same intensity in both patterns allows merging the two diffraction patterns. Here, the correction for the partiality should be taken into account as well. The more Bragg spots the patterns share, the more robustly the scaling can be performed, which results in a better merged dataset. This approach can be extended to multiple diffraction patterns, allowing to take into account all Bragg spots that are pairwise shared among the patterns.

Due to the thick Ewald sphere, many more Bragg spots are recorded in pink beam experiments than in monochromatic serial crystallography. As a result, many more peaks are shared among the diffraction patterns. This allows a very robust scaling and merging in pink beam experiments (Rossmann, 2014).

4.3.6 Time-resolved crystallography

An important application of serial crystallography is the study of chemical reactions - the so-called time-resolved SX. A chemical reaction of every molecule in the crystal is initiated, and after a fixed time the state of the reaction is recorded by the X-ray laser. The initiation of the reaction can be accomplished e.g. by a laser pulse of visible light or diffusion of chemicals into the crystal. Using different time intervals between initiation and recording, a time-resolved measurement of the reaction can be obtained. Here, the exposure time limits the time resolution. The exposure time itself is limited by the intensity of the X-ray source. Exposure times in the range of milliseconds are common for synchrotron radiation sources (Botha et al., 2015; Martin-Garcia et al., 2017; Nogly et al., 2015; Stellato et al., 2014). Meents et al. demonstrated 2017 that utilizing the large intensity of the pink beam and a low-background setup, room-temperature serial crystallography with 100 ps exposure time is possible.

Moreover, a structure could be determined from as little as 50 diffraction patterns.

4.3.7 Conclusion

The use of a pink beam is highly beneficial for serial crystallography (Meents et al., 2017). In synchrotrons, it allows utilizing a two orders of magnitude higher total intensity than what is available for a monochromatic beam. It significantly facilitates the partiality estimation and the intensity scaling of diffraction patterns (Rossmann, 2014). Due to the higher flux, pink beam diffraction experiments can have significantly shorter exposure time than monochromatic beam experiments, allowing better resolution for time-resolved experiments.

As a result of the wide bandwidth, the model of an Ewald sphere needs to be extended to a thick Ewald sphere. This model is more challenging from a data processing point of view since it allows overlaps of Bragg spots. These challenges have already been solved for the most part in the last decades.

Another problem that is inherent to the broad bandwidth is the larger background noise compared to a monochromatic beam. However, it has been shown that the significance of this problem can be drastically reduced by an optimized experimental setup (Meents et al., 2017).

The reason why pink beam serial crystallography is not widely used yet is that there did not exist reliable automatic indexing algorithms for it. The results for the pink beam publication by Meents et al. 2017 e.g. were obtained by semi-manual indexing with the Precognition software (Ren et al., 1999), which included several days of an experts' labor for a few dozens usable diffraction patterns. The following chapter shows the details of the pink beam indexing problem and presents an automatic indexing algorithm that can reliably index pink beam diffraction patterns, enabling extensive use of pink beam serial crystallography.

Chapter 5

PinkIndexer - A universal indexer for pink-beam X-ray and electron diffraction snapshots

Large parts of this chapter are taken from the pinkIndexer publication (Gevorkov et al., 2020).

5.1 Introduction

The use of a pink beam is highly beneficial for serial crystallography. However, the automated analysis of pink-beam diffraction patterns has found to be problematic, with only 15 % of patterns successfully indexed in the demonstration of Meents et al. 2017 in a semi-automatic manner. Therefore, there was a motivation to create a new robust algorithm to index snapshot diffraction patterns recorded with a quasi-collimated beam of arbitrary bandwidth, with the requirement to index weak or incomplete patterns, using approximately-known unit cell parameters. In meeting this goal, the algorithm "pinkIndexer" was developed. Moreover, pinkIndexer can also be applied to several other data collection methods. In addition to superior performance in processing pink-beam diffraction compared with the state of the art, the algorithm indexes more patterns in monochromatic serial crystallography datasets than all other programs tested and is successful in indexing snapshot crystal diffraction patterns recorded with electrons.

In the case of a broad bandwidth X-ray beam, indexing is complicated by the uncertainty of the particular incident wavelength that gave rise to a given Bragg spot, while for electron diffraction the short wavelength results in an almost flat Ewald sphere for which the determination of unknown 3D lattice parameters

is ill-conditioned. Indeed the main bottleneck in pink-beam and electron serial crystallography analysis has been the indexing step.

Automatic indexing algorithms implemented in widely-used software including MOSFLM (Powell, 1999), XDS (Kabsch, 1993; Kabsch, 2010) and DirAx (Duisenberg, 1992) were originally devised for data collected in a rotation series with monochromatic radiation. They typically perform poorly when presented with individual pink-beam or electron snapshot diffraction patterns due to their reliance on the particular conditions of monochromatic rotation measurements. Recent algorithms designed for indexing snapshot diffraction patterns encountered in serial crystallography include TakeTwo (Ginn et al., 2016), FELIX (Beyerlein et al., 2017), and XGANDALF (Gevorkov et al., 2019). These all assume monochromatic radiation and do not fare much better than other indexers when processing polychromatic diffraction patterns. Several indexing approaches have been developed for polychromatic crystal diffraction, also referred to as Laue diffraction (Moffat, 1997). These include an approach due to Jacobson 1986 that requires the use of an energy-resolving position-sensitive detector; the Daresbury software suite for indexing Laue patterns (Campbell et al., 1998; Helliwell et al., 1989), and the Precognition software (Ren et al., 1999) based on searching arcs of reflections so that prominent zone axes can be identified; geometric approaches of Carr 1993 and Wenk 1997; and the LaueUtil toolkit (Kalinowski et al., 2011) which carries out a clustering analysis of possible orientations that map lattice vectors to observed peaks. The latter algorithm requires measurements of a crystal at several known relative orientations and is therefore not suited to serial crystallography. Of these, the current state-of-the-art software for indexing single wide-bandwidth diffraction patterns of macromolecular crystals is Precognition. However, while this works well for patterns recorded with a very wide spectrum (e.g. that of a wiggler or bending magnet where the bandwidth is more than 10 % of the nominal X-ray energy), it becomes less reliable as the number of Bragg spots decreases as occurs with either reduced spectral width (less than 5 % of the nominal X-ray energy) or with small crystals, where only several tens of Bragg reflections are observed.

Here the principles and performance of the new general indexing algorithm, pinkIndexer, are presented. As described in Sec. 5.2, the algorithm maps observed Bragg reflections into trajectories of possible lattice orientations. The most likely orientation is then determined as the orientation in which most trajectories intersect. As such, pinkIndexer covers the cases of monochromatic serial X-ray crystallography, X-ray crystallography using the unmodified spectrum of an undulator of 1 % to 25 % bandwidth and approximately 1 Å wavelength, and serial electron crystallography at approximately 0.01 Å wavelength. These cases are evaluated in Sec. 5.3. The algorithm can be employed in automated

processing of serial crystallography datasets, for example using the CrystFEL software suite (White, 2019; White et al., 2012).

5.2 The pinkIndexer Algorithm

5.2.1 Diffraction geometry

Consider elastic scattering from an object by a plane monochromatic wave characterised by a wave-vector $\boldsymbol{k}_0$ with a wavelength $\lambda = 1/|\boldsymbol{k}_0|$. In the kinematic approximation, the strength of scattering in a direction $\boldsymbol{k}_1$ is given by the magnitude of the Fourier transform component of the object at a spatial frequency $\boldsymbol{q}$ equal to the momentum transfer $\boldsymbol{k}_1 - \boldsymbol{k}_0$ (James, 1950). Elastic scattering ($|\boldsymbol{k}_1| = |\boldsymbol{k}_0|$) confines the observable spatial frequencies $\boldsymbol{q}$ to the Ewald sphere, shown as a circle in figure 5.1. Each pixel in a detector placed in the far field measures a particular direction given by the unit vector $\widehat{\boldsymbol{k}}_1$, which unambiguously maps to a point $\boldsymbol{q}$ in reciprocal space. The spatial frequency spectrum of a crystal of infinite extent is a lattice of points that are commonly referred to as reciprocal lattice points (RLPs), shown as black dots in figure 5.1. As can be seen in that figure, a diffraction spot observed in a particular direction $\widehat{\boldsymbol{k}}_1$ is unambiguously mapped to a particular RLP on the Ewald sphere (green dot in figure 5.1).

Consider now the case where the radiation source emits a finite but continuous distribution of wavelengths within some known range. Instead of a single Ewald sphere as in the case of monochromatic illumination, each incident wavelength produces an Ewald sphere with a radius inversely proportional to that wavelength. Thus a volume of reciprocal space can be excited in a diffraction experiment, contributing to the two-dimensional diffraction pattern. This volume is depicted in figure 5.2, bounded by the red and blue Ewald spheres (longest to shortest wavelengths in the range). There is a significant difference to the monochromatic case: with a polychromatic source, a particular scattering direction $\widehat{\boldsymbol{k}}_1$ no longer maps to a single point in reciprocal space. There may be many diffracted wave-vectors, each with a different wavelength (and hence different wave-vector magnitude and different placement in the Ewald-sphere construction), but pointing in the same direction $\widehat{\boldsymbol{k}}_1$ and thus arriving at the same point on the detector. These wave-vectors are depicted by the red, purple, and blue arrows in figure 5.2. Turning this around, for a given diffraction direction $\widehat{\boldsymbol{k}}_1$, there are many points $\boldsymbol{q}$ in reciprocal space that contribute to the diffracted intensity. All these points lie on a straight line segment (green line in figure 5.2), the extension of which passes through the origin of reciprocal space. The line segment can be described by $\frac{1}{\lambda}(\widehat{\boldsymbol{k}}_1 - \widehat{\boldsymbol{k}}_0) \mid \lambda \in [\lambda_{\min}, \lambda_{\max}]$.

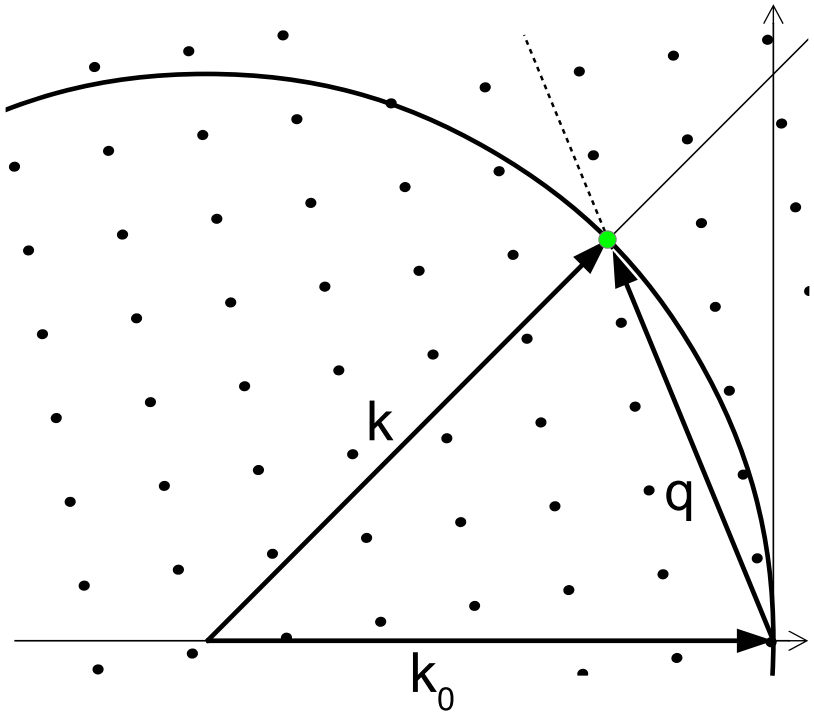


Figure 5.1: *A two-dimensional representation of reciprocal space, showing RLPs of a crystal and the Ewald sphere for monochromatic diffraction.* $\boldsymbol{k}_0$ *is the wave vector of the incident beam,* $\boldsymbol{k}_1$ *is the outgoing wave vector elastically scattered from the structure in the object with spatial frequency* $\boldsymbol{q}$*. The diffraction intensity measured by a detector in the direction of vector* $\boldsymbol{k}_1$ *corresponds to the intensity at the green RLP.*

It can be seen that in the case of broad bandwidth, a point on the detector integrates signal from a line segment in reciprocal space, in contrast to a single point in the monochromatic case. The RLP which generates a Bragg peak observed at some position on the detector may therefore lie anywhere on the corresponding line segment. Let's call this line segment the Uncertainty Line Segment (ULS), shown in green in figure 5.2. The main challenge for analyzing broad-bandwidth snapshot crystal diffraction patterns is to determine the point along the ULS where the RLP[1] lies which generated the observed Bragg peak. This is equivalent to identifying the wavelength that excited the measured RLP. Note that if more than one RLP lies on the ULS, they will contribute to the observed intensity, excited by different wavelengths. The bandwidth in that case would be too broad to distinguish those particular reflections in the peak

[1]Or multiple RLPs

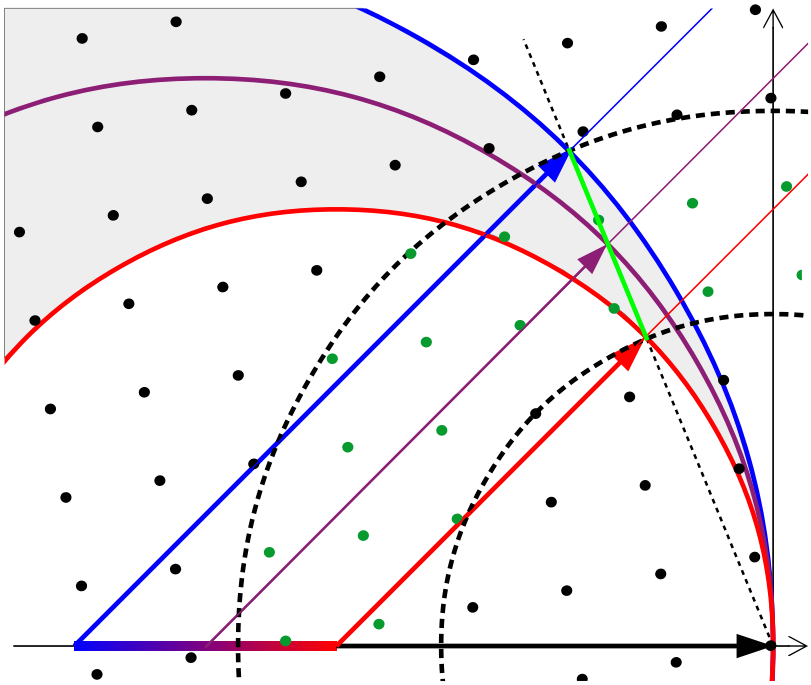


Figure 5.2: *A two-dimensional representation of reciprocal space depicting Ewald spheres for polychromatic diffraction spanning the longest wavelength shown in red to the shortest in blue. The volume between the red and blue spheres is excited and the intensity measured at the detector in the direction of the highlighted vectors corresponds to the integral of diffraction intensities along an uncertainty line segment (ULS) (green line). Candidate reciprocal lattice points (dark green dots) lie in the volume between shells depicted by the dashed circles. Note that the red, purple and blue vectors are parallel and therefore the corresponding photons arrive at the same pixel of the detector. However, since the photons have different wavelengths, they do not interfere coherently and their intensities are simply added together.*

finding stage without energy resolving detectors. It is nevertheless possible to separate the summed intensities after indexing and integration (Shrive et al., 1990; Zurek et al., 1985).

Since the crystal orientation is not known, and thus the orientation of the reciprocal lattice is also not known, candidate reciprocal lattice points may lie anywhere in the volume between $\boldsymbol{q}$-shells centered at the origin with radii set by the scattering direction and range of wavelengths as depicted by the dashed circles in figure 5.2. Let's call the RLPs that can match a ULS by rotation of the reciprocal lattice "candidate RLPs" (candidates to predict the particular Bragg spot). The candidate RLPs are plotted in dark green in figure 5.2.

5.2.2 Determining the crystal orientation

The task of indexing is to find the crystal orientation which gives rise to a particular measured diffraction pattern and then to assign indices to predicted reflection locations. In practice, this is achieved by finding the crystal orientation which best predicts the set of Bragg peaks observed on the detector. It is assumed that the unit cell parameters of the crystal are known. PinkIndexer determines the likely crystal orientation as follows: 1) For each Bragg spot observed on the detector, find all RLPs of a crystal that can be intersected by the Bragg spot's ULS by rotation around the origin (candidate RLPs); 2) For each observed Bragg spot, find all rotations of the crystal that place at least one candidate RLP onto the corresponding ULS. This is equivalent to finding all orientations of the reciprocal lattice that could predict the measured Bragg spot; 3) Find the orientation which predicts the most Bragg spots from the list of candidate orientations for all Bragg peaks observed in the pattern. The orientation which correctly predicts the most observed Bragg spots will be the chosen indexing solution; finally, 4) Refine the lattice parameters and other experimental parameters to further improve the agreement of predicted and observed Bragg peaks (If the original parameters were not accurate, one could repeat steps 1 to 4 using the refined parameters). Once the crystal orientation is determined it is, of course, possible to predict the location and wavelength of all potential reflections including absent or weak reflections not present in the set of observed Bragg peaks. These can then be included in the integration of the observed intensities for structure determination.

The main challenge lies in making the search outlined above tractable and robust. As we will now discuss, for each candidate RLP $\boldsymbol{h}$ there is an infinite set of reciprocal lattice rotations which place it onto its particular ULS. All this family of rotations is identified by constructing a rotation operation in two steps: first the reciprocal lattice is rotated such that the vector $\boldsymbol{h}$ of the RLP is rotated by an angle π around the axis $\hat{\boldsymbol{m}}$ that bisects $\boldsymbol{h}$ and $\hat{\boldsymbol{q}}$ as shown in figure 5.3. This rotation brings the candidate RLP onto the ULS. Next, the reciprocal lattice is rotated around $\hat{\boldsymbol{q}}$ by a rotation of ϕ (see figure 5.3). Since the rotated candidate RLP lies on the ULS it is invariant to the second rotation and thus all rotations ϕ are potential orientations of the lattice. This construction is only valid for one particular candidate RLP and a particular ULS. The particular RLP might not actually give rise to the Bragg spot which generated the ULS. That is, none of the orientations of the lattice generated by the operations $R_{\hat{\boldsymbol{q}}}(\phi)\, R_{\hat{\boldsymbol{m}}}(\pi)$ might be the indexing solution. In a triclinic lattice only one candidate RLP, as well as other RLPs lying on the same straight line through the origin, generate the correct indexing solution. In lattices with higher symmetry, multiple candidate RLPs generate correct indexing solutions.

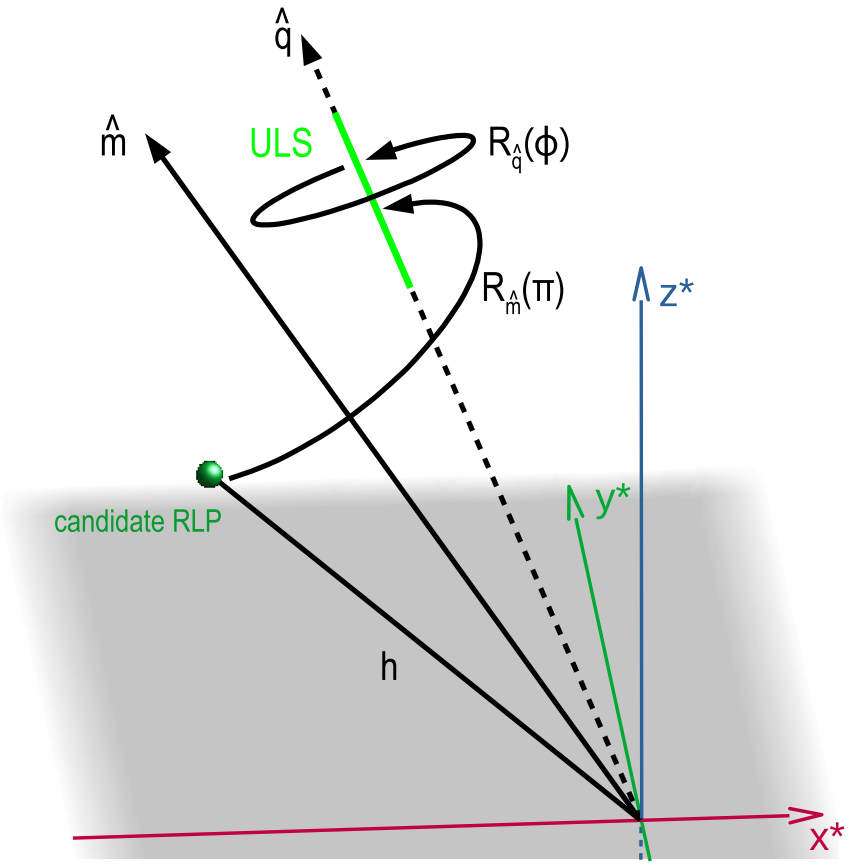


Figure 5.3: *All possible rotations for which a candidate reciprocal lattice point (RLP) intersects with a given uncertainty line segment (ULS) are uniquely described by the rotation* $R = R_{\hat{\boldsymbol{q}}}(\phi)R_{\hat{\boldsymbol{m}}}(\pi)$. *The first rotation* $R_{\hat{\boldsymbol{m}}}(\pi)$ *rotates the lattice by the constant angle of* π *around the bisecting vector* $\hat{\boldsymbol{m}}$ *of* $\boldsymbol{h}$ *and* $\hat{\boldsymbol{q}}$. *The second rotation* $R_{\hat{\boldsymbol{q}}}(\phi)$ *rotates the lattice around the axis* $\hat{\boldsymbol{q}}$ *of the ULS by an angle* ϕ

To determine common orientations that bring a number of RLPs onto ULSs a three-dimensional vector space is defined that contains curves parameterized by $\hat{\boldsymbol{m}}$, $\hat{\boldsymbol{q}}$ and the rotation angle ϕ, satisfying the reflection conditions stated above that place a particular candidate RLP onto a particular ULS (corresponding to an observed Bragg spot). The vector space consists of three dimensions since it is spanned by three variables describing rotations, such as the three Euler angles. In this 3D space, all candidate RLPs for a particular Bragg spot will form a set of non-intersecting curves[2]. We call this collection a rotogram. By combining

[2]The parameter along the curves is ϕ.

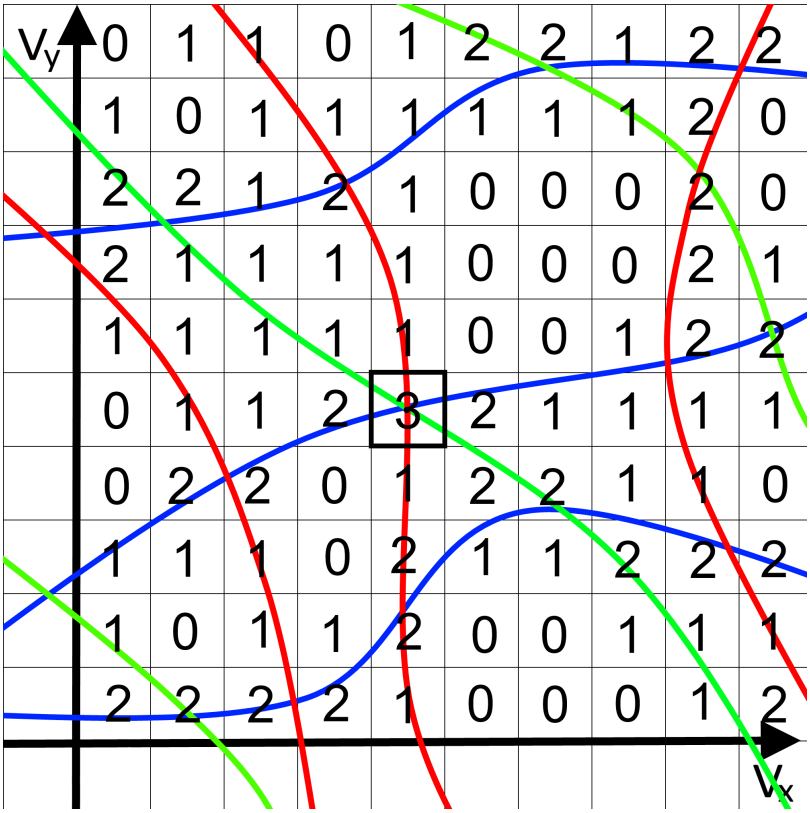


Figure 5.4: *Schematic sketch of a rotogram in a 2D vector space. Each point in the rotogram corresponds to a rotation (defined by the constants v_x and v_y) of a sample reciprocal lattice. Each color marks the positions in the rotogram, that predict one particular Bragg spot. The position that is marked by most colors corresponds to the orientation that predicts most Bragg spots correctly—a good indexing solution. The parameter along the curves is the rotation angle ϕ around $\hat{\boldsymbol{q}}$.*

rotograms for all measured Bragg spots, a total rotogram for a diffraction pattern is formed, depicted schematically in figure 5.4. The point in a rotogram with the highest density of overlapping curves provides the lattice orientation that predicts most of the observed Bragg spots. This point represents the rotation of the lattice onto that of the measured crystal, i.e. it is the indexing solution. The task of crystal orientation determination is therefore now reduced to one of finding the point in rotation space with the largest number of intersecting lines.

5.2.3 Algorithm details

In practice, many additional issues arise when dealing with data from a real experiment that complicate the indexing process, such as spurious intensity peaks resulting from experimental noise, or multiple crystals in the beam contributing to the same diffraction pattern. The robustness of the algorithm to these factors becomes a critical issue.

The pinkIndexer uses the same basic approach as another indexing method for monochromatic crystal diffraction patterns, FELIX (Beyerlein et al., 2017), which similarly parameterizes possible orientations as curves in a 3D rotation space. Felix is based on a method implemented by Schmidt 2014 in the program GrainSpotter and previously more generally described by Morawiec 2005. Both methods are similar to the Hough and Radon transforms that operate on 2D parameter spaces. With such approaches, the choice of the mapping function is crucial for the performance and simplicity of the algorithm. Well-known mappings for 3D rotations to a 3D space are: the Euler-angles representation, the axis-angle representation, the Gibbs representation, and the modified Rodrigues parameters (Terzakis et al., 2018). The pinkIndexer algorithm employs a novel mapping function by which it achieves drastic reduction of complexity and the number of necessary parameters compared to FELIX (which uses Rodrigues parameters), while at the same time increasing the noise tolerance. The following features of the transform are desired for robustness and efficient construction of the rotogram:

- adjacent voxels in the rotogram correspond to similar rotations;
- rotations are distributed uniformly across a volume of the 3D rotation space;
- the results of the transformation are efficiently discretized in cuboid samples (i.e. on an orthogonal lattice);
- the transform is computable in an efficient way

Since none of the well-known examples sufficiently fulfill these requirements, another transform is proposed here, that better fulfills the requirements and is the major factor in the quality of the pinkIndexer algorithm. In this scheme, a single rotation operation $R_{\hat{\boldsymbol{e}}}(\theta)$ is determined from the composite rotation $R_{\hat{\boldsymbol{q}}}(\phi)\, R_{\hat{\boldsymbol{m}}}(\pi) = R_{\hat{\boldsymbol{e}}}(\theta)$. This rotation then is mapped to the point $\boldsymbol{v}$ in the rotogram given by $\boldsymbol{v} = \arctan(\theta/4)\, \hat{\boldsymbol{e}}$, where $\hat{\boldsymbol{e}}$ is the rotation axis, θ is the rotation angle and $\arctan(\theta/4)$ is a non-linear scaling factor. Compared with the well-known axis-angle representation which maps a rotation to $\theta\, \hat{\boldsymbol{e}}$ (i.e. the length of the vector encodes the rotation angle), this definition only slightly

increases the computational burden and inherits its property of adjacent voxels corresponding to similar rotations. The nonlinear scaling of rotation angle to the length of the vector gives a more uniform distribution of points in the rotogram than the axis-angle representation. The new transform maps all possible rotations to a finite-size ball of radius $\arctan(\pi/4)$ and is in some sense the opposite[3] of the modified Rodrigues parameters, $\tan(\theta/4)\,\hat{\boldsymbol{e}}$.

The construction of $R_{\hat{\boldsymbol{e}}}(\theta)$ from $R_{\hat{\boldsymbol{q}}}(\phi)R_{\hat{\boldsymbol{m}}}(\pi)$ is achieved in a computationally inexpensive way by employing the composition law for finite rotations first derived by Olinde Rodrigues (Altmann, 1989; Pujol, 2013). This describes the consecutive operations of two general rotations $R_{\hat{\boldsymbol{m}}}(\psi)$ and $R_{\hat{\boldsymbol{q}}}(\phi)$ to give $R_{\hat{\boldsymbol{e}}}(\theta) = R_{\hat{\boldsymbol{q}}}(\phi)\,R_{\hat{\boldsymbol{m}}}(\psi)$ by solving

$$\begin{aligned} \cos\frac{\theta}{2} &= \cos\frac{\phi}{2}\cos\frac{\psi}{2} - \sin\frac{\phi}{2}\sin\frac{\psi}{2}\,\hat{\boldsymbol{q}}\cdot\hat{\boldsymbol{m}} \\ \sin\frac{\theta}{2}\,\hat{\boldsymbol{e}} &= \sin\frac{\phi}{2}\cos\frac{\psi}{2}\hat{\boldsymbol{q}} - \cos\frac{\phi}{2}\sin\frac{\psi}{2}\,\hat{\boldsymbol{m}} + \sin\frac{\phi}{2}\sin\frac{\psi}{2}\,\hat{\boldsymbol{q}}\times\hat{\boldsymbol{m}}. \end{aligned} \tag{5.1}$$

For our problem, the first rotation axis is the bisector $\hat{\boldsymbol{m}} = (\hat{\boldsymbol{h}} + \hat{\boldsymbol{q}})/\left\|\hat{\boldsymbol{h}} + \hat{\boldsymbol{q}}\right\|_2$, and $\hat{\boldsymbol{q}}$ is the direction of the ULS. This choice of $\hat{\boldsymbol{m}}$ allows setting $\psi = \pi$ such that the equations simplify to

$$\begin{aligned} \cos\frac{\theta}{2} &= -\sin\frac{\phi}{2}\,\hat{\boldsymbol{q}}\cdot\hat{\boldsymbol{m}} \\ \sin\frac{\theta}{2}\,\hat{\boldsymbol{e}} &= -\cos\frac{\phi}{2}\,\hat{\boldsymbol{m}} + \sin\frac{\phi}{2}\,\hat{\boldsymbol{q}}\times\hat{\boldsymbol{m}}. \end{aligned} \tag{5.2}$$

Setting the parameters $c_1 = \sin(\phi/2)$, $c_2 = -\cos(\phi/2)$, $d_1 = \hat{\boldsymbol{q}}\cdot\hat{\boldsymbol{m}}$, and $\boldsymbol{d}_2 = \hat{\boldsymbol{q}}\times\hat{\boldsymbol{m}}$ we obtain

$$\begin{aligned} \cos\frac{\theta}{2} &= -c_1\,d_1 \\ \sin\frac{\theta}{2}\,\hat{\boldsymbol{e}} &= c_2\,\hat{\boldsymbol{m}} + c_1\,\boldsymbol{d}_2, \end{aligned} \tag{5.3}$$

which can be solved as

$$\begin{aligned} \theta &= 2\arccos(-c_1 d_1) \\ \hat{\boldsymbol{e}} &= \frac{c_2\,\hat{\boldsymbol{m}} + c_1\,\boldsymbol{d}_2}{\sin(\theta/2)},\ \theta \neq n\cdot 2\pi \end{aligned} \tag{5.4}$$

For $\theta = n\cdot 2\pi$, $\hat{\boldsymbol{e}}$ can be chosen as an arbitrary unit vector. For machines where $1/\sqrt{x}$ is implemented in hardware, replacing $1/\sin(\theta/2)$ by $1/\sqrt{1-(c_1d_1)^2}$ can lead to faster execution.

[3]The tangent is replaced by its inverse - the arctangent.

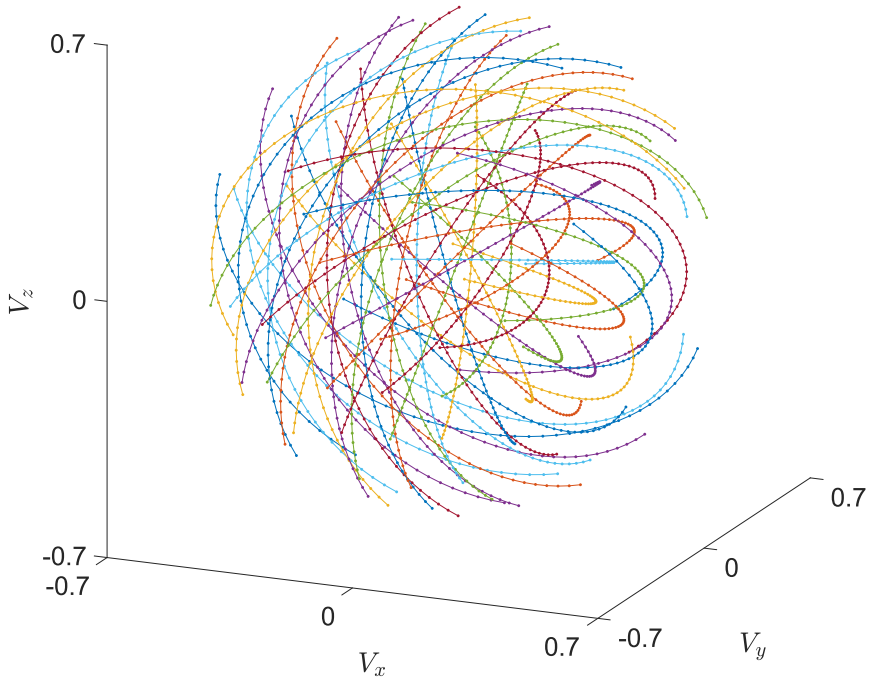


Figure 5.5: *Plot of the curves* $\boldsymbol{v} = \arctan(\theta/4)\,\hat{\boldsymbol{e}}$ *for a particular Bragg point direction* $\hat{\boldsymbol{q}}$ *and a set of* 56 *candidate RLPs in a cubic lattice.* $\phi \in [-2\pi, 0)$

An example of a rotogram for a particular Bragg peak is shown in figure 5.5, for 56 candidate RLPs on a cubic lattice and a relative bandwidth of 6.5 %. Each of the non-intersecting 56 coloured curves is a plot of the vector $\boldsymbol{v} = \arctan(\theta/4)\,\hat{\boldsymbol{e}}$ for a full rotation of the lattice around $\hat{\boldsymbol{q}}$. For a given point $\boldsymbol{v}$ in the plot, the corresponding rotation angle of the lattice to bring the RLP onto the ULS defined by $\hat{\boldsymbol{q}}$ is $\theta = 4\tan(\|\boldsymbol{v}\|_2)$, and the rotation axis is $\hat{\boldsymbol{v}}$. As seen from eqn. 5.2, each trajectory lies in the plane containing the orthogonal vectors $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{q}} \times \hat{\boldsymbol{m}}$, which is to say the plane normal to $\hat{\boldsymbol{m}} \times \hat{\boldsymbol{q}} \times \hat{\boldsymbol{m}} = \hat{\boldsymbol{q}} - d_1\,\hat{\boldsymbol{m}}$. The trajectories form closed curves in the vector space over the range of ϕ from 0 to 4π, but only a single rotation of the lattice is required. To keep the rotogram volume as small as possible the range $\phi \in [-2\pi, 0)$ is chosen. In the example of figure 5.5 the curves were sampled over that range at steps of 0.1 rad and while the curves are not necessarily uniformly sampled in the vector space, the choice of $\boldsymbol{v} = \arctan(\theta/4)\,\hat{\boldsymbol{e}}$ generates curves that more uniformly fill the space than the axis-angle representation $\theta\,\hat{\boldsymbol{e}}$ or any other of the constructions tested here.

In the implementation of pinkIndexer, rotograms are not calculated continuously as shown in figure 5.5 but computed on discrete sets of $N \times N \times N$ voxels that circumscribe the ball of radius $\arctan(\pi/4)$. For each Bragg spot the voxel array is initialized with zeroes and voxels are set to 1 that are intersected by

the curve $\boldsymbol{v} = \arctan(\theta/4)\,\hat{\boldsymbol{e}}$ for each of the candidate RLPs, with a uniform sampling of ϕ that is chosen to ensure that the curve is contiguous across the voxels. This is accomplished by computing the parameters $c_1 = \sin(\phi/2)$ and $c_2 = -\cos(\phi/2)$ at those values once for the whole rotogram and using values of the parameters $d_1 = \hat{\boldsymbol{q}} \cdot \hat{\boldsymbol{m}}$ and $\boldsymbol{d}_2 = \hat{\boldsymbol{q}} \times \hat{\boldsymbol{m}}$ that need to be computed once per curve. To make the discretization of ϕ smoother, the flagged voxels are dilated by setting all of their 26 neighbouring voxels to 1. This reduces the effective resolution of the rotogram, but increases the noise tolerance. The rotogram indicates all orientations of the crystal that predict the respective Bragg spot.

By adding each Bragg spot's rotogram, a total rotogram is created where the value of each voxel gives the number of Bragg spots predicted by the corresponding orientation. The voxel with the maximum value thus indicates the most likely lattice orientation that provides the correct indexing solution. The task of indexing is thus reduced to finding the location of this maximum. Since the rotogram is discrete the determined indexing solution is approximate. A subsequent refinement is carried out to increase the precision of the indexing solution, in which the lattice basis is refined to minimize the mean Euclidean distance between the ULS and the respective closest RLP using a gradient descent approach. Only the RLPs close enough to a ULS are used for this refinement to improve noise tolerance.

5.2.4 Implementation details

The algorithm has been implemented such that parameters have effects that are easy to understand. Besides experimental settings like detector distance and pixel geometry, beam parameters, and crystal lattice parameters, pinkIndexer requires a relative tolerance, set by the parameter *tolerance*, to decide when a peak is correctly fitted. Additional parameters trade off fitting performance against execution time. The parameter *consideredPeaksCount* specifies the number of found Bragg spots that are used to compute the initial indexing solution from the maximum of the rotogram. All Bragg spots are considered during refinement. The parameter *angleResolution* sets the resolution of the rotogram in terms of number of voxels N spanning $-\arctan(\pi/4)$ to $\arctan(\pi/4)$. Choosing larger voxels (lower resolution) leads to a faster calculation but lower precision in the initial step of determining the orientation from the rotogram. The second step of refining the orientation is controlled by the parameter *refinementType*. Refinement can be performed by a gradient descent method, fitting all parameters of the lattice or keeping the cell parameters constant and just refining the orientation. All parameters take descriptive values wherever possible.

5.3 Evaluation of algorithm performance

The performance of the pinkIndexer algorithm is evaluated on data from macromolecular crystal diffraction experiments utilizing three different types of radiation: monochromatic X-rays, pink X-rays, and electrons. For the evaluation, the CrystFEL(White et al., 2012) software suite 0.8.0+50a3cb06 with modifications to include the pinkIndexer library and enable prediction for wide-bandwidth and electron beams is used.

5.3.1 Monochromatic X-ray beam Crystallography

The performance of pinkIndexer in treating monochromatic serial femtosecond X-ray diffraction data was compared with the indexers MOSFLM (Powell, 1999), XDS (Kabsch, 1993) (Kabsch, 2010), DirAx (Duisenberg, 1992), TakeTwo (Ginn et al., 2016), FELIX (Beyerlein et al., 2017) and XGANDALF (Gevorkov et al., 2019) using the indexamajig program from the CrystFEL(White et al., 2012) software suite. For the test, all CrystFEL optimizations were turned off by using the options `-no-retry -no-refine -no-multi -no-check-cell -no-check-peaks`. Only one indexing solution per pattern was accepted. Indexing solutions that differed from the original indexing solution by less than 3° were counted as correct. The diffraction dataset was retrieved from the CXIDB (Maia, 2012), entry 21, from SFX measurements of a G-protein coupled receptor (the serotonin 5-HT_{2B} receptor bound to ergotamine) (Liu et al., 2013).

Comparing algorithms using real data provides results that indicate their performance under real conditions. However, unlike when using simulated data, the true indexing solution is unknown. The indexing solutions can be tested for correctness by comparing the predicted Bragg spots to the found ones. This is a precise method when there are many Bragg spots, but when the number of found spots is small there can be several incorrect orientations of a crystal that predict the found spots well enough to pass the indexing test. Following a practice that was introduced earlier to compare indexing algorithms (see section 3.3.1) semi-simulated datasets with different numbers of Bragg spots were created by removing spots from patterns with large numbers of spots (which have reliable indexing solutions). As previously, the indexers were tested in two modes of Bragg peak removal. In one mode the sets contained patterns with only 5 to 50 Bragg spots selected randomly from the patterns. In the other mode, the sets of patterns contained 5 to 50 Bragg spots only at low resolution[4]. The comparison is graphed in figure 5.6. All algorithms performed well when there

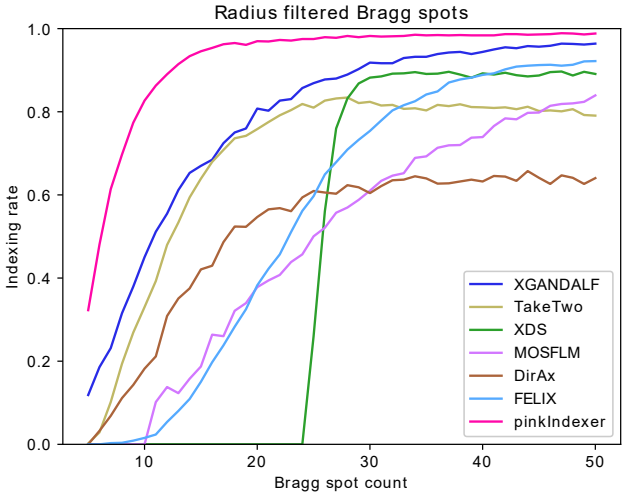
[4]The datasets were created by removing Bragg spots at high diffraction angles from the diffraction patterns.

Indexer name	Indexed patterns	total execution time [mm:ss] / per pattern [ms]
MOSFLM	452	00:17 / 17
XDS	400	00:22 / 22
DirAx	394	00:12 / 12
TakeTwo	545	11:02 / 662
FELIX	656	01:05 / 65
XGANDALF	724	00:19 / 19
pinkIndexer fast mode	757	04:15 / 255
pinkIndexer precise mode	816	22:16 / 1336

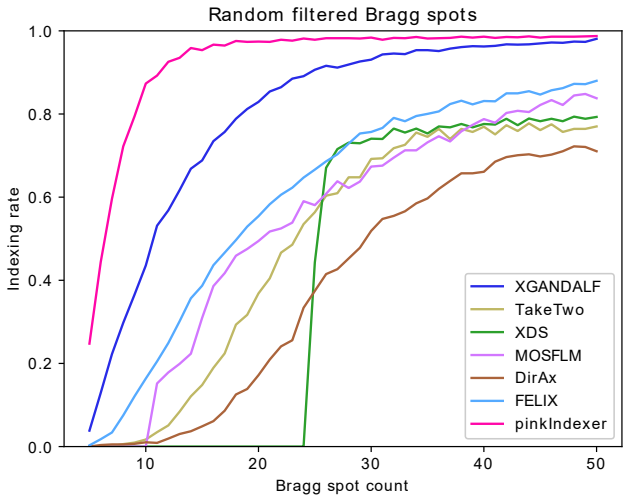
Table 5.1: *Indexing results and execution times for monochromatic indexers and pinkIndexer.*

were sufficient measured Bragg spots to determine the crystal lattice. With both cases of randomly-distributed Bragg spots and low-resolution Bragg spots, the pinkIndexer algorithm outperformed all others over the whole range of Bragg spot counts. The settings of pinkIndexer in these tests were chosen to favor precision over speed (*angleResolution* = "dense"). In all cases, the lattice parameters were specified to the indexing algorithm. No additional tuning of the indexing algorithms was performed apart from an option that allows FELIX to index patterns with as few as 5 peaks.

The average times for the various algorithms to index monochromatic diffraction patterns are given in Table 5.1, computed by indexing a set of 1000 diffraction patterns chosen randomly from the same dataset as used above. To ensure a fair comparison, all indexers were called from CrystFEL with the `-no-retry` flag set. No attempt was made to index multiple crystals per pattern. The program was executed on a dual-socket INTEL Xeon E5-2698 v4 CPU (2.20MHz, 20 cores, 50MB cache, 512GB RAM). pinkIndexer was tested with settings to maximize the speed ("fast mode" in Table 5.1, *angleResolution* = "loose") and with settings to maximize the yield ("precise mode", *angleResolution* = "dense") which took about five times longer. Settings in between are also possible. Even in the fast mode, the algorithm takes considerably longer than other algorithms except for TakeTwo. The slower speed of pinkIndexer is because the algorithm is memory intense and makes extensive use of the random memory access pattern. Nevertheless, due to its high indexing success rate, pinkIndexer can be profitably used as a fallback option for monochromatic diffraction patterns that cannot be indexed by other indexers. This can be implemented in CrystFEL by placing pinkIndexer last in the list of indexers.



(a)



(b)

Figure 5.6: *Comparison of pinkIndexer to other indexing algorithms, using the CrystFEL software suite. Each original experimental diffraction pattern had more than 50 Bragg spots and was indexable with MOSFLM. From every pattern, a number of Bragg peaks were selected to be used for indexing (indicated along the abscissa) either (a) by selecting those with the smallest scattering angles or (b) randomly. The indexing results were compared to the original MOSLFM indexing solution using all Bragg spots. Indexing solutions that differed from the original indexing solution by less than 3° were counted as correct.*

5.3.2 Pink-Beam Serial Crystallography

Diffraction patterns collected using pink-beam radiation (1 % to 25 % relative bandwidth) contain many observed peaks, which means that indexing solutions can easily be verified by comparing the predicted to the observed Bragg spots. To evaluate pinkIndexer using real pink-beam serial crystallography data the dataset of Proteinase K crystal diffraction from Meents 2017 measured at the 14-ID-B (BioCARS) beamline at the Advanced Photon Source (APS) was used, utilizing the full polychromatic spectrum of an undulator harmonic. A representative diffraction pattern is depicted in figure 5.7. Although the FWHM of the incident X-ray beam spectrum was 5 % of the mean photon energy, the tails of the spectrum extended up to 25 % of the mean photon energy. The dataset contained 999 patterns that had been classified as crystal diffraction "hits" based on the detection of at least 35 Bragg spots in the original work of Meents 2017. Of these, 667 patterns were successfully indexed by pinkIndexer, with 428 determined to contain a single lattice, 168 with two lattices, and 71 with three or more lattices. This gave a total of 1005 indexed lattices. The vast majority of the 332 patterns that could not be indexed appeared to be falsely identified as crystal diffraction patterns due to fitting peaks to noise.

The comparison of polychromatic indexing programs in an automated way for serial crystallography datasets is complicated by these programs not being able to be called from within the CrystFEL software suite. Indeed, the analysis of the pink-beam serial crystallography data carried out in Meents 2017 could only be achieved in a semi-manual way using the software Precognition. This resulted in the indexing of 140 patterns of the 999 (Meents et al., 2017), and only single-crystal diffraction patterns could be indexed. This comparison shows that for serial crystallography, pinkIndexer provides an order of magnitude more indexable patterns than the current state of the art software. Moreover, pinkIndexer can deal with multiple crystals per pattern and is fully automatic, thus making serial crystallography with a pink beam much easier. Figure 5.7 shows two crystals contributing to the pattern which are both indexed correctly. We have also successfully used the pinkIndexer for serial crystallography data from (Tolstikova et al., 2019) measured with X-rays with 2.5 % relative bandwidth produced using a multilayer monochromator.

5.3.3 Serial Electron Crystallography

Electron crystallography poses a challenge for conventional indexing algorithms due to the flatness of the Ewald sphere caused by the short de Broglie wavelength of the electrons. To demonstrate the applicability of pinkIndexer to serial electron diffraction data, a rotation series dataset from (Cruz et al., 2017) was treated

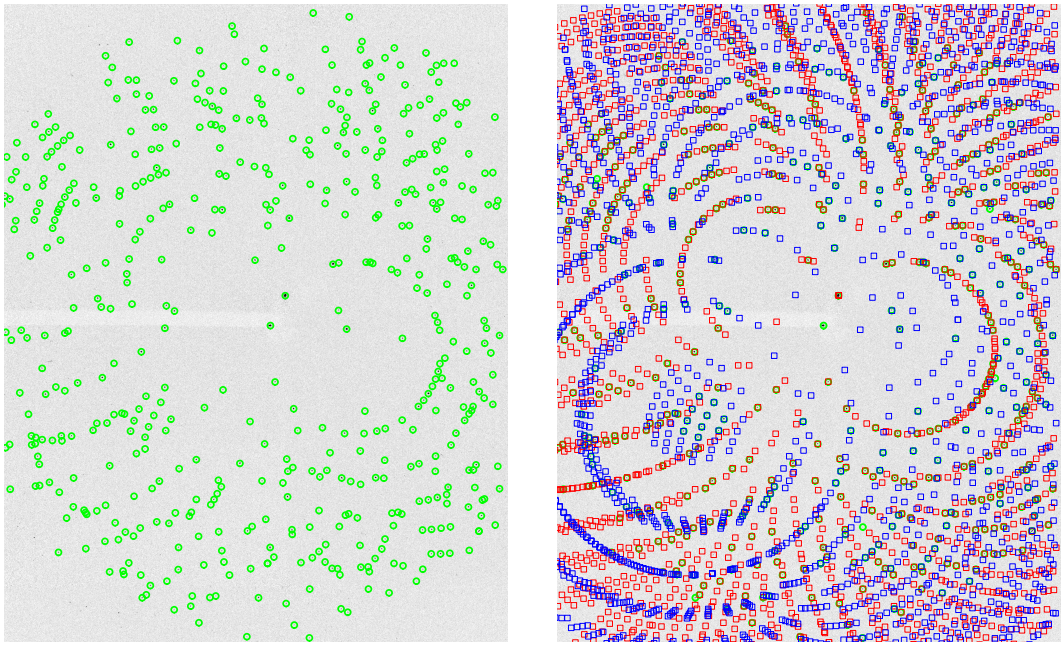


Figure 5.7: *A snapshot pink beam diffraction pattern of two Proteinase K crystals measured coincidentally at the 14-ID-B (BioCARS) beamline at APS. The pattern is overlaid with the detected peaks (left) and with detected as well as predicted peaks (right). Both crystals were indexed by pinkIndexer (red and blue). Note the agreement between prediction and measurement.*

like serial data by indexing each pattern individually. The known rotation increment available for this data set was used to check the correctness of the indexing solutions. The results are displayed in figure 5.8. All patterns from the dataset were indexed correctly, as can be seen from the linear increment of the determined rotation angle. The maximum deviation of the angle determined by indexing from a linear fit was 0.14°, which represents an upper bound to the indexing precision since goniometer errors may also contribute. This result enables serial electron crystallography using randomly-oriented crystals exposed in individual data frames as performed in SFX measurements (Bücker et al., 2020).

5.4 Conclusion

The indexer presented in this paper has been developed for pink-beam serial crystallography using the full polychromatic spectrum of an undulator harmonic

at a synchrotron radiation facility. Starting from known unit cell parameters, the pinkIndexer algorithm works by mapping all possible rotations of candidate reciprocal lattice points onto line segments in reciprocal space that correspond to Bragg peaks of unknown wavelength. By examining these mappings in a novel rotation space the most likely lattice orientation can be found. The main limitation of pinkIndexer is its lower speed compared to many other existing algorithms for monochromatic diffraction analysis and the requirement of knowing the cell parameters of the studied crystals. The benefit, however, is its higher success rate in indexing snapshot diffraction patterns than all other algorithms tested here.

By generalizing over different wavelengths and spectral characteristics, the algorithm presented here has the ability to open up emerging and as yet unexplored avenues of serial crystallography. The higher X-ray flux of the polychromatic beam enables exposure times as short as those emitted from a single electron bunch in the storage ring, while the broad bandwidth leads to a high fraction of fully-integrated Bragg peaks recorded in a snapshot pattern and a greater coverage of reciprocal space. These advantages have long been appreciated for time-resolved Laue diffraction experiments at synchrotron facilities (Moffat, 1997), macromolecular crystallography at neutron facilities (Blakeley et al., 2008), and have motivated the generation of pulses with broader bandwidth at FEL facilities (Dejoie et al., 2013; White et al., 2013). As demonstrated here, the pinkIndexer program overcomes difficulties previously encountered in automatically analyzing thousands of polychromatic diffraction patterns. Additionally, the generality of the algorithm also makes it useful for indexing monochromatic serial crystallography. In this case, we found that pinkIndexer demonstrates a superior success rate in indexing diffraction patterns, especially for the tricky case of a small number of detected Bragg spots. It has also been shown that the approach works well for indexing snapshot diffraction patterns recorded with very short wavelengths, which is usually the situation in electron diffraction. The method might additionally find application in neutron diffraction and could be slightly modified to treat the case of convergent beam diffraction.

5.5 Code availability

The pinkIndexer is implemented in C++ and released as an open source library under the LGPLv3 license. This library can be compiled independently or together with the program suite CrystFEL (White et al., 2012). The full processing pipeline, including indexing, high-precision prediction, and integration, is realized as a part of CrystFEL. The source code can be downloaded at `https://stash.desy.de/users/gevorkov/repos/pinkindexer`.

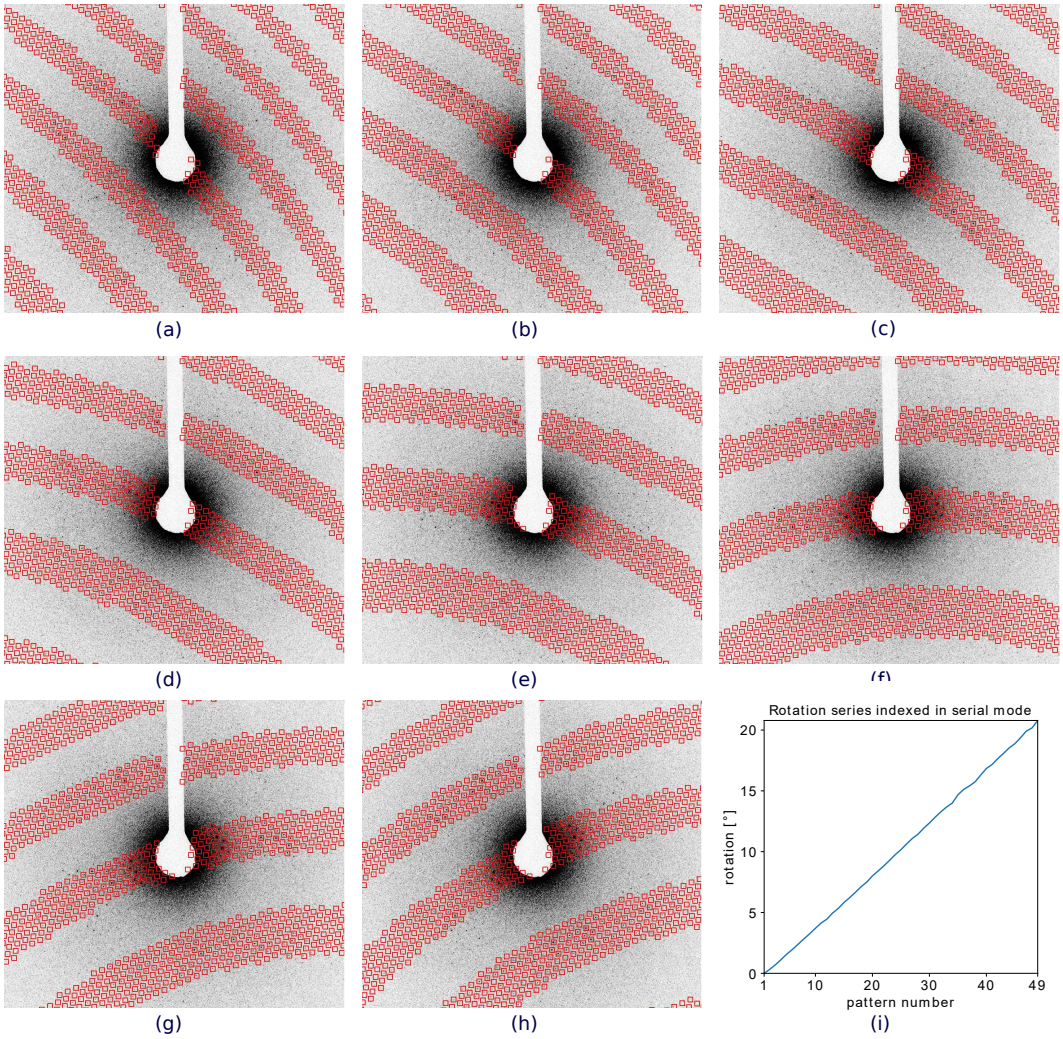


Figure 5.8: *(a) - (h) Patterns 30–37 from the electron diffraction rotation series data of Proteinase K in (Cruz et al., 2017). The red squares mark the locations of the predicted Bragg spots after indexing. (i) Plot of the crystal rotation angle as derived from the indexing result of the pinkIndexer.*

Chapter 6

Indexing in convergent beam serial crystallography

6.1 Introduction

In chapter 4, it has been found that it is highly beneficial to record a large part of the reciprocal space in each diffraction pattern in serial crystallography experiments. So far, we have only considered a polychromatic beam to achieve the required thickness of the Ewald sphere. Another option to tune the beam for a thick Ewald sphere is the beam convergence[1]. Instead of Ewald spheres with different radii, it is possible to generate Ewald spheres with the same radius but different centers, i.e. a superposition of flat waves of the same wavelength but different propagation directions. A simple example of such a setup is a convergent beam that is created from a monochromatic parallel beam by the aid of a spherical lens.

Convergent beam diffraction has been used with electrons for over half a century (Kossel et al., 1939; MacGillavry, 1940), mostly on very thin samples for materials science (Vijayalakshmi et al., 2003). Recent developments in X-ray optics (Murray et al., 2019) made large convergence angles of up to 2° possible and thus enabled interesting applications of a convergent beam for serial X-ray crystallography. Two applications are of primary interest (Spence et al., 2014): The first one aims to exploit the benefits of a thick Ewald sphere on inherently monochromatic sources such as free-electron lasers (FELs). The hope here is to drastically reduce the number of required diffraction patterns for a full dataset due to easier partiality estimation. The second application is only possible if the convergent beam is coherent, i.e. if waves from different directions have a known phase offset. Similar to the pink beam case, Bragg spots will partially overlap

[1]The term beam divergence is also often used to describe this effect.

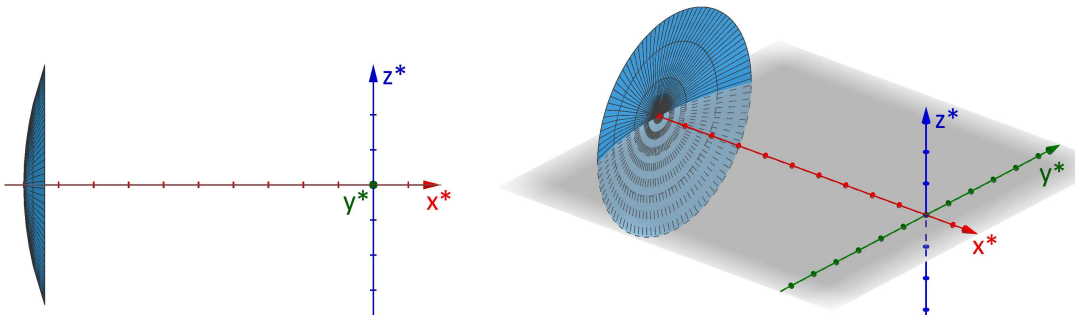


Figure 6.1: *Two views of Ewald spheres'* centers *of a rotationally symmetric monochromatic convergent beam. The coordinate system is selected such, that the center ray of the convergent beam cone points in the direction of the x-axis. The beam is monochromatic, so the Ewald spheres all have the same radius of* $\frac{1}{\lambda}$*. The centers of Ewald spheres available in the beam thus are on the intersection of a cone along the* x^**–axis with an apex angle equal to the convergence angle and a sphere around the origin with radius* $\frac{1}{\lambda}$*. The figure shows two views of the intersection which is a spherical cap.*

due to the thickness of the Ewald sphere. Contrary to the pink beam case (see equation 4.2), these overlaps are coherent and thus encode the phase difference of the Bragg spots. The knowledge of this phase difference can help to solve the phase problem, as has been studied in electron diffraction (Mansfield, 1989).

In this chapter, the geometry of the convergent beam diffraction is explained and it is shown that a slight modification of the pinkIndexer algorithm enables it to index convergent beam diffraction data.

6.2 Convergent beam diffraction geometry

Here, the coordinate system is selected such that the center ray of the beam cone points in the direction of the x-axis. The beam is assumed to be rotationally symmetric around the x-axis, i.e. it is a perfect cone. The beam is monochromatic, so the Ewald spheres all have the same radius of $\frac{1}{\lambda}$. The centers of the Ewald spheres that are available in the beam thus are on the intersection of a cone along the x^*–axis with an apex angle equal to the convergence angle and a sphere around the origin with radius $\frac{1}{\lambda}$. The intersection is a spherical cap, see figure 6.1.

The resulting thick Ewald sphere is rotationally symmetric around the x^*-axis. Figure 6.2 shows a 2D central slice of the thick Ewald sphere. The slice through the spherical cap that marks all Ewald sphere centers available in the

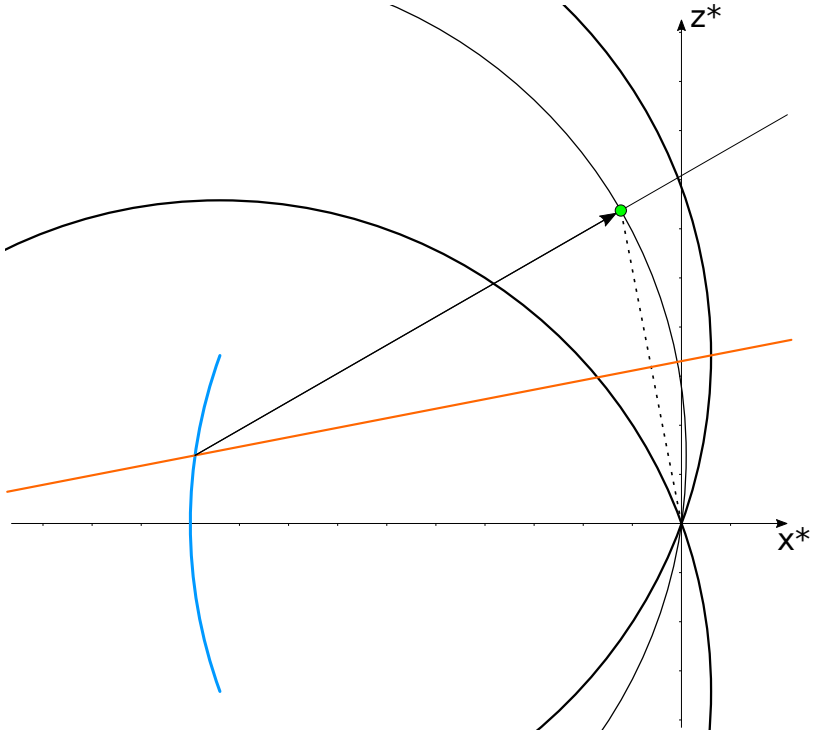


Figure 6.2: *Center slice through the rotationally symmetric thick Ewald sphere of a convergent monochromatic beam. The light-blue circular arc marks the centers of Ewald spheres available in the beam. The green point is an excited RLP, and the orange line is the bisector of this RLP and the origin. The bisector marks all positions of possible Ewald spheres that excite the RLP. The intersection of the orange line and the light-blue circular arc is the center of the Ewald sphere that excites the RLP. The wave vector of the diffracted wave is sketched as a vector from the Ewald sphere center to the RLP.*

beam is sketched as a light-blue circular arc. Given a reciprocal lattice point in the thick Ewald sphere (green point in figure 6.2), all centers of Ewald spheres that excite this RLP must be located on the bisector of the line segment between the RLP and the origin. That is true because, by definition, only points on the bisector have the same distance to the origin and the RLP and thus define centers of Ewald spheres that excite the RLP. Figure 6.2 shows the bisector as an orange line in the 2D slice. All Ewald spheres that excite the RLP are located on the intersection of the bisector and the spherical cap. Contrary to the pink beam case, it is not sufficient to only consider the 2D sketch where the intersection is a point. In the convergent beam case, the Ewald spheres intersect among each other in 3D, and thus, usually more than one Ewald sphere excites the RLP.

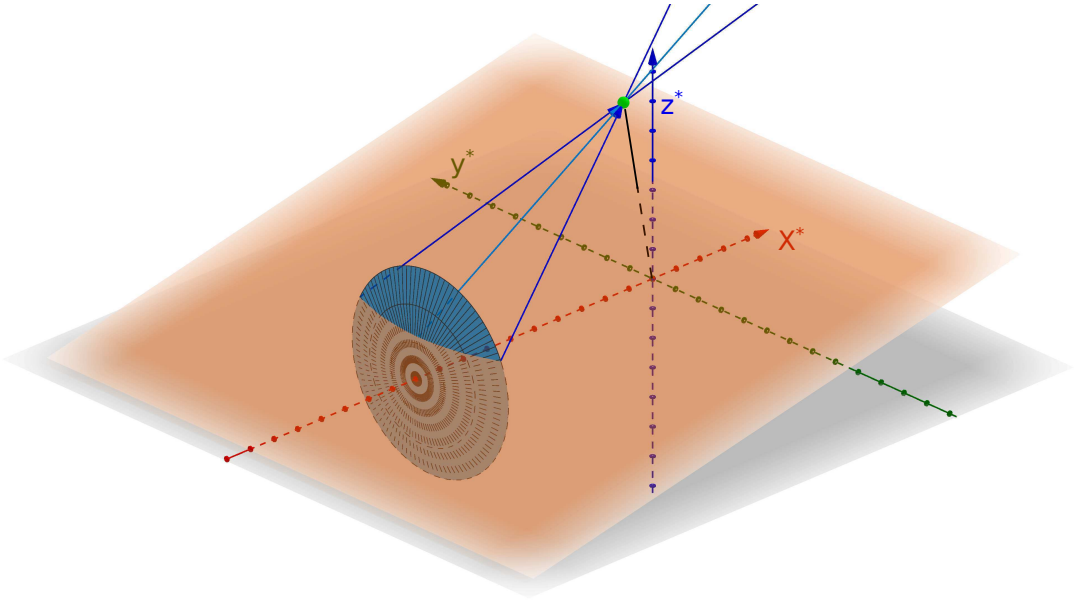


Figure 6.3: *Calculation of the shape of a Bragg spot generated by an RLP in the thick Ewald sphere of a convergent beam. The orange plane is the bisector of the RLP and the origin. It marks all centers of Ewald spheres that excite the RLP. The light-blue spherical cap marks all centers of Ewald spheres available in the beam. The intersection of the bisector and the spherical cap is a circular arc that marks all Ewald sphere centers of the beam that excite the RLP. The resulting Bragg spot is a curve, three points of this curve are marked by the light and dark blue rays.*

In 3D, the bisector of two points is a plane. The bisector of the green RLP and the origin is marked as an orange plane in figure 6.3. The intersection of the bisector and the spherical cap is a circular arc. Note that this is a different circular arc than the one in figure 6.2. All points of this circular arc are centers of Ewald spheres that excite the RLP. Each Ewald sphere center on this circular arc thus defines an outgoing wave vector with the RLP (see arrows in figure 6.3). All outgoing wave vectors have different directions, so contrary to the pink beam case, a reciprocal lattice point of a perfect crystal in a convergent beam does not create a point-shaped Bragg spot on a 2D detector. It is evident from figure 6.3 that the shape of the point must be a continuous curve. To identify the exact shape of the curve, the circular arc of the intersection is expanded to a full circle. This simulates a convergence of 2π, i.e. the beam covers all possible directions - which, of course, is far from realistic but leads to useful insights.

The purple circle in figure 6.4 (a) shows all centers of Ewald spheres of a convergent monochromatic beam of maximum convergence that excite the green

RLP. It is the intersection of the bisector and a sphere of radius $\frac{1}{\lambda}$ around the origin. Figure 6.4 (b) additionally shows some diffracted wave vectors that are generated by Ewald spheres with centers on the purple circle. From the construction, it is clear that the rays are located on a cone.

To simulate the Bragg spot shape, the rays have to be sketched in real space[2], this is done in figure 6.5. Due to the symmetry in the construction, the purple circle can be drawn in the real space as well. A ray from a point on the purple circle through the RLP in reciprocal space corresponds to a ray from the origin of the real space to the opposite[3] point on the purple circle in real space. The ray in the real space is the shifted version of the ray in reciprocal space. From here, it is clear that the set of diffracted rays in the real space has the shape of a cone that has its top at the origin and sides defined by the purple circle. This allows computing the exact shape of the Bragg spot: For a planar detector perpendicular to the x-axis, the form of the Bragg spot is the intersection of the diffracted cone and a plane, i.e. a (part of a) parabola, hyperbola or an ellipse. For reasonably small convergence angles of the beam, the Bragg spot will have the form of a hyperbola (see figure 6.6), since for a parabola or ellipse the Bragg spot has to be far in the positive x^* region of the reciprocal space, which is only reachable with an unrealistically large convergence angle.

For currently achievable convergence angles of approximately 2°, only the very bottom part of the hyperbola is visible, which is almost indistinguishable from a straight line. Let RLP' be the vector from the origin to the RLP projected on the y^*-z^*-plane, then the visible line on the detector is orthogonal to RLP'. In general, for a rotationally symmetric convergent beam as it is assumed it to be, the Bragg spot shape is symmetric to the plane spanned by RLP' and the x^*-axis.

6.3 Indexing convergent beam diffraction patterns

As with every serial crystallography experiment, the diffraction patterns need to be indexed to be able to merge them. In the simple case of a parallel monochromatic beam, there is only one Ewald sphere. Consequently, for one point on the detector, there exists only one point (i.e. a *zero-dimensional* object) in reciprocal space that can cause diffraction in the respective direction. For a pink beam, the Ewald sphere centers available in the beam are located on a line

[2]The diffraction directions are the same in both spaces, so the rays simply have to be sketched from the origin of the real space.

[3]i.e. point-reflected by the center of the circle

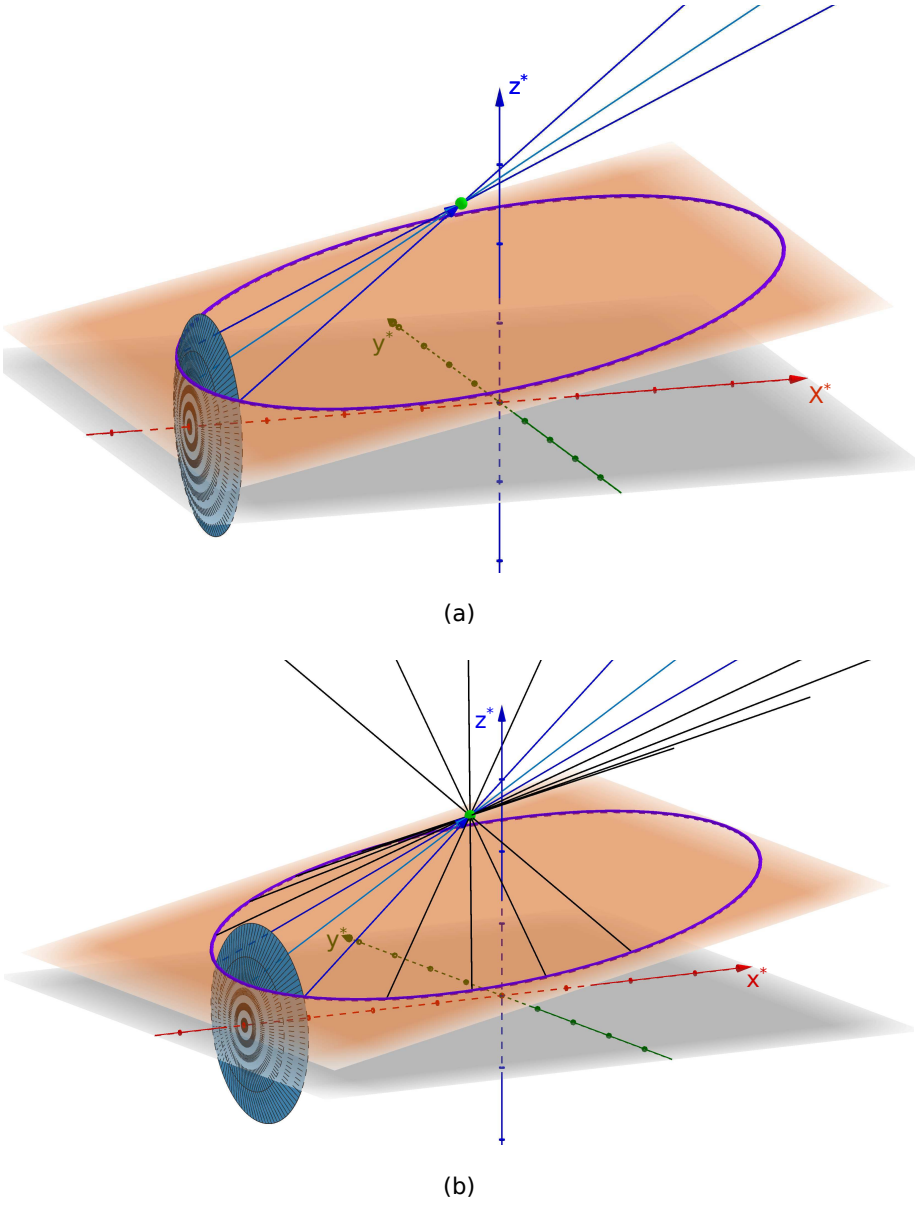


Figure 6.4: *Extension of the circular arc of Ewald sphere centers, that excite the RLP, to a whole circle. The purple circle in (a) and (b) shows all centers of Ewald spheres of a convergent monochromatic beam of maximum convergence that excite the green RLP. This purple circle is the intersection of the bisector and a sphere of radius* $\frac{1}{\lambda}$ *around the origin. (b) additionally shows exemplary diffracted wave vectors that are generated by Ewald spheres with centers on the purple circle. The rays form a cone.*

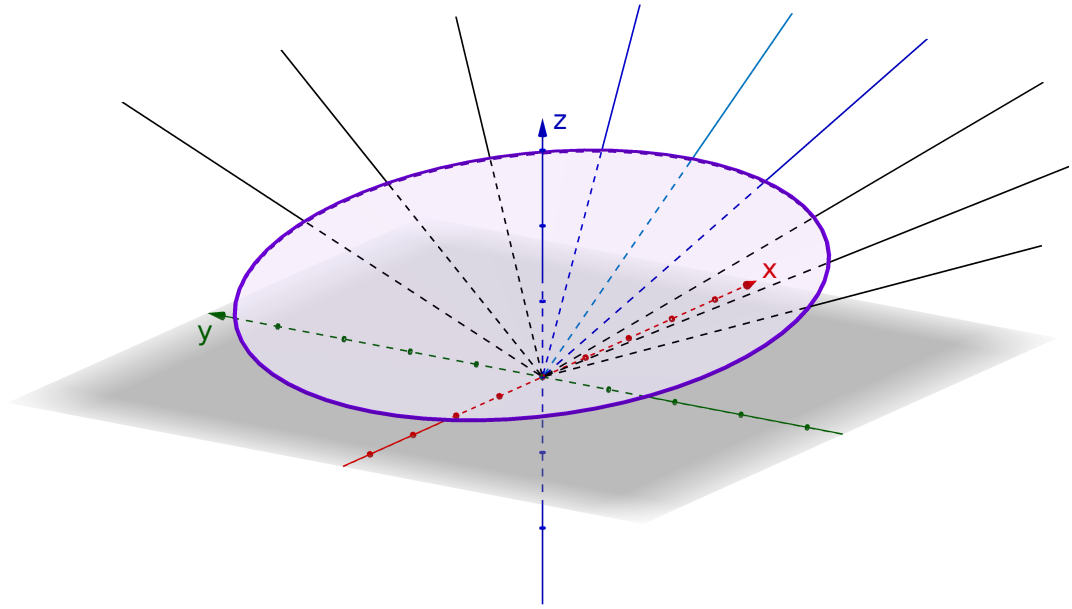


Figure 6.5: *Real space representation of the rays in figure 6.4. This representation allows simulating the Bragg spot shape. It shows the directions in which nonzero intensity can be measured. Due to the symmetry in the construction, the purple circle from figure 6.4 can be drawn in the real space as well. Other than in figure 6.4, the rays don't start on the circle but at the origin. They intersect the point on the circle at the opposite (i.e. inverted) position of the starting position in figure 6.4. Through this symmetry in the construction, the rays in the real space form a cone as well. The light and dark blue rays correspond to the light and dark blue rays in the previous figures.*

segment, consequently, for a point on the detector, there are many points in the reciprocal space that can cause diffraction in the relevant direction. They all are located on the uncertainty line segment, i.e. a *one-dimensional* object. This uncertainty significantly complicates the indexing and led to the development of the pinkIndexer algorithm. In the convergent beam case it can be easily seen that all points of the reciprocal space that cause diffraction in a specific direction are located on a spherical cap - let's call it the uncertainty cap. It will later be shown that this is a shifted version of the spherical cap of Ewald spheres that are available in the convergent beam. The uncertainty cap is a surface, i.e. a *two-dimensional* object. At first glance, it is tempting to think that in this case, due to the increase in dimensionality indexing becomes even more complicated, but it is the other way round. The reason for this is the extended Bragg spot shape. As the previous section showed, the Bragg spot in the convergent beam

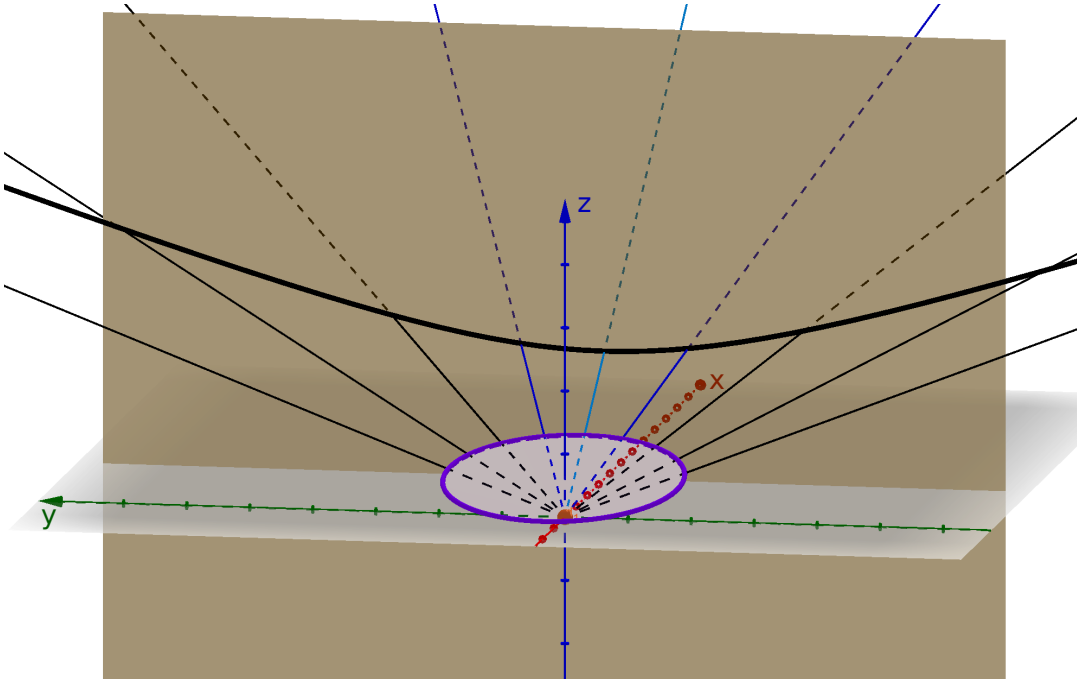


Figure 6.6: *Intersection of the diffracted cone and a planar detector orthogonal to the x-axis. Nonzero diffraction can be measured on the surface of a cone. If the diffraction is measured with a planar detector orthogonal to the x-axis, the measured Bragg spot has the shape of the intersection of the detector and the diffracted cone - in the typical case a (part of a) hyperbola (see the black line). If not maximum divergence is available, the measurable part of the hyperbola is reduced. The dark and light-blue rays show the measurable part for the convergent beam marked by the light-blue spherical cap in the previous figures. For currently achievable convergence angles of* $\approx 2°$*, only the very bottom part of the hyperbola is visible, which is almost indistinguishable from a straight line.*

case is actually a curve - let's call it a Bragg curve. The following paragraphs will show that a Bragg curve contains much more information than a single point, and this information can be exploited for indexing.

An RLP that gives rise to a Bragg curve must give rise to each point on the Bragg curve. This means that the uncertainty caps of all points on the Bragg curve must intersect in the RLP. By inspection it can be seen that all uncertainty caps of a Bragg curve intersect in exactly two points: the RLP and the origin. To find these intersections, the intersection of at least three uncertainty caps need to be computed. Two uncertainty caps are not sufficient, as they intersect in a spherical arc. A third one is needed to define an intersection point. Three points on the Bragg curve thus allow reconstructing the exact location of the

RLP that caused the Bragg curve. Given the 3D locations of the RLPs for the Bragg curves, the indexing is simplified to finding a lattice that fits to a set of points in 3D. This can be done with the XGANDALF algorithm or one of the many algorithms developed for conventional crystallography.

Unfortunately, this only holds for extremely low noise experiments. Uncertainties in the detector geometry, imperfect crystals and quantization noise of the pixel array detector make it necessary to choose the three points on the Bragg curve very far away from each other and thus require extremely high convergence angles. This is the case because the uncertainty caps of points that are close to each other on the detector are very similar. Thus small changes in the position of the points can cause large changes in the intersection of the caps. In practice, the full exploitation of the Bragg spot shape is not feasible, yet.

Instead of exploiting the whole shape of the Bragg curve, its symmetry can be exploited. The Bragg curve is mirror-symmetrical to the plane defined by the RLP and the x^*-axis. As a result, finding the center of the Bragg curve by the noise-resistant center-of-mass calculation allows identifying the plane spanned by the RLP and the x^*-axis. For the identification of the RLPs position within this plane, it is now sufficient to consider the intersection of the thick Ewald sphere and this plane. As can be seen in figure 6.2, the RLP in this plane gives rise to exactly one point on the detector - the intersection of this plane and the Bragg curve. Given the direction of the outgoing wave corresponding to this point of intersection, the points in the reciprocal space that can cause diffraction in this direction can be determined. This is done in the same way as in the pink beam case (see figure 5.2) by appending the wave vectors in the direction of the measured Bragg spot to the centers of Ewald spheres available in the beam. Figure 6.7 shows the determined points in green. They are located on a circular arc. In the style of the ULS in the pink beam model, let's call this arc the "uncertainty circular arc" (UCA), as it is uncertain which point on this arc gave rise to the measured diffraction. With this approach of exploiting only the symmetry instead of the whole shape of the Bragg curve, we lose information that manifests in a less well-defined position of the RLP, but we gain noise tolerance, which is important for real experiments.

Given the UCA for all Bragg spots, a very similar algorithm to the one used for pinkIndexer can be applied to index the pattern. The only difference to the pinkIndexer is that unlike the ULS in the pink beam model, the UCA in the convergent beam model is not on a line that passes the origin. This means that the rotation axis $\hat{\mathbf{m}}$ needs to be recomputed for each candidate RLP dependent on its distance to the origin, i.e. the pre-computation of the parameters d_1 and $\mathbf{d}_2$ in equation (5.3) is not possible. Apart from slightly increasing the execution time, this change does not affect the rest of the algorithm, so a major portion of

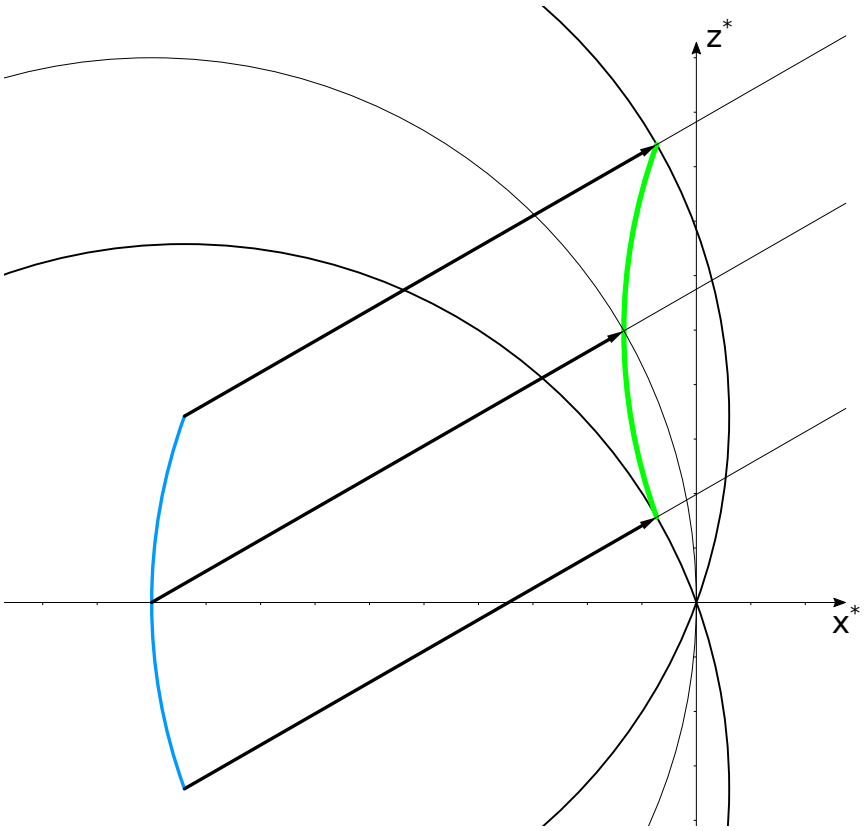


Figure 6.7: *Derivation of the uncertainty circular arc (UCA) in the reciprocal space of a convergent beam experiment. The figure shows a slice through the reciprocal space. The slice is selected such that it is a mirror plane for the diffracted Bragg curve. Here, the RLP is located in the x^*-z^*-plane. The arrows are wave vectors that point in the direction of the intersection of the slice with the Bragg curve. The arrows mark three examples of Ewald spheres and the respective excited points of the reciprocal space that can cause diffraction in the measured direction. The green curve is the UCA. It marks all points in the thick Ewald sphere that can cause diffraction in the measured direction. The UCA is a shifted version of the light-blue curve that marks the centers of all Ewald spheres that are available in the convergent beam.*

the pinkIndexer code can be reused.

The required changes in the pinkIndexer code were implemented to create the convergentIndexer. As a small-convergence-angle approximation, the circular arc shape of the UCA was replaced by a vertical line. For small angles, the introduced error is negligible, but it reduces the program execution time and code complexity. The program was successfully tested on simulated data, see e.g.

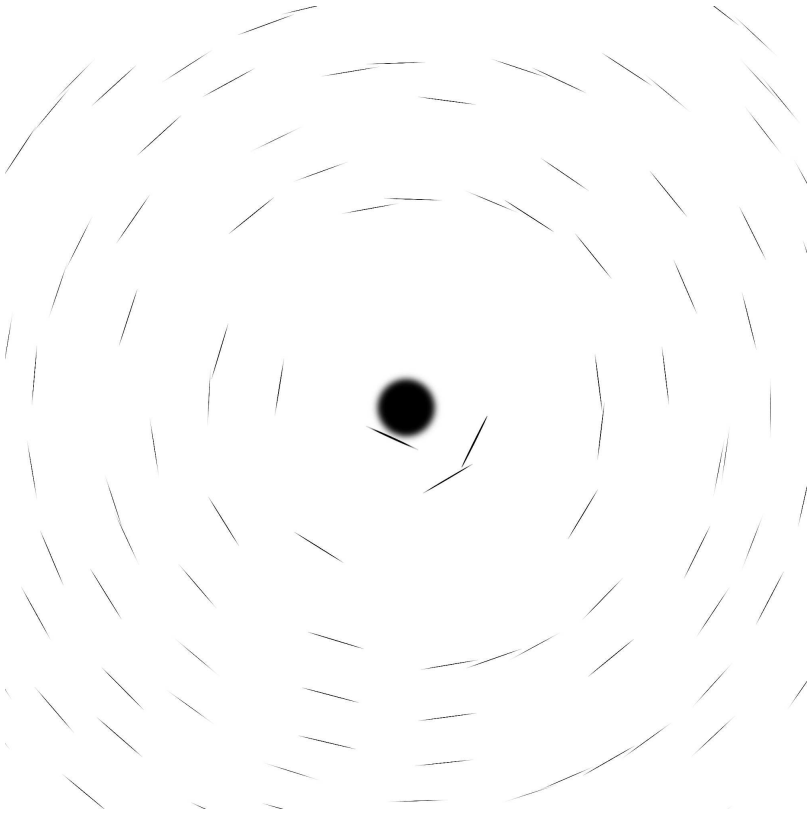


Figure 6.8: *Simulated diffraction pattern with* 1.15° *convergence radius. The diffracted radiation is concentrated in Bragg curves instead of Bragg spots. The convergence angle is not large enough to see the hyperbolic shape of the Bragg curves. The large spot in the center is the direct convergent beam. This pattern can be indexed by the convergentIndexer.*

figure 6.8. The lattice identified during the indexing process corresponded to the one used for simulating the diffraction pattern. Analysis is therefore ready for experimental data. The performance on real data is assumed to be similar to the pinkIndexer.

6.4 Conclusion

Convergent beam diffraction is a powerful technique, as it allows the inherently monochromatic free-electron lasers to benefit from a thick Ewald sphere. Moreover, it has the potential to facilitate the phasing problem, as different Bragg curves can coherently overlap and thus reveal information about their phase difference. Recent developments in X-ray optics (Murray et al., 2019)

enabled a significantly large convergence angle to make this technique useful for crystallography.

With a model for the thick Ewald sphere of a convergent monochromatic beam, it has been shown that the Bragg spot caused by a reciprocal lattice point of a perfect crystal is not a point as in conventional monochromatic and pink beam experiments but rather a curve. It has further been shown that the Bragg curve on a planar detector has the shape of a part of a conic section. In the typical experimental setup where the detector is orthogonal to the central ray of the beam and convergence of a few degrees, the Bragg curve has the shape of a hyperbola.

It has been shown that utilizing the information about the shape of the Bragg curve, the position of the respective RLP can, in theory, be unambiguously determined. This allows indexing noise-free still diffraction patterns in the conventional way. Due to the extremely low noise tolerance, the technique of exploiting the shape of the Bragg curve is not applicable in real experiments.

A way to achieve noise-tolerant exploitation of the Bragg curve shape is to exploit its symmetry. This gain in noise tolerance is paid for by a larger uncertainty about the RLP location. It has been shown that this uncertainty can be handled by a slight adaptation of the pinkIndexer algorithm. The algorithm has been implemented and successfully tested on simulated data.

Chapter 7

Conclusion

7.1 Contribution of this thesis to the current research

Even after a century of highly successful application, crystallography remains a driving force in structural biology and other fields that require knowledge about molecular structure. Most of today's known biological molecule structures have been solved by the conventional crystallography approach of rotating a large crystal in the radiation beam to record diffraction from all orientations. This approach is well studied, but the maximum radiation dose limit restricts it to the difficult-to-produce large crystals.

In the recent decade, the method of serial crystallography has been developed (Chapman et al., 2011). Instead of rotating a crystal in the beam, the total radiation dose is spent on one still diffraction pattern. This method not only significantly decreased the required crystal size, but it also allows us to exceed the maximum dose limit by employing the diffraction before destruction principle with ultrashort pulses, e.g. with X-ray free-electron lasers. Moreover, the development of serial crystallography facilitated time-resolved measurements for biologically interesting timescales, thereby enabling the direct measurement of fast chemical reactions.

A large part of the data processing required for serial crystallography can be inherited from conventional crystallography. Nevertheless, two stages of the data processing pipeline are significantly more difficult with serial crystallography: The indexing stage and the integration/merging stage. The indexing is complicated by the lack of different views on the same crystal, and the integration/merging step is complicated by the significantly reduced noise-tolerance in the partiality estimation due to the lack of crystal movement during the exposure.

Indexing is a necessary step in the data processing pipeline for serial crystallography experiments. Diffraction patterns that cannot be indexed do not contribute to the final result, i.e. the respective crystal and experimental time is wasted. The lack of an indexer for a serial crystallography method completely prevents the data of an experiment to be processed. This Ph.D. thesis entirely concentrated on the development of indexing algorithms for serial crystallography and thus improving or even enabling emerging methods in this field.

To tackle the indexing problem for X-ray serial crystallography, indexing algorithms originally devised for conventional crystallography have been reused. Although in general successful, the indexing results clearly leave room for improvement. Especially patterns with a low number of detected Bragg spots suffer from low indexing rates, thus reducing the number of utilizable diffraction patterns.

The indexing algorithm **XGANDALF** (see chapter 3) was explicitly designed for indexing still monochromatic X-ray diffraction patterns for serial crystallography experiments. It outperforms the current state-of-the-art indexers that were mostly adapted from conventional crystallography. Compared with those programs, XGANDALF gives higher indexing rates and higher indexing precision and can be used both with and without prior unit-cell parameters. The execution time of the implementation is of the same order of magnitude as currently-used indexing algorithms and, with mean indexing times of about 20 ms, is fast enough to allow real-time feedback in experiments. Compared to the available indexers, the algorithm successfully indexes more patterns in test serial crystallography data sets and is more robust to multiple lattices in a single diffraction pattern, see section 3.3. The program has already been used in serial crystallography experiments by several other groups with very positive results.

To tackle the second disadvantage that came up with serial crystallography, the partiality problem, the pink beam serial crystallography has been invented. It has been shown that the use of a wide bandwidth beam can drastically reduce the number of required diffraction patterns for a complete dataset (Meents et al., 2017). Besides the positive effect on the partiality estimation, a pink beam uses the synchrotron radiation more efficiently and lowers the exposure time that limits the time resolution in time-resolved experiments. Although having great benefits for the partiality estimation, pink beam serial crystallography suffered from the lack of an automatic indexing algorithm that prevented its efficient exploitation. With state-of-the-art methods, each diffraction pattern needs manual interaction for successful indexing.

The indexing algorithm **pinkIndexer** (see chapter 5) has been developed for pink beam serial crystallography. The pinkIndexer program overcomes difficulties previously encountered in automatically analyzing thousands of pink

beam diffraction patterns. This removes the bottleneck in processing the data from pink beam serial diffraction experiments. Additionally, the generality of the algorithm also makes it useful for indexing monochromatic serial crystallography diffraction patterns. In this case, pinkIndexer demonstrates a superior success rate in indexing diffraction patterns, especially for the tricky case of a small number of detected Bragg spots. The main limitation of pinkIndexer is its longer execution time compared to many other existing algorithms for monochromatic diffraction analysis and the requirement of knowing the cell parameters of the studied crystals. The benefit, however, is its higher success rate in indexing snapshot diffraction patterns than all other algorithms tested.

Not all radiation sources can produce a pink beam and thus benefit from its positive effects. Prominent examples of such radiation sources are most of the free-electron lasers that enable the use of the diffraction before destruction principle. Recent developments in X-ray optics (Murray et al., 2019) enabled the creation of a convergent beam that, similarly to the pink beam method, can facilitate the partiality problem for these sources. Moreover, the convergent beam might facilitate solutions to the phase problem, which is a long-lasting problem in crystallography. Similar to the pink beam method, convergent beam serial crystallography lacked an indexing algorithm and thus could not be exploited.

The geometry for convergent beam serial crystallography has been analyzed in this thesis. It has been shown that utilizing the information about the shape of the Bragg curve theoretically allows indexing noise-free still diffraction patterns in the conventional way. It has been found that a practical way to achieve noise-tolerant exploitation of the Bragg curve shape is to exploit its symmetry. This gain in noise-tolerance is paid for by a significant loss of information that the indexing algorithm can exploit. It has been shown that this loss of information can be handled by a slight adaptation of the pinkIndexer algorithm. The algorithm has been implemented and successfully tested on simulated data. The resulting **convergentIndexer** (see chapter 6) enables the indexing of serial crystallography diffraction patterns recorded with a rotationally symmetric convergent beam.

The adaption of the pinkIndexer algorithm to the convergent beam case not only enables the key data processing step of indexing for an emerging technique, but it also demonstrates the generality of the pinkIndexer algorithm. This generality anticipates the applicability of the pinkIndexer for future experimental methods.

Serial crystallography started as a method for X-ray radiation but recently has been transferred to electron beams. The use of electrons instead of X-rays has the benefit of a higher ratio of elastic to inelastic scattering and significantly

lower costs for the radiation source. Instead of national-scale facilities such as synchrotrons or FELs, lab-scale machines such as transmission electron microscopes (TEMs) can be efficiently used for serial crystallography. Although being monochromatic and having the option of a parallel beam, serial electron crystallography as well suffered from a lack of indexing algorithms. The extremely short wavelength of the electron beam prevented the use of conventional indexing algorithms.

It was shown that the **pinkIndexer** algorithm works well for indexing serial electron crystallography data. This enabled the first successful data processing for serial electron crystallography data (Bücker et al., 2020).

From the two large disadvantages in serial crystallography, i.e. complicated indexing and merging, the merging step has been tackled by the pink beam and convergent beam methods. The electron beam method has tackled the ease of applicability of serial crystallography. All three methods were not operable due to unsolved indexing problems. Solving the indexing problem for all of these methods, this thesis **increased the yield** for monochromatic crystallography and **enabled** data processing three upcoming and highly promising methods for pink beam, convergent beam, and electron beam serial crystallography.

7.2 Future perspectives

The implementation of XGANDALF allows to easily put blocks together to develop new indexers for monochromatic serial crystallography. The currently used algorithm is only one of the various possible implementations and is not optimal for all cases. This implementation is a compromise between computation time and accuracy, balanced towards accuracy. Using the XGANDALF framework, indexers can be developed that exceed the performance of the current algorithm in particular for corner cases like high-throughput screening of large numbers of crystals.

Although highly optimized, the execution time of the pinkIndexer program is not real-time capable for current experiment setups with the currently available computation resources. It is worthwhile to examine improvements of the execution time by reordering the data structure of the rotogram to allow better performance with the required memory-access pattern. More improvements can be achieved by carefully exploiting symmetries in the applied trigonometric functions and a cache-friendly reordering of the algorithm by splitting the rotogram into smaller parts.

Currently, all Bragg spots are equally weighted in the rotogram of the pinkIndexer. A suitable weighting of the Bragg spots e.g. by intensity or resolu-

tion may increase the indexing rate by cost of a significantly increased execution time. For particularly difficult cases, this option might be favorable.

Another way of thinking about indexing with a known unit cell is to see it as a special case of the rigid registration problem. Several solutions to this problem have been developed in recent years (Li et al., 2018; Parra Bustos et al., 2014). Comparing these solutions to the pinkIndexer approach can lead to improvements in either of the fields.

A large part of the motivation for the pink beam and convergent beam methods is the facilitated scaling and merging step due to the thick Ewald sphere. This, in theory, should allow reconstructing structures from significantly fewer diffraction patterns than in monochromatic SX. Nevertheless, with current state-of-the-art open source software, the number of required patterns still is relatively high (Tolstikova et al., 2019). The main problem is that overlapping peaks are discarded and that the partiality computation is still rudimentary. To take full advantage of the thick Ewald sphere, an advanced partiality model that is suitable for a thick Ewald sphere needs to be applied and incorporated in the post-refinement step that uses all diffraction patterns to refine the extracted data from every single pattern. Apart from that, overlapping Bragg spots need to be disentangled using either the known methods (Ren et al., 1995; Shrive et al., 1990) or more sophisticated ones. This is especially important for very large bandwidth experiments.

As pink beam SX experiments become more popular, new beamlines will be planned that allow utilizing large-bandwidth X-ray beams. Here, care needs to be taken for the design of the spectrum shape. From the indexing point of view, neither a narrow nor a wide spectrum pose a problem anymore. For the partiality calculation, wider spectra and thus thicker Ewald spheres are favorable as well as uniform and smooth spectra, i.e. ones with very small intensity gradients over the spectrum. For the problem of overlapping spots and for experiments with high background noise, a narrow spectrum is preferred. Taking an optimistic point of view, where the data processing part will significantly improve in the next few years and careful experiment setups will reduce the background noise, the trade-off should be weighted towards wide and smooth spectra for the new beamlines. This is opposite to the common undulator development that concentrated on sharp spectra and thus needs significant efforts in beamline design.

For convergent beam SX experiments the indexing problem is significantly facilitated by a rotationally symmetric beam, so the experiment setup should be designed to allow this kind of beam shape. For experiment setups where a rotationally symmetric beam is not possible but the beam shape is well known, the orientation and length of the Bragg curves can be used to restrict the

uncertainty of the RLP location. Another option to design the experimental setup for easy indexing is to shape the intensity profile of the beam. This allows encoding the position of the RLP in the intensity along the Bragg curve. A simple example of such beam shaping is to enforce zero intensity in specific directions by blocking parts of the beam with metal wires. These zones of zero intensity create characteristic gaps in the Bragg curves that can be used for the identification of the RLP positions.

Two decades ago the use of chirped pulses for Laue crystallography was proposed to allow sub-pulse resolution for synchrotrons (Moffat, 2002). With new X-ray sources and the ability to index still pink beam diffraction patterns, this idea could experience a revival for both synchrotrons and free-electron X-ray lasers. Applications can be e.g. shorter time-resolutions for time-resolved experiments or dose fractionation for radiation damage studies. Moreover, the use of chromatic lenses can generate a spatially moving beam from an energy-chirped pulse. This spatially moving beam can help to outrun secondary radiation damage on relatively large crystals by sweeping the focal spot over the crystal and thus creating a locally shorter pulse duration. These methods need significant effort in both, experimental setup design and data processing.

7.3 Closing words

The indexing algorithms developed in this Ph.D. thesis solve the indexing problem for all serial crystallography methods known to the author at the time of writing this thesis and provide the tools to efficiently tackle upcoming challenges. The resulting programs are all open source. These developments are meant to be used by the serial crystallography community to widen our understanding of the processes of life and nature.

Bibliography

Altmann, S. L. (1989). "Hamilton, Rodrigues, and the Quaternion Scandal". In: *Mathematics Magazine* 62.5, p. 291. DOI: 10.2307/2689481. URL: https://doi.org/10.2307/2689481.

Barty, A. et al. (2014). "Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data". In: *Journal of Applied Crystallography* 47.3, pp. 1118–1131. DOI: 10.1107/s1600576714007626. URL: https://doi.org/10.1107/s1600576714007626.

Bergwerf, H. (2014). *The MolView Project*. http://molview.org/. Accessed: 2019-11-26.

Berman, H. M. et al. (2000). "The protein data bank". In: *Nucleic Acids Res* 28, pp. 235–242.

Beyerlein, K. R. et al. (2017). "FELIX: an algorithm for indexing multiple crystallites in X-ray free-electron laser snapshot diffraction images". In: *Journal of Applied Crystallography* 50.4, pp. 1075–1083. DOI: 10.1107/s1600576717007506. URL: https://doi.org/10.1107/s1600576717007506.

Blakeley, M. P. et al. (2008). "Neutron crystallography: opportunities, challenges, and limitations". In: *Current Opinion in Structural Biology* 18.5, pp. 593–600. DOI: 10.1016/j.sbi.2008.06.009. URL: https://doi.org/10.1016/j.sbi.2008.06.009.

Botha, S. et al. (2015). "Room-temperature serial crystallography at synchrotron X-ray sources using slowly flowing free-standing high-viscosity microstreams". In: *Acta Crystallographica Section D Biological Crystallography* 71.2, pp. 387–397. DOI: 10.1107/s1399004714026327. URL: https://doi.org/10.1107/s1399004714026327.

Boutet, S. et al. (2012). "High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography". In: *Science* 337.6092, pp. 362–364. DOI: 10.1126/science.1217737. URL: https://doi.org/10.1126/science.1217737.

Bücker, R. et al. (2020). "Serial protein crystallography in an electron microscope". In: *Nature Communications* 11.1. DOI: 10.1038/s41467-020-

14793-0. URL: https://doi.org/10.1038/s41467-020-14793-0.

Buerger, M. J. (1940). "The Correction of X-Ray Diffraction Intensities for Lorentz and Polarization Factors". In: *Proceedings of the National Academy of Sciences* 26.11, pp. 637–642. DOI: 10.1073/pnas.26.11.637. URL: https://doi.org/10.1073/pnas.26.11.637.

Campbell, J. W. et al. (1998). "LAUEGEN version 6.0 and INTLDM". In: *Journal of Applied Crystallography* 31.3, pp. 496–502. DOI: 10.1107/s0021889897016683. URL: https://doi.org/10.1107/s0021889897016683.

Carr, P. D. et al. (1993). "The determination of unit-cell parameters from a Laue diffraction pattern". In: *Journal of Applied Crystallography* 26.3, pp. 384–387. DOI: 10.1107/s0021889892012391. URL: https://doi.org/10.1107/s0021889892012391.

Chapman, H. et al. (2011). "Femtosecond X-ray protein nanocrystallography". In: *Nature* 470.7332, pp. 73–77. DOI: 10.1038/nature09750. URL: https://doi.org/10.1038/nature09750.

Chapman, H. N. et al. (2014). "Diffraction before destruction". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1647, p. 20130313. DOI: 10.1098/rstb.2013.0313. URL: https://doi.org/10.1098/rstb.2013.0313.

Clarke, J. (2007). *Insertion Devices*. URL: https://cas.web.cern.ch/sites/cas.web.cern.ch/files/lectures/daresbury-2007/clarke.pdf (visited on 02/12/2012).

Clifton, I. J. et al. (1991). "Experimental strategies in Laue crystallography". In: *Journal of Applied Crystallography* 24.4, pp. 267–277. DOI: 10.1107/s0021889890013863. URL: https://doi.org/10.1107/s0021889890013863.

Cohen, A. E. et al. (2014). "Goniometer-based femtosecond crystallography with X-ray free electron lasers". In: *Proceedings of the National Academy of Sciences* 111.48, pp. 17122–17127. DOI: 10.1073/pnas.1418733111. URL: https://doi.org/10.1073/pnas.1418733111.

Crowther, R. A. (1972). *The Molecular Replacement Method, edited by MG Rossmann*.

Cruz, M. J. de la et al. (2017). "Atomic-resolution structures from fragmented protein crystals with the cryoEM method MicroED". In: *nature methods* 14.4, p. 399. DOI: 10.15785/sbgrid/289. URL: http://data.sbgrid.org/dataset/289.

Dejoie, C. et al. (2013). "Using a non-monochromatic microbeam for serial snapshot crystallography". In: *Journal of Applied Crystallography* 46.3, pp. 791–

794. DOI: 10.1107/s0021889813005888. URL: https://doi.org/10.1107/s0021889813005888.

DePonte, D. P. et al. (2008). "Gas dynamic virtual nozzle for generation of microscopic droplet streams". In: *Journal of Physics D: Applied Physics* 41.19, p. 195505. DOI: 10.1088/0022-3727/41/19/195505. URL: https://doi.org/10.1088/0022-3727/41/19/195505.

Dorset, D. (1992). "Direct phasing in electron crystallography: determination of layer silicate structures". In: *Ultramicroscopy* 45.1, pp. 5–14. DOI: 10.1016/0304-3991(92)90033-g. URL: https://doi.org/10.1016/0304-3991(92)90033-g.

— (1996). "Direct Phasing in Protein Electron Crystallography - Phase Extension and the Prospects forAb InitioDeterminations". In: *Acta Crystallographica Section A Foundations of Crystallography* 52.3, pp. 480–489. DOI: 10.1107/s0108767396001420. URL: https://doi.org/10.1107/s0108767396001420.

Duisenberg, A. J. M. (1992). "Indexing in single-crystal diffractometry with an obstinate list of reflections". In: *Journal of Applied Crystallography* 25.2, pp. 92–96. DOI: 10.1107/s0021889891010634. URL: https://doi.org/10.1107/s0021889891010634.

Gevorkov, Y. et al. (2019). "XGANDALF - extended gradient descent algorithm for lattice finding". In: *Acta Crystallographica Section A Foundations and Advances* 75.5, pp. 694–704. DOI: 10.1107/s2053273319010593. URL: https://doi.org/10.1107/s2053273319010593.

Gevorkov, Y. et al. (2020). "pinkIndexer - a universal indexer for pink-beam X-ray and electron diffraction snapshots". In: *Acta Crystallographica Section A Foundations and Advances* 76.2, pp. 121–131. DOI: 10.1107/s2053273319015559. URL: https://doi.org/10.1107/s2053273319015559.

Gildea, R. J. et al. (2014). "New methods for indexing multi-lattice diffraction data". In: *Acta Crystallographica Section D Biological Crystallography* 70.10, pp. 2652–2666. DOI: 10.1107/s1399004714017039. URL: https://doi.org/10.1107/s1399004714017039.

Ginn, H. M. et al. (2015). "A revised partiality model and post-refinement algorithm for X-ray free-electron laser data". In: *Acta Crystallographica Section D Biological Crystallography* 71.6, pp. 1400–1410. DOI: 10.1107/s1399004715006902. URL: https://doi.org/10.1107/s1399004715006902.

Ginn, H. M. et al. (2016). "TakeTwo: an indexing algorithm suited to still images with known crystal parameters". In: *Acta Crystallographica Section D Structural Biology* 72.8, pp. 956–965. DOI: 10.1107/s2059798316010706. URL: https://doi.org/10.1107/s2059798316010706.

Helliwell, J. R. et al. (1989). "The recording and analysis of synchrotron X-radiation Laue diffraction photographs". In: *Journal of Applied Crystallography* 22.5, pp. 483–497. DOI: 10.1107/s0021889889006564. URL: https://doi.org/10.1107/s0021889889006564.

Hendrickson, W. (1991). "Determination of macromolecular structures from anomalous diffraction of synchrotron radiation". In: *Science* 254.5028, pp. 51–58. DOI: 10.1126/science.1925561. URL: https://doi.org/10.1126/science.1925561.

Hestenes, M. et al. (1952). "Methods of conjugate gradients for solving linear systems". In: *Journal of Research of the National Bureau of Standards* 49.6, p. 409. DOI: 10.6028/jres.049.044. URL: https://doi.org/10.6028/jres.049.044.

Jacobson, R. A. (1986). "An orientation-matrix approach to Laue indexing". In: *Journal of Applied Crystallography* 19.5, pp. 283–286. DOI: 10.1107/s0021889886089343. URL: https://doi.org/10.1107/s0021889886089343.

James, R. (1950). *The Optical Principles of the Diffraction of X-rays*. Crystalline state. Bell. URL: https://books.google.de/books?id=cIsynQEACAAJ.

Kabsch, W. (1976). "A solution for the best rotation to relate two sets of vectors". In: *Acta Crystallographica Section A* 32.5, pp. 922–923. DOI: 10.1107/s0567739476001873. URL: https://doi.org/10.1107/s0567739476001873.

— (1993). "Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants". In: *Journal of Applied Crystallography* 26.6, pp. 795–800. DOI: 10.1107/s0021889893005588. URL: https://doi.org/10.1107/s0021889893005588.

Kabsch, W. (2010). "XDS". In: *Acta Crystallographica Section D Biological Crystallography* 66.2, pp. 125–132. DOI: 10.1107/s0907444909047337. URL: https://doi.org/10.1107/s0907444909047337.

— (2014). "Processing of X-ray snapshots from crystals in random orientations". In: *Acta Crystallographica Section D Biological Crystallography* 70.8, pp. 2204–2216. DOI: 10.1107/s1399004714013534. URL: https://doi.org/10.1107/s1399004714013534.

Kalinowski, J. A. et al. (2011). "TheLaueUtiltoolkit for Laue photocrystallography. I. Rapid orientation matrix determination for intermediate-size-unit-cell Laue data". In: *Journal of Applied Crystallography* 44.6, pp. 1182–1189. DOI: 10.1107/s0021889811038143. URL: https://doi.org/10.1107/s0021889811038143.

Karplus, P. A. et al. (2012). "Linking Crystallographic Model and Data Quality". In: *Science* 336.6084, pp. 1030–1033. DOI: 10.1126/science.1218231. URL: https://doi.org/10.1126/science.1218231.

Kirian, R. A. et al. (2010). "Femtosecond protein nanocrystallography - data analysis methods". In: *Optics Express* 18.6, p. 5713. DOI: 10.1364/oe.18.005713. URL: https://doi.org/10.1364/oe.18.005713.

Kirian, R. A. et al. (2011). "Structure-factor analysis of femtosecond microdiffraction patterns from protein nanocrystals". In: *Acta Crystallographica Section A Foundations of Crystallography* 67.2, pp. 131–140. DOI: 10.1107/s0108767310050981. URL: https://doi.org/10.1107/s0108767310050981.

Kossel, W. et al. (1939). "Elektroneninterferenzen im konvergenten Bündel". In: *Annalen der Physik* 428.2, pp. 113–140. DOI: 10.1002/andp.19394280204. URL: https://doi.org/10.1002/andp.19394280204.

Kroon-Batenburg, L. et al. (2015). "Accounting for partiality in serial crystallography using ray-tracing principles". In: *Acta Crystallographica Section D Biological Crystallography* 71.9, pp. 1799–1811. DOI: 10.1107/s1399004715011803. URL: https://doi.org/10.1107/s1399004715011803.

Li, C. et al. (2019). "SPIND: a reference-based auto-indexing algorithm for sparse serial crystallography data". In: *IUCrJ* 6.1, pp. 72–84. DOI: 10.1107/s2052252518014951. URL: https://doi.org/10.1107/s2052252518014951.

Li, X. et al. (2018). "Fast and globally optimal rigid registration of 3d point sets by transformation decomposition". In: *arXiv preprint arXiv:1812.11307*.

Liu, W. et al. (2013). "Serial Femtosecond Crystallography of G Protein-Coupled Receptors". In: *Science* 342.6165, pp. 1521–1524. DOI: 10.1126/science.1244142. URL: https://doi.org/10.1126/science.1244142.

Lyubimov, A. Y. et al. (2016). "IOTA: integration optimization, triage and analysis tool for the processing of XFEL diffraction images". In: *Journal of Applied Crystallography* 49.3, pp. 1057–1064. DOI: 10.1107/s1600576716006683. URL: https://doi.org/10.1107/s1600576716006683.

MacGillavry, C. H. (1940). "Diffraction of Convergent Electron Beams". In: *Nature* 145.3666, pp. 189–190. DOI: 10.1038/145189a0. URL: https://doi.org/10.1038/145189a0.

Maia, F. R.N. C. (2012). "The Coherent X-ray Imaging Data Bank". In: *Nature Methods* 9.9, pp. 854–855. DOI: 10.1038/nmeth.2110. URL: https://doi.org/10.1038/nmeth.2110.

Mansfield, J. (1989). "Practical phase identification by convergent beam electron diffraction". In: *Journal of Electron Microscopy Technique* 13.1, pp. 3–15.

DOI: 10.1002/jemt.1060130104. URL: https://doi.org/10.1002/jemt.1060130104.

Martin-Garcia, J. M. et al. (2017). "Serial millisecond crystallography of membrane and soluble protein microcrystals using synchrotron radiation". In: *IUCrJ* 4.4, pp. 439–454. DOI: 10.1107/s205225251700570x. URL: https://doi.org/10.1107/s205225251700570x.

McIntyre, G. J. et al. (2006). "High-speed neutron Laue diffraction comes of age". In: *Physica B: Condensed Matter* 385-386, pp. 1055–1058. DOI: 10.1016/j.physb.2006.05.338. URL: https://doi.org/10.1016/j.physb.2006.05.338.

Meents, A. et al. (2017). "Pink-beam serial crystallography". In: *Nature Communications* 8.1. DOI: 10.1038/s41467-017-01417-3. URL: https://doi.org/10.1038/s41467-017-01417-3.

Mezza, D. et al. (2016). "New calibration circuitry and concept for AGIPD". In: *Journal of Instrumentation* 11.11, pp. C11019–C11019. DOI: 10.1088/1748-0221/11/11/c11019. URL: https://doi.org/10.1088/1748-0221/11/11/c11019.

Moffat, K. et al. (1984). "X-ray Laue Diffraction from Protein Crystals". In: *Science* 223.4643, pp. 1423–1425. DOI: 10.1126/science.223.4643.1423. URL: https://doi.org/10.1126/science.223.4643.1423.

Moffat, K. (1997). "[22] Laue diffraction". In: *Methods in Enzymology*. Elsevier, pp. 433–447. DOI: 10.1016/s0076-6879(97)77024-1. URL: https://doi.org/10.1016/s0076-6879(97)77024-1.

— (2002). "The frontiers of time-resolved macromolecular crystallography: movies and chirped X-ray pulses". In: *Faraday Discussions* 122, pp. 65–77. DOI: 10.1039/b201620f. URL: https://doi.org/10.1039/b201620f.

Morawiec, A et al. (2005). "On Algorithms for Indexing of K-line Diffraction Patterns". In: *Archives of Metallurgy and Materials* 50, pp. 47–56.

Morgan, A. J. et al. (2015). "High numerical aperture multilayer Laue lenses". In: *Scientific Reports* 5.1. DOI: 10.1038/srep09892. URL: https://doi.org/10.1038/srep09892.

Murray, K. T. et al. (2019). "Multilayer Laue lenses at high X-ray energies: performance and applications". In: *Optics Express* 27.5, p. 7120. DOI: 10.1364/oe.27.007120. URL: https://doi.org/10.1364/oe.27.007120.

Myles, D. et al. (1997). "Neutron Laue diffraction in macromolecular crystallography". In: *Physica B: Condensed Matter* 241-243, pp. 1122–1130. DOI:

10.1016/s0921-4526(97)00808-9. URL: https://doi.org/10.1016/s0921-4526(97)00808-9.

Nave, C. (2014). "Matching X-ray beam and detector properties to protein crystals of different perfection". In: *Journal of Synchrotron Radiation* 21.3, pp. 537–546. DOI: 10.1107/s1600577514003609. URL: https://doi.org/10.1107/s1600577514003609.

Nogly, P. et al. (2015). "Lipidic cubic phase serial millisecond crystallography using synchrotron radiation". In: *IUCrJ* 2.2, pp. 168–176. DOI: 10.1107/s2052252514026487. URL: https://doi.org/10.1107/s2052252514026487.

Parra Bustos, A. et al. (2014). "Fast rotation search with stereographic projections for 3D registration". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3930–3937.

Pithan, L. (2017). "On the role of external stimuli to tailor growth of organic thin films". PhD thesis. DOI: 10.18452/17749.

Powell, H. R. (1999). "The Rossmann Fourier autoindexing algorithm inMOSFLM". In: *Acta Crystallographica Section D Biological Crystallography* 55.10, pp. 1690–1695. DOI: 10.1107/s0907444999009506. URL: https://doi.org/10.1107/s0907444999009506.

Pujol, J. (2013). "The Rodrigues equations for the composition of finite rotations: a simple ab initio derivation and some consequences". In: *Applied Mechanics Reviews* 65.5, p. 054501.

Raventós, M. et al. (2019). "Laue three dimensional neutron diffraction". In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-41071-x. URL: https://doi.org/10.1038/s41598-019-41071-x.

Ren, Z. et al. (1995). "Deconvolution of Energy Overlaps in Laue Diffraction". In: *Journal of Applied Crystallography* 28.5, pp. 482–494. DOI: 10.1107/s0021889895003219. URL: https://doi.org/10.1107/s0021889895003219.

Ren, Z. et al. (1999). "Laue crystallography: coming of age". In: *Journal of Synchrotron Radiation* 6.4, pp. 891–917. DOI: 10.1107/s0909049599006366. URL: https://doi.org/10.1107/s0909049599006366.

Roedig, P. et al. (2017). "High-speed fixed-target serial virus crystallography". In: *Nature Methods* 14.8, pp. 805–810. DOI: 10.1038/nmeth.4335. URL: https://doi.org/10.1038/nmeth.4335.

Rossmann, M. G. et al. (1962). "The detection of sub-units within the crystallographic asymmetric unit". In: *Acta Crystallographica* 15.1, pp. 24–31. DOI: 10.1107/s0365110x62000067. URL: https://doi.org/10.1107/s0365110x62000067.

Rossmann, M. G. (2014). "Serial crystallography using synchrotron radiation". In: *IUCrJ* 1.2, pp. 84–86. DOI: 10.1107/s2052252514000499. URL: https://doi.org/10.1107/s2052252514000499.

Sauter, N. K. et al. (2009). "Autoindexing the diffraction patterns from crystals with a pseudotranslation". In: *Acta Crystallographica Section D Biological Crystallography* 65.6, pp. 553–559. DOI: 10.1107/s0907444909010725. URL: https://doi.org/10.1107/s0907444909010725.

Schlichting, I. (2015). "Serial femtosecond crystallography: the first five years". In: *IUCrJ* 2.2, pp. 246–255. DOI: 10.1107/s205225251402702x. URL: https://doi.org/10.1107/s205225251402702x.

Schmidt, S. (2014). "GrainSpotter: a fast and robust polycrystalline indexing algorithm". In: *Journal of Applied Crystallography* 47.1, pp. 276–284. DOI: 10.1107/s1600576713030185. URL: https://doi.org/10.1107/s1600576713030185.

Semaev, I. (2001). "A 3-Dimensional Lattice Reduction Algorithm". In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 181–193. DOI: 10.1007/3-540-44670-2_13. URL: https://doi.org/10.1007/3-540-44670-2_13.

Semechko, A. (2015). *Suite of functions to perform uniform sampling of a sphere*. URL: https://de.mathworks.com/matlabcentral/fileexchange/37004-suite-of-functions-to-perform-uniform-sampling-of-a-sphere (visited on 03/26/2015).

Shrive, A. K. et al. (1990). "Laue film integration and deconvolution of spatially overlapping reflections". In: *Journal of Applied Crystallography* 23.3, pp. 169–174. DOI: 10.1107/S0021889890000346. URL: https://doi.org/10.1107/S0021889890000346.

Smeets, S. et al. (2018). "Serial electron crystallography for structure determination and phase analysis of nanocrystalline materials". In: *Journal of Applied Crystallography* 51.5, pp. 1262–1273. DOI: 10.1107/s1600576718009500. URL: https://doi.org/10.1107/s1600576718009500.

Spence, J. C. H. et al. (2004). "Single Molecule Diffraction". In: *Physical Review Letters* 92.19. DOI: 10.1103/physrevlett.92.198102. URL: https://doi.org/10.1103/physrevlett.92.198102.

Spence, J. C. H. et al. (2012). "X-ray lasers for structural and dynamic biology". In: *Reports on Progress in Physics* 75.10, p. 102601. DOI: 10.1088/0034-4885/75/10/102601. URL: https://doi.org/10.1088/0034-4885/75/10/102601.

Spence, J. C. H. et al. (2014). "Coherent convergent-beam time-resolved X-ray diffraction". In: *Philosophical Transactions of the Royal Society B: Biological*

Sciences 369.1647, p. 20130325. DOI: 10.1098/rstb.2013.0325. URL: https://doi.org/10.1098/rstb.2013.0325.

Stellato, F. et al. (2014). "Room-temperature macromolecular serial crystallography using synchrotron radiation". In: *IUCrJ* 1.4, pp. 204–212. DOI: 10.1107/s2052252514010070. URL: https://doi.org/10.1107/s2052252514010070.

Steller, I. et al. (1997). "An Algorithm for Automatic Indexing of Oscillation Images using Fourier Analysis". In: *Journal of Applied Crystallography* 30.6, pp. 1036–1040. DOI: 10.1107/s0021889897008777. URL: https://doi.org/10.1107/s0021889897008777.

Terzakis, G. et al. (2018). "Modified Rodrigues Parameters: An Efficient Representation of Orientation in 3D Vision and Graphics". In: *Journal of Mathematical Imaging and Vision* 60.3, pp. 422–442. ISSN: 1573-7683. DOI: 10.1007/s10851-017-0765-x. URL: https://doi.org/10.1007/s10851-017-0765-x.

Thompson, D. (1996). "The reciprocal lattice as the Fourier transform of the direct lattice". In: *American Journal of Physics* 64.3, pp. 333–334. DOI: 10.1119/1.18243. URL: https://doi.org/10.1119/1.18243.

Tolstikova, A. et al. (2019). "1 kHz fixed-target serial crystallography using a multilayer monochromator and an integrating pixel detector". In: *IUCrJ* 6.5. DOI: 10.1107/s205225251900914x. URL: https://doi.org/10.1107/s205225251900914x.

Unwin, P. et al. (1975). "Molecular structure determination by electron microscopy of unstained crystalline specimens". In: *Journal of Molecular Biology* 94.3, pp. 425–440. DOI: 10.1016/0022-2836(75)90212-0. URL: https://doi.org/10.1016/0022-2836(75)90212-0.

Valafar, H. et al. (2012). "Structure and Dynamics of Proteins from Nuclear Magnetic Resonance Spectroscopy". In: *Protein Structure*. InTech. DOI: 10.5772/38682. URL: https://doi.org/10.5772/38682.

Vijayalakshmi, M. et al. (2003). "Convergent beam electron diffraction - A novel technique for materials characterisation at sub-microscopic levels". In: *Sadhana* 28.3-4, pp. 763–782. DOI: 10.1007/bf02706458. URL: https://doi.org/10.1007/bf02706458.

Wan, W. et al. (2012). "Structure projection reconstruction from through-focus series of high-resolution transmission electron microscopy images". In: *Ultramicroscopy* 115, pp. 50–60. DOI: 10.1016/j.ultramic.2012.01.013. URL: https://doi.org/10.1016/j.ultramic.2012.01.013.

Wang, X. (2008). "Method of Steepest Descent and its Applications". In: *IEEE Microwave Wireless Components Lett* 12.

Weierstall, U. (2014). "Liquid sample delivery techniques for serial femtosecond crystallography". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1647, p. 20130337. DOI: 10.1098/rstb.2013.0337. URL: https://doi.org/10.1098/rstb.2013.0337.

Wenk, H. R. et al. (1997). "Laue Orientation Imaging". In: *Journal of Synchrotron Radiation* 4.2, pp. 95–101. DOI: 10.1107/s090904959601432x. URL: https://doi.org/10.1107/s090904959601432x.

White, T. A. (2019). "Processing serial crystallography data with CrystFEL: a step-by-step guide". In: *Acta Crystallographica Section D Structural Biology* 75.2, pp. 219–233. DOI: 10.1107/s205979831801238x. URL: https://doi.org/10.1107/s205979831801238x.

White, T. A. et al. (2012). "CrystFEL: a software suite for snapshot serial crystallography". In: *Journal of Applied Crystallography* 45.2, pp. 335–341. DOI: 10.1107/s0021889812002312. URL: https://doi.org/10.1107/s0021889812002312.

White, T. et al. (2013). "Crystallographic data processing for free-electron laser sources". In: *Acta Crystallographica Section D Biological Crystallography* 69.7, pp. 1231–1240. DOI: 10.1107/s0907444913013620. URL: https://doi.org/10.1107/s0907444913013620.

Wiedorn, M. O. et al. (2018). "Megahertz serial crystallography". In: *Nature Communications* 9.1. DOI: 10.1038/s41467-018-06156-7. URL: https://doi.org/10.1038/s41467-018-06156-7.

Winter, G. et al. (2018). "DIALS: implementation and evaluation of a new integration package". In: *Acta Crystallographica Section D Structural Biology* 74.2, pp. 85–97. DOI: 10.1107/s2059798317017235. URL: https://doi.org/10.1107/s2059798317017235.

Zurek, S et al. (1985). "Unscrambling of harmonic reflection intensities from spots on Laue patterns: results on pea lectin". In: *Info. Quarterly for Protein Crystallography* 16, pp. 37–40.

Summary

Crystallography allows experimentally gaining information about the atomic structure of molecules by exposing crystals of them to radiation and analyzing the diffraction. The diffraction pattern of one orientation of the crystal in the radiation beam reveals a small part of the crystals 3D Fourier transform. Information from diffraction patterns of many different orientations needs to be merged to recover the full 3D structure. For this, the knowledge of the crystal orientation in the beam is crucial. The identification of the crystal orientation is called indexing. In conventional crystallography, a crystal is rotated in the beam with a known angular velocity. This allows combining information from all diffraction patterns for indexing.

The rotation method poses a limit on the minimum crystal size and the time resolution of the measurement. To overcome both, serial crystallography (SX) has been developed. SX allows using large numbers of small crystals instead of one big crystal. The price for the relaxed requirement on crystal size is significantly more complicated indexing and merging of the diffraction patterns.

This thesis presents the algorithm XGANDALF, it is meant to replace the state-of-the-art indexing algorithms for SX. The algorithm extends the widely used Fourier methods by tuneable basis functions, a non-linear noise filtering, and a heuristic approach to find precise indexing solutions with short program execution times. Rigorous tests show significant improvement in the indexing performance on various datasets for all tested figures of merit. XGANDALF is already widely used by the SX community. The complicated merging in SX is tackled by the new method of pink beam SX. It allows using significantly fewer crystals for the same merge quality and enables time-resolved measurements in cheaper radiation sources than before. Another technique, the convergent beam SX, also simplifies the merging procedure and can be applied to high-performance sources where pink beam SX is not realizable. Popular sources that can benefit from this method are X-ray free electron lasers. The third method, the electron SX, allows using significantly cheaper radiation sources for SX, thus making SX available to a large community and reducing latency for beamtime assignment. These methods have in common that they are highly promising. Many groups have collected data with these methods without the ability to analyze it. The bottleneck is the lack of an automatic indexing algorithm.

This thesis presents the algorithm pinkIndexer, it is capable of indexing all of the above SX methods. It inverts the brute-force approach with a technique that is similar to the Hough and Radon transforms. By enabling effective automatic indexing, this algorithm enables the application of all of the three above techniques. This gives the SX community all over the world new powerful tools to widen our understanding of the processes of life and nature.