Title
Reconstruction algorithms for x-ray nanocrystallography via solution of the twinning problem

Author
Donatelli, Jeffrey J.

Publication Date
2013

Peer reviewed|Thesis/dissertation

Reconstruction algorithms for x-ray nanocrystallography via solution of the twinning problem

by

Jeffrey J. Donatelli

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James A. Sethian, Chair
Professor F. Alberto Grünbaum
Professor Roger W. Falcone

Spring 2013

Reconstruction algorithms for x-ray nanocrystallography via solution of the twinning problem

Abstract

Reconstruction algorithms for x-ray nanocrystallography via solution of the twinning problem

by

Jeffrey J. Donatelli

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor James A. Sethian, Chair

X-ray nanocrystallography is an emerging technique for imaging nanoscale objects that alleviates the large crystallization requirement of conventional crystallography by collecting diffraction patterns from a large ensemble of smaller and easier to build nanocrystals, which are typically delivered to the x-ray beam via a liquid jet. In order to determine the structure of an imaged object, several parameters must first be determined, including the crystal sizes, incident photon flux densities, and crystal orientations. Autoindexing techniques, which have been used extensively to orient conventional crystals, only determine the orientation of the nanocrystals up to symmetry of the crystal lattice, which is often greater than the symmetry of the diffraction information, resulting in what is known as the twinning problem. In addition, the image data is corrupted by large degrees of shot noise due to low collected signal, background signal due to the liquid jet and detector electronics, as well as other sources of noise. Furthermore, diffraction only measures the magnitudes of the Fourier transform of the object and, thus, one must recover phase information in order to invert the data and recover a three-dimensional reconstruction of the constituent molecular structure. Previous approaches for handling the twinning problem have mainly relied on having a known similar structure available, which may not be present for fundamentally new structures. We present a series of techniques to determine the crystal sizes, incident photon flux densities, and crystal orientations in the presence of large amounts of noise common in experiments. Additionally, by using a new sampling strategy, we demonstrate that phase information can be computed from nanocrystallographic diffraction images using only Fourier magnitude information, via a compressive phase retrieval algorithm. We demonstrate the feasibility of this new approach by testing it on simulated data with parameters and noise levels common in current experiments.

Contents

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor Jamie Sethian for his support and guidance, for introducing me to x-ray science, and for being an integral part of both my personal and professional development. I would not be the person that I am today without Jamie.

I would like to thank Stefano Marchesini for introducing me to the x-ray nanocrystallography reconstruction problem, for guiding me through the basics of x-ray imaging, and for several stimulating conversations.

I am grateful to my committee members, consisting of Jamie Sethian, Alberto Grünbaum, and Roger Falcone, for their valuable feedback and suggestions.

I would like to acknowledge the members of the Mathematics Group at LBL, past and present, for making the lab a fun and enlightening place to work. In particular, I would like to thank Michaeel Kazi, Danielle Maddix, Ben Preskill, and Robert Saye for helping me maintain sanity with several needed distractions, including pictography, lexicographic combinatorics, and polyhedral dynamics.

This research was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC02-05CH11231, by the Division of Mathematical Sciences of the National Science Foundation, a Department of Energy Computational Science Graduate Fellowship, and used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Finally, I would like to thank my family for their motivation and support throughout the years.

Chapter 1

Introduction

A traditional method for obtaining high resolution atomic structure information from nanoscale objects is through conventional x-ray crystallography. In this technique, a large number of copies of the target object are arranged into a large, typically about 1 mm in size, periodic crystal structure, in order to increase the strength of the collected signal, and diffraction images are collected from the sample as it is rotated. In general, the pixel intensities of a diffraction pattern measure the magnitude of the three-dimensional Fourier transform of the sample's electron density along a spherical slice in frequency space. Due to the translational property of the Fourier transform, the periodic crystal structure induces the formation of several sharp bright spots of intensity, known as Bragg peaks, whose location and intensity values are used to ultimately invert the data and reconstruct the electron density. The missing phase information in the data is commonly recovered through phasing techniques such as anomalous diffraction, where the wavelength is varied through an absorption edge; isomorphous replacement, which requires a duplicate crystal to be made with the inclusion of heavy atoms in the crystal structure; or molecular replacement, which attempts to modify a previously known structure to match the collected diffraction data.

While conventional x-ray crystallography has been successful in determining the structure of numerous objects, it is limited to samples which can be formed into large crystals, a laborious process that can take up to several years to perform for certain structures. Additionally, the crystal samples are commonly plagued with imperfections that may hinder the reconstruction process. An appealing alternative, made possible by recent advances in light source technology, is x-ray nanocrystallography, which uses a large ensemble of easier to build nanocrystals, typically less than 1 µm in length, which are, for example, delivered to the x-ray beam via a liquid jet [4, 6, 12, 39, 44], as illustrated in Figure 1.1. In particular, x-ray nanocrystallography allows one to image structures which are resistant to large crystallization, such as membrane proteins. However, the beam power density required to retrieve a sufficient amount of signal from nanocrystals is large enough to destroy the crystal during the imaging process. Therefore, ultrafast pulses, typically under 70 fs, are required

to ensure that the data is collected before damage effects come into play.

The use of nanocrystals introduces several practical difficulties into the reconstruction procedure. For instance, due to the small crystal sizes, the Bragg peaks are smeared out and signal between peaks becomes noticeable. Due to the delivery system and short pulses, we cannot integrate out the shape transform via rotational averaging methods, as is done in conventional crystallography. Therefore, only partial peak reflections are typically measured, resulting in reduced and noisy collected intensities. Further sources of uncertainty and error are caused by large variations in crystal sizes, shot noise, background signal introduced by the disordered water molecules in the liquid jet, changes in signal intensity induced by beam fluctuations and partial collisions of the nanocrystals with the x-ray beam, and the fact that orientations of the crystals are unknown during the data collection process.

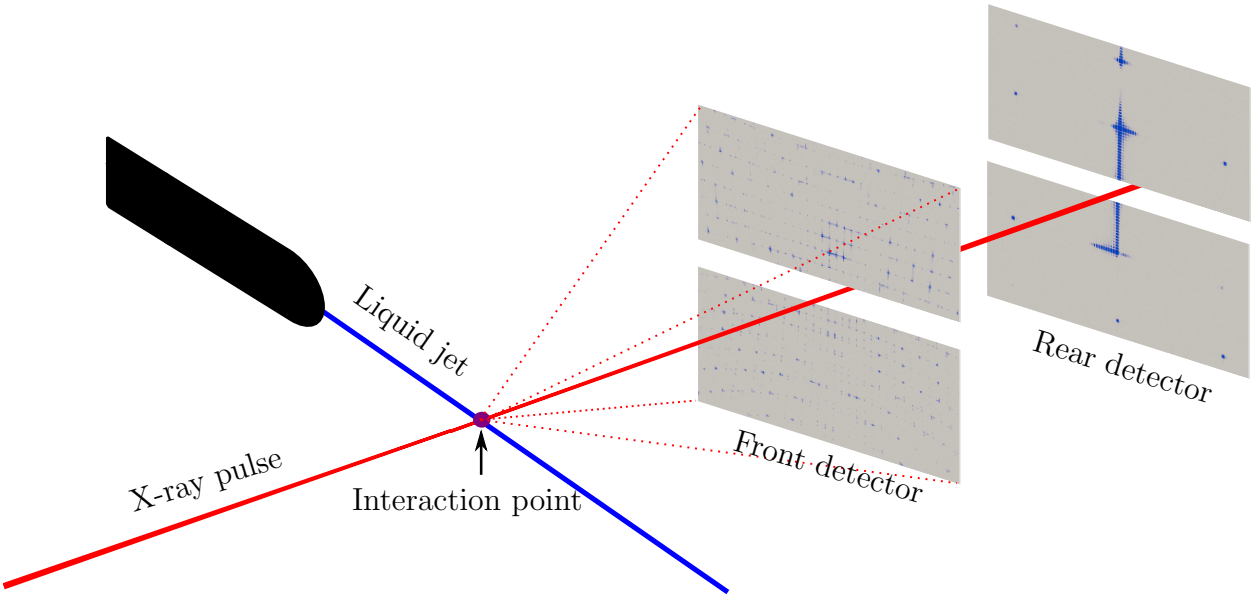


Figure 1.1: Set up of an x-ray nanocrystallography experiment. A liquid jet (blue) is often used to delivery the nanocrystal samples to the x-ray pulse (red). Wide and small angle diffraction data can be collected by utilizing both front and rear detectors.

If the crystal orientations were known, then the noise and variation in the peak measurements could be averaged out, allowing one to proceed to invert the data to retrieve the electron density of the object. In theory, location of a sufficient number of Bragg peaks in an image can be used to determine the orientation of the crystal up to symmetry of its periodic lattice, a process known as autoindexing. While autoindexing has been performed extensively to increase the accuracy of the orientations of conventional crystals, a few fundamental issues still remain in its use for the orientation of nanocrystals [47]. One issue is the robustness of autoindexing in the presence of partial and non-Bragg reflections. Further-

more, autoindexing only narrows down the orientation of an image to a list of possibilities, whose size is the order of the crystal's lattice rotational symmetry group. This leads to an ambiguity in orienting the images when the diffraction pattern does not posses the same symmetry, known as the twinning problem. Additionally, phase information is missing from the images and must be retrieved. While there are many techniques to recover the missing phase information, most of them become infeasible without knowledge of the orientations up to symmetry of the diffraction data. Consequently, in the presence of the twinning problem, current use of nanocrystallography data has been mainly limited to approaches that can work with this twinned data, such as molecular replacement, which requires comparison to a similar known structure, and, thus, cannot be used to discover fundamentally new structures *ab initio*. Previous attempts at removing the twinning ambiguity have proven unsuccessful, largely due to the excessive variance in intensities over the ensemble of images [68, 74].

In this thesis, we develop a new approach to x-ray nanocrystallography reconstruction which relies on solving the twinning problem. Our approach is robust to the amount of noise and uncertainty common in experiments. In particular, our framework directly seeks out the unknown image parameters in order to decrease the amount of variance in the magnitude values. We begin by developing a new technique to increase the precision of the partial orientation information computed in autoindexing, by utilizing reflections between Bragg peaks in order to increase the accuracy of the computed ambiguous orientations in the face of partial reflections and low peak counts. Then, by using a high-resolution low angle image, such as from the rear detector in Figure 1.1, we compute the approximate crystal sizes by using a combination of Fourier analysis and image segmentation. Next, we model the distribution of Fourier magnitudes for each peak via a multi-modal Gaussian distribution by using a multi-stage expectation maximization algorithm that alternates between correcting for the unknown incident photon flux densities and calculating the model parameters. We then use these multi-modal models to build a weighted graphical model of the magnitude concurrency, which describes how often two magnitudes occur within the same image. Solving the twinning problem is formulated as finding the maximum edge clique in this graphical model. Although the maximum edge weight clique problem is, in general, NP-hard, we develop an approximate greedy approach which runs in quadratic time, is exact for the twinning problem in the absence of noise, and highly accurate in the presence of large amounts of noise. The solution to this clique problem then determines the detwinned orientations, up to symmetry of the utilized diffraction data. Additionally, we show that if one determines the orientations up to symmetry of the data on lines connecting adjacent Bragg points, then phase information can be recovered by only utilizing Fourier magnitude information via a compressive phase retrieval algorithm.

This work has the potential to enhance x-ray nanocrystallography experiments by increasing the accuracy of the processed data, decreasing the total number of images required, and allowing for the use of additional phasing methods in the presence of the twinning problem. In particular, this framework allows phase recovery with conventional techniques that

typically require knowledge of the orientations up to symmetry of the data, such as anomalous dispersion and isomorphous replacement. This approach can also aid in the current use of molecular replacement by allowing one to test models against the full set of detwinned data. Alternatively, this framework allows one to compute the phases with only Fourier magnitude information, which does not require the extra setup and measurements needed for experimental phasing nor knowledge of a similar structure.

In Chapter 2, we discuss relevant background information including basic notation and theorems, the mathematical formulation of elastic scattering, autoindexing, techniques and theory for phase retrieval, and the x-ray nanocrystallography reconstruction problem. In Chapter 3, we describe the theory and algorithms behind our x-ray nanocrystallography reconstruction framework. Finally, in Chapter 4, we present a detailed analysis of our methods and demonstrate the feasibility of our approach by reconstructing molecular structure from realistic simulated data.

Chapter 2

Background

2.1 Overview

In this chapter we discuss background material relevant to x-ray nanocrystallography reconstruction.

First, in Section 2.2, we establish some basic notation and theorems that will be commonly referred to throughout the thesis.

Next, in Section 2.3, we present a mathematical formulation of x-ray diffractive imaging. We begin by formulating the basic equations for diffraction due to elastic scattering. Then we discuss various symmetries present in crystals and describe how this affects diffraction from nanocrystals. Additionally, we show how diffraction can be modeled in terms of atomic scattering factors and various noise processes.

Then, in Section 2.4, we discuss current autoindexing techniques, which allow one to deduce the lattice properties and orientations, up to lattice symmetry, of a crystal, by analyzing the periodicity of the recorded reflections in its diffraction pattern. In particular, we describe how this results in an ambiguity in determining the orientations in x-ray nanocrystallography, which leads to the twinning problem when the diffraction pattern displays less symmetry than the lattice.

Afterward, in Section 2.5, we give an overview of the phase problem, which must be solved in order to reconstruct a sample's molecular structure from its diffraction data. Moreover, we present theory and algorithms for the technique of computational phase retrieval, which allows one to solve the phase problem using only Fourier magnitude information. While this computational phase retrieval approach has largely been infeasible in conventional crystallography, we describe how it may be applicable to solving the phase problem in nanocrystallography, due to the increased sampling rate of the Fourier magnitudes provided from the

measurable non-Bragg peaks in nanocrystal diffraction images.

Finally, in Section 2.6, we formulate the reconstruction problem in x-ray nanocrystallography and describe the algorithmic challenges in solving it. In particular, we discuss how, in presence of the twinning problem, reconstruction has been limited to techniques that rely on having a similar known structure available, which may be infeasible for studying fundamentally new objects.

2.2 Basic Notation and Theorems

Here we describe our notation and several key theorems that will be used.

We denote the group of real-valued orthogonal matrices in d dimensions by $O(d)$ and the group of real-valued orthogonal matrices with determinant one by $SO(d)$. Given $\mathbf{N} = (N_1, \ldots, N_d) \in \mathbb{N}^d$, we define $|\mathbf{N}| = \prod_{j=1}^d N_j$ and $\mathbb{Z}_{\mathbf{N}} = \mathbb{Z}_{N_1} \times \cdots \times \mathbb{Z}_{N_d}$, where $\mathbb{Z}_{N_j} = \{0, 1, \ldots, N_j - 1\}$. We use the following conventions for operations between two vectors $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = (y_1, \ldots, y_d)$: $\mathbf{x}^{\mathbf{y}} = \prod_{j=1}^d x_j^{y_j}$ and $\frac{\mathbf{x}}{\mathbf{y}} = (\frac{x_1}{y_1}, \ldots, \frac{x_d}{y_d})$.

We will refer to the following function spaces: $L^1(\mathbb{R}^d)$ denotes the Banach space of complex-valued Lebesgue integrable functions on $\mathbb{R}^d$ with norm $||f||_{L^1} = \int_{\mathbb{R}^d} |f(\mathbf{x})| d\mathbf{x}$, $L^2(\mathbb{R}^d)$ denotes the Hilbert space of complex-valued square integrable functions on $\mathbb{R}^d$ with norm $||f||_{L^2} = \int_{\mathbb{R}^d} |f(\mathbf{x})|^2 d\mathbf{x}$, $\ell^2(\mathbb{Z}^d)$ denotes the Hilbert space of complex-valued square summable functions defined on $\mathbb{Z}^d$ with norm $||f||_{\ell^2} = \sum_{\mathbf{n} \in \mathbb{Z}^d} |f(\mathbf{n})|^2$, and $\ell^2(\mathbb{Z}_{\mathbf{N}})$ denotes the Hilbert space of complex-valued functions defined on $\mathbb{Z}_{\mathbf{N}}$ with norm $||f||_{\ell^2} = \sum_{\mathbf{n} \in \mathbb{Z}_{\mathbf{N}}} |f(\mathbf{n})|^2$.

We will make use of functions $f : \mathbb{R}^d \to \mathbb{C}$, $f : \mathbb{Z}^d \to \mathbb{C}$, and $f : \mathbb{Z}_{\mathbf{N}} \to \mathbb{C}$. In each of these cases, we define the *Fourier transform* as a unitary operator. We will typically refer to the domain of the original function as *real space* and the domain of the Fourier transformed function as *Fourier space* or *reciprocal space*.

Definition 1. *The Fourier transform of $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ is given by*

$$(\mathcal{F}f)(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-2\pi i \mathbf{x} \cdot \boldsymbol{\xi}} d\mathbf{x}, \text{ for all } \boldsymbol{\xi} \in \mathbb{R}^d \tag{2.1}$$

with inverse given by

$$(\mathcal{F}^* \hat{f})(\mathbf{x}) = f(\mathbf{x}) = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi}) e^{2\pi i \mathbf{x} \cdot \boldsymbol{\xi}} d\boldsymbol{\xi}, \text{ for all } \mathbf{x} \in \mathbb{R}^d. \tag{2.2}$$

The above definition of the Fourier transform can be extended to $L^1(\mathbb{R}^d) + L^2(\mathbb{R}^d)$ by using the density of $L^1(\mathbb{R}^d)$ in $L^2(\mathbb{R}^d)$.

Definition 2. *The discrete-time Fourier transform (DTFT) of $f \in \ell^2(\mathbb{Z}^d)$ is given by*

$$(\mathcal{F}f)(\boldsymbol{\xi}) = \hat{f}(\boldsymbol{\xi}) = \sum_{\mathbf{n}\in\mathbb{Z}^d} f(\mathbf{n})e^{-2\pi i \mathbf{n}\cdot\boldsymbol{\xi}}, \text{ for all } \boldsymbol{\xi} \in [0,1]^d, \tag{2.3}$$

with inverse given by

$$(\mathcal{F}^*\hat{f})(\mathbf{n}) = f(\mathbf{n}) = \int_{[0,1]^d} \hat{f}(\boldsymbol{\xi})e^{2\pi i \mathbf{n}\cdot\boldsymbol{\xi}} d\boldsymbol{\xi}, \text{ for all } \mathbf{n} \in \mathbb{Z}^d. \tag{2.4}$$

Definition 3. *The discrete Fourier transform (DFT) of $f \in \ell^2(\mathbb{Z}_{\mathbf{N}})$ is given by*

$$(\mathcal{F}f)(\mathbf{k}) = \hat{f}(\mathbf{k}) = \frac{1}{\sqrt{|\mathbf{N}|}} \sum_{\mathbf{n}\in\mathbb{Z}_{\mathbf{N}}} f(\mathbf{n})e^{-2\pi i \mathbf{n}\cdot(\mathbf{k}/\mathbf{N})}, \text{ for all } \mathbf{k} \in \mathbb{Z}_{\mathbf{N}}, \tag{2.5}$$

with inverse given by

$$(\mathcal{F}^*\hat{f})(\mathbf{n}) = f(\mathbf{n}) = \frac{1}{\sqrt{|\mathbf{N}|}} \sum_{\mathbf{k}\in\mathbb{Z}_{\mathbf{N}}} \hat{f}(\mathbf{k})e^{2\pi i \mathbf{n}\cdot(\mathbf{k}/\mathbf{N})}, \text{ for all } \mathbf{n} \in \mathbb{Z}_{\mathbf{N}}. \tag{2.6}$$

Theorem 1 (Parseval's Theorem). *For $f, g \in L^2(\mathbb{R}^d)$, $f, g \in \ell^2(\mathbb{Z}^d)$, or $f, g \in \ell^2(\mathbb{Z}_{\mathbf{N}})$, we have that*

$$||f|| = ||\hat{f}|| \text{ and } \langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle. \tag{2.7}$$

By realizing that a compactly supported function is the DTFT of some ℓ^2 function, we arrive at the following result.

Theorem 2 (Shannon-Nyquist Theorem). *If $f \in L^2(\mathbb{R}^d)$ satisfies* $\text{supp}(f) \subseteq [-\frac{L_1}{2}, \frac{L_1}{2}] \times \cdots \times [-\frac{L_d}{2}, \frac{L_d}{2}]$ *then f is uniquely determined by $\{\hat{f}(\frac{n_1}{L_1}, \ldots, \frac{n_d}{L_d}) : n_1, \ldots, n_d \in \mathbb{Z}\}$.*

The squared norm of the Fourier transform of a function $|\hat{f}|^2$ is known as the *power spectrum* of f. The power spectrum of a function is related to its *autocorrelation*, which we now define.

Definition 4. *The autocorrelation of $f \in L^2(\mathbb{R}^d)$ is given by*

$$(Af)(\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{y})\overline{f(\mathbf{y}-\mathbf{x})}d\mathbf{y}, \text{ for all } \mathbf{x} \in \mathbb{R}^d. \tag{2.8}$$

Definition 5. *The autocorrelation of $f \in \ell^2(\mathbb{Z}^d)$ is given by*

$$(Af)(\mathbf{n}) = \sum_{\mathbf{m}\in\mathbb{Z}^d} f(\mathbf{m})\overline{f(\mathbf{m}-\mathbf{n})}, \text{ for all } \mathbf{n} \in \mathbb{Z}^d. \tag{2.9}$$

Definition 6. *The autocorrelation of $f \in \ell^2(\mathbb{Z}_{\mathbf{N}})$ is given by*

$$(Af)(\mathbf{n}) = \frac{1}{\sqrt{|\mathbf{N}|}} \sum_{\mathbf{m}\in\mathbb{Z}_{\mathbf{N}}} f(\mathbf{m})\overline{f(\mathbf{m}-\mathbf{n})}, \text{ for all } \mathbf{n} \in \mathbb{Z}_{\mathbf{N}}. \tag{2.10}$$

Theorem 3 (Wiener-Khinchin Theorem). *For $f \in L^2(\mathbb{R}^d)$, $f \in \ell^2(\mathbb{Z}^d)$, or $f \in \ell^2(\mathbb{Z}_{\mathbf{N}})$, we have that*

$$\widehat{Af} = |\hat{f}|^2. \tag{2.11}$$

Note that the DTFT or DFT of a function can be realized as a complex analytic function restricted to the torus $\mathbb{T}^d$ by identifying $\mathbf{z} = e^{2\pi i\boldsymbol{\xi}}$ or $\mathbf{z} = e^{2\pi i\mathbf{k}}$. In particular, we may analytically continue the DTFT and DFT to a function defined on $\mathbb{C}^d \setminus \{0\}$, known as the *Z-transform.*

Definition 7. *The $Z-$transform of $f \in \ell^2(\mathbb{Z}^d)$ is given by*

$$(\mathcal{Z}[f])(\mathbf{z}) = \sum_{\mathbf{n}\in\mathbb{Z}^d} f(\mathbf{n})\mathbf{z}^{-\mathbf{n}}, \text{ for all } \mathbf{z} \in \mathbb{C}^d. \tag{2.12}$$

Definition 8. *The $Z-$transform of $f \in \ell^2(\mathbb{Z}_{\mathbf{N}})$ is given by*

$$(\mathcal{Z}[f])(\mathbf{z}) = \frac{1}{\sqrt{|\mathbf{N}|}} \sum_{\mathbf{n}\in\mathbb{Z}_{\mathbf{N}}} f(\mathbf{n})\mathbf{z}^{-\mathbf{n}}, \text{ for all } \mathbf{z} \in \mathbb{C}^d. \tag{2.13}$$

The *x-ray transform* projects a function to a lower dimensional space and is related to the restriction of the function's Fourier transform.

Definition 9. *The x-ray transform of $f \in L^1(\mathbb{R}^d)$, with $d \geq 2$, to a hyperplane Σ passing through the origin with normal ℓ is given by*

$$(P_\ell f)(\mathbf{x}) = \int_{-\infty}^{\infty} f(A_\Sigma \mathbf{x} + z\ell)dz, \text{ for all } \mathbf{x} \in \mathbb{R}^{d-1}, \tag{2.14}$$

where $A_\Sigma : \mathbb{R}^{d-1} \to \mathbb{R}^d$ parametrizes Σ.

Theorem 4 (Fourier Projection-Slice Theorem). *For $f \in L^1(\mathbb{R}^d)$ and $\ell = (0, \ldots, 0, 1)$, we have that*

$$\widehat{P_\ell f} = \hat{f}|_{\mathbb{R}^{d-1}\times\{0\}}, \tag{2.15}$$

using the canonical parametrization of $\mathbb{R}^{d-1} \times \{0\}$.

We will also make use of generalized functions, such as the Dirac delta function δ. The above definitions and theorems can be extended to many of these generalized functions, e.g., tempered distributions.

2.3 Mathematical Formulation of X-ray Nanocrystallography

2.3.1 Elastic Scattering

The continuous diffraction pattern $I_c : \mathbb{R}^2 \to \mathbb{R}$ due to elastic scattering from an object with electron density $\rho : \mathbb{R}^3 \to \mathbb{R}$, rotated by $R \in SO(3)$, using a fully coherent x-ray beam with wavelength λ and incident photon flux density J at a plane with distance D from the interaction point and normal in the direction of the incident beam, is described by [72]:

$$I_c(x, y) = J\, r_e^2 P(x, y) |\hat{\rho}(R q_\lambda(x, y))|^2 d\Omega(x, y), \tag{2.16}$$

where $q_\lambda : \mathbb{R}^2 \to \mathbb{R}^3$ maps the detector plane onto a spherical slice of frequency space, known as the *Ewald sphere*, depicted in Figure 2.1, and is given by

$$q_\lambda(x, y) = \frac{1}{\lambda} \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2+D^2}} \\ \frac{y}{\sqrt{x^2+y^2+D^2}} \\ \frac{D}{\sqrt{x^2+y^2+D^2}} - 1 \end{pmatrix}, \tag{2.17}$$

r_e^2 is the electron cross-section, $d\Omega(x, y) = \frac{D}{(x^2+y^2+D^2)^{3/2}}$ is the solid angle subtended by a point, and $P : \mathbb{R}^2 \to \mathbb{R}$ is a polarization factor that depends on the polarization type of the beam. For example, for horizontal polarization, $P(x, y) = 1 - \frac{x^2}{x^2+D^2}$. For elastic scattering, the values of $\hat{\rho}$ are often called the *structure factors*. Since the Fourier transform of a real valued function displays *Friedel symmetry*, $\hat{\rho}(\boldsymbol{\xi}) = \overline{\hat{\rho}(-\boldsymbol{\xi})}$, the associated magnitudes are inversion symmetric, i.e., $|\hat{\rho}(\boldsymbol{\xi})| = |\hat{\rho}(-\boldsymbol{\xi})|$.

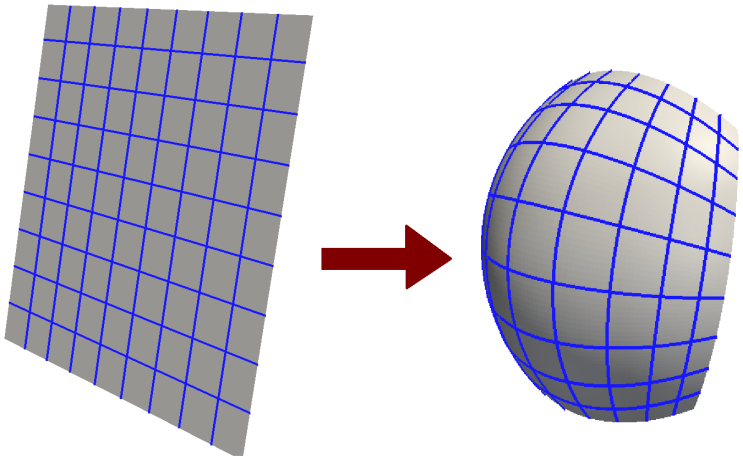


Figure 2.1: The detector plane (left) measures power spectrum information along a spherical slice of reciprocal space (right).

In practice, a detector measures the integral of (2.16) over pixels of size $dx \times dx$, yielding $I : \mathbb{Z}^2 \to \mathbb{R}$

$$I_{m,n} = I(x,y) = \int_x^{x+dx} \int_y^{y+dx} J\, r_e^2 P(a,b) |\hat{\rho}(Rq_\lambda(a,b))|^2 d\Omega(a,b) \tag{2.18}$$

$$\approx J\, r_e^2 P(x,y) \Delta\Omega(x,y) |\hat{\rho}(Rq_\lambda(x,y))|^2 \tag{2.19}$$

where $\Delta\Omega(x,y) = \frac{D\,dx^2}{(x^2+y^2+D^2)^{3/2}}$ is the solid angle subtended by a pixel, $x = m\,dx$, and $y = n\,dx$. We will slightly abuse notation by treating (2.19) as equality.

2.3.2 Crystal Lattice Theory

In x-ray crystallography, one collects a series of diffraction patterns from a periodic crystal made up of the target object. The three-dimensional crystal lattice structure may be described by its *Bravais lattice characteristic* $(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3), \mathbf{h}_j \in \mathbb{R}^3$, depicted in Figure 2.2. In particular, the infinite lattice $\mathcal{L}$ consists of all points which are integer combinations of the Bravais vectors:

$$\mathcal{L} = \left\{ \sum_{j=1}^{3} n_j \mathbf{h}_j : n_j \in \mathbb{Z} \right\}, \tag{2.20}$$

where (n_1, n_2, n_3), known as the *Miller indices*, describe a position within the lattice.

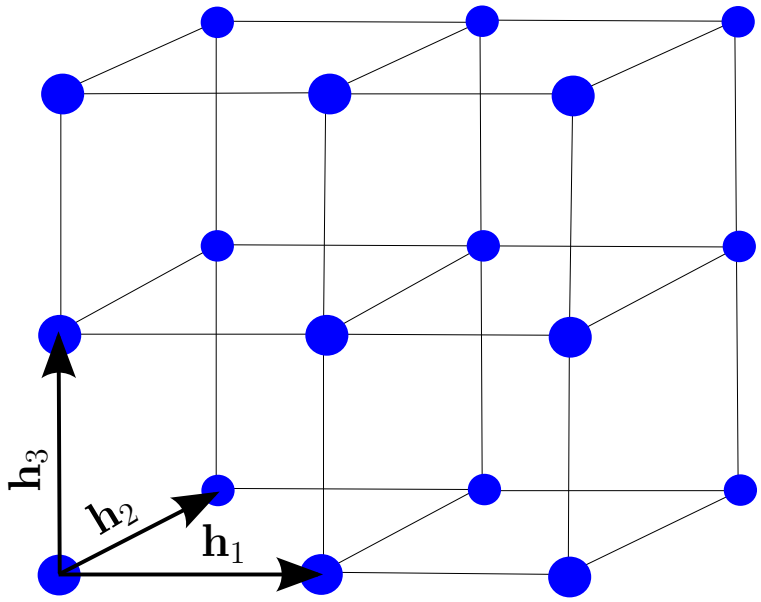


Figure 2.2: Example of a crystal lattice generated from integer combinations of the Bravais characteristic vectors $(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$.

Crystal lattices are categorized by their *point groups*, i.e., the set of linear operators that map L to itself and leave some point fixed. In particular, we define the *lattice symmetry group* $\mathcal{S}(\mathcal{L})$ of a lattice to be the set of orthogonal linear operators which preserve the lattice structure:

$$\mathcal{S}(\mathcal{L}) = \{Q \in O(3) : Q\mathcal{L} = \mathcal{L}\}. \tag{2.21}$$

Similarly, we define the *lattice rotational symmetry group* to be restriction of the lattice symmetry group to rotations, $\mathcal{S}_R(\mathcal{L}) = \mathcal{S}(\mathcal{L}) \cap SO(3)$. In three dimensions, crystal lattices can be classified by their symmetry groups into to one of 7 possible *lattice systems* [30]. For each crystal lattice, we can define an associated *Dirac comb* $\Delta_{\mathcal{L}}$, which is a sum of Dirac delta functions supported on the lattice points:

$$\Delta_{\mathcal{L}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{L}} \delta(\mathbf{x} - \mathbf{y}), \text{ for all } \mathbf{x} \in \mathbb{R}^3. \tag{2.22}$$

Each crystal lattice has a dual, known as the *reciprocal lattice* $\hat{\mathcal{L}}$, which is given by the support of the Dirac comb's Fourier transform $\hat{\Delta}_{\mathcal{L}}$. Note that the Fourier transform of the Dirac comb of $\mathcal{L}$ is the Dirac comb of $\hat{\mathcal{L}}$ up to a multiplicative constant, i.e., $\hat{\Delta}_{\mathcal{L}} = (|\mathbf{h}_1||\mathbf{h}_2||\mathbf{h}_3|)^{-1}\Delta_{\hat{\mathcal{L}}}$. The Bravais lattice characteristic $(\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \hat{\mathbf{h}}_3)$ of the reciprocal lattice can be expressed in terms of the original Bravais vectors:

$$\hat{\mathbf{h}}_1 = \frac{\mathbf{h}_2 \times \mathbf{h}_3}{\mathbf{h}_1 \cdot (\mathbf{h}_2 \times \mathbf{h}_3)}, \quad \hat{\mathbf{h}}_2 = \frac{\mathbf{h}_3 \times \mathbf{h}_1}{\mathbf{h}_2 \cdot (\mathbf{h}_3 \times \mathbf{h}_1)}, \quad \hat{\mathbf{h}}_3 = \frac{\mathbf{h}_1 \times \mathbf{h}_2}{\mathbf{h}_3 \cdot (\mathbf{h}_1 \times \mathbf{h}_2)}. \tag{2.23}$$

In practice, a crystal lattice $\mathcal{L}_{\mathcal{C}}$ consists of only a finite part of its associated infinite lattice. In this case, the associated Dirac comb's Fourier transform $\hat{\Delta}_{\mathcal{L}_{\mathcal{C}}}$, known as the *shape transform*, is no longer a sum of delta functions, but, instead, is a smeared out version of $\hat{\Delta}_{\mathcal{L}}$. In particular, if we assume that the finite crystal lattice can be described as $\mathcal{L}_{\mathcal{C}} = \{\sum_{j=1}^{3} n_j \mathbf{h}_j : n_j \in \mathbb{Z}_{N_j}\}$, then its associated shape transform $S : \mathbb{R}^3 \to \mathbb{C}$ is given by

$$S(\boldsymbol{\xi}) = \prod_{j=1}^{3} \frac{e^{2\pi i N_j \mathbf{h}_j \cdot \boldsymbol{\xi}} - 1}{e^{2\pi i \mathbf{h}_j \cdot \boldsymbol{\xi}} - 1}. \tag{2.24}$$

In diffractive imaging, one typically works with the squared norm of the shape function, which can be expressed as

$$|S(\boldsymbol{\xi})|^2 = \prod_{j=1}^{3} \frac{\sin^2(\pi N_j \mathbf{h}_j \cdot \boldsymbol{\xi})}{\sin^2(\pi \mathbf{h}_j \cdot \boldsymbol{\xi})}. \tag{2.25}$$

2.3.3 Space Groups

While the crystal lattice structure describes the translational symmetry of the crystal, the arrangement of molecules may display extra symmetry within each periodic unit, known as

the *unit cell*. The full symmetry of the crystal is described by the *space group* of the crystal. In particular, within a unit cell, there are often multiple copies of the molecule, which can be described in terms of rotations, reflections, improper rotations, translations, glide planes, and screw axes applied to the molecule in some reference position, depicted in Figure 2.3. In general, these symmetry operations can be described by mapping a reference $\rho(\mathbf{x})$ to the symmetry elements $\rho(\mathbf{y})$ by

$$\mathbf{y} = M\mathbf{x} + D, \tag{2.26}$$

where M is a matrix and D is a vector.

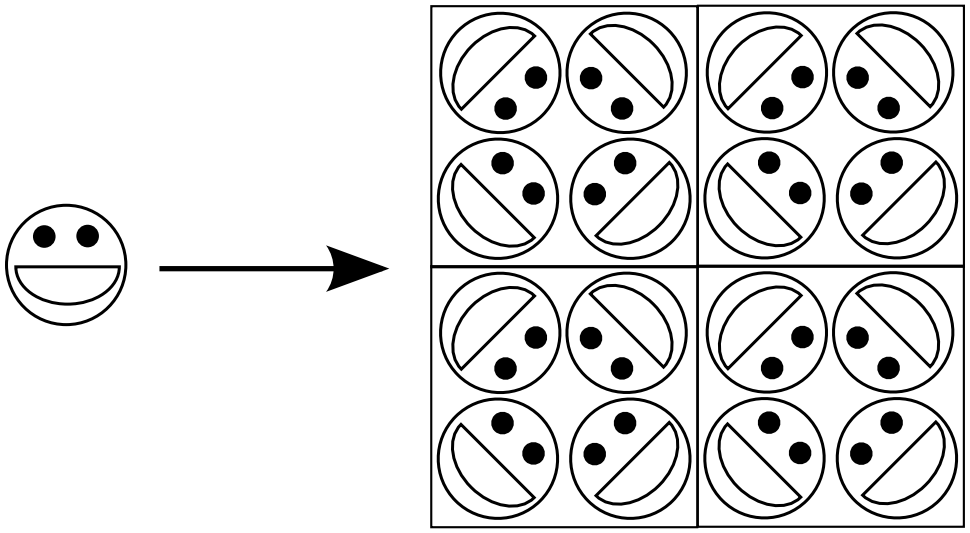


Figure 2.3: A molecule (left) is arranged into a periodic unit cell (right) via a set of affine transformations.

Combining the possible unit cell symmetries with the different lattice systems yields 230 possible space groups [30]. The point group symmetries of a crystal $\mathcal{C}$ with electron density ρ_C, can be described by its *crystal point symmetry group* $\mathcal{S}(\mathcal{C})$:

$$\mathcal{S}(\mathcal{C}) = \{Q \in O(3) : \rho_{\mathcal{C}}(Q\mathbf{x}) = \rho_{\mathcal{C}}(\mathbf{x}), \text{ for all } \mathbf{x} \in \mathbb{R}^3\}. \tag{2.27}$$

We also define the *crystal rotational symmetry group* as $\mathcal{S}_R(\mathcal{C}) = \mathcal{S}(\mathcal{C}) \cap SO(3)$. Note that while the point symmetry group of the crystal is a subset of the symmetry group of the lattice, $\mathcal{S}(\mathcal{C}) \subseteq \mathcal{S}(\mathcal{L})$, they are not necessarily equal.

The space group of a crystal introduces another form of symmetry on its diffraction pattern, known as *Laue symmetry*. In particular, the *Laue symmetry group* $\mathcal{S}_L(\mathcal{C})$ of a crystal $\mathcal{C}$ is given by the set of orthogonal operators which preserve the structure factor magnitudes of the crystal at its reciprocal lattice points:

$$\mathcal{S}_L(\mathcal{C}) = \{Q \in O(3) : |\hat{\rho}_{\mathcal{C}}(Q\boldsymbol{\xi})| = |\hat{\rho}_{\mathcal{C}}(\boldsymbol{\xi})|, \text{ for all } \boldsymbol{\xi} \in \mathcal{L}\}. \tag{2.28}$$

We also define the *Laue rotational symmetry group* as $\mathcal{S}_{L,R}(\mathcal{C}) = \mathcal{S}_L(\mathcal{C}) \cap SO(3)$. The Laue symmetry group is at least as big as the crystal point symmetry group but never bigger than the lattice symmetry group, i.e., $\mathcal{S}(\mathcal{C}) \subseteq \mathcal{S}_L(\mathcal{C}) \subseteq \mathcal{S}(\mathcal{L})$.

2.3.4 Crystal Diffraction

For simplicity, we make the assumption that the crystal lattice can be expressed as $\mathcal{L}_\mathcal{C} = \{\sum_{j=1}^{3} n_j \mathbf{h}_j : n_j \in \mathbb{Z}_{N_j}\}$. In this case, the electron density $\rho_\mathcal{C}$ of a crystal can be expressed in terms of the electron density ρ of one of its unit cells by

$$\rho_\mathcal{C}(x) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} \rho(x + n_1\mathbf{h}_1 + n_2\mathbf{h}_2 + n_3\mathbf{h}_3) \tag{2.29}$$

$$= \rho(x) * \Delta_{\mathcal{L}_\mathcal{C}}. \tag{2.30}$$

Therefore, by Equation (2.19), the diffraction pattern of a crystal due to elastic scattering is given by

$$I(x,y) = J\, r_e^2 P(x,y) \Delta\Omega(x,y) |\hat{\rho}(Rq_\lambda(x,y))|^2 |S(Rq_\lambda(x,y))|^2. \tag{2.31}$$

For a large crystal, as is used in conventional crystallography, the shape function approaches the Dirac comb associated to the reciprocal lattice, up to a constant factor. Therefore the diffraction images of a large crystal consist of a series of bright spots, known as *Bragg peaks*, concentrated at the reciprocal lattice points. However, for nanocrystallography, the crystal sizes are small enough for one to notice the spread of the shape function, which are approximately Gaussian around a peak and oscillate outward. In this case, measurements close to, but not directly at, a Bragg peak are known as *partial reflections* and have a decreased amount of collected signal. Additionally, the signal at pixels corresponding to lines in between reciprocal lattice points is often noticeable in nanocrystal diffraction images. An example of a nanocrystal diffraction image is given in Figure 2.4.

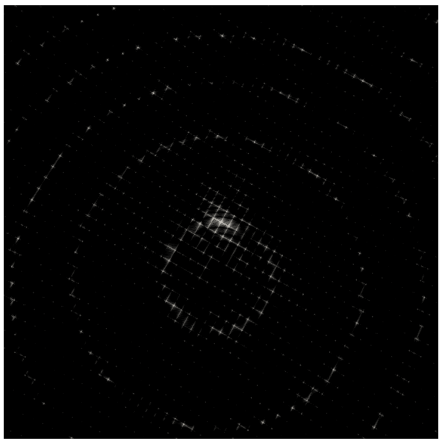


Figure 2.4: Simulated x-ray nanocrystallography diffraction image (log scale).

2.3.5 Atomic Scattering and Dispersion Factors

The structure factors F for a molecule can be expressed as the superposition of the *atomic scattering factors* $\{f_{o,a_k}\}$ of its atoms $\{a_k\}$:

$$F(q) = \sum_k f_{o,a_k}(q) e^{2\pi i \mathbf{x}_k \cdot q}. \tag{2.32}$$

The atomic scattering factors for each atom are well known and documented in [75].

In most cases, diffraction due to elastic scattering is the dominant signal in the collected images. However, when the wavelength of the beam is near an absorption edge of one of the atoms of the sample, effects from absorption of the x-ray photons, known as *anomalous dispersion*, become noticeable. In particular, this phenomenon induces both a magnitude and phase shift in the diffraction signal and, consequently, breaks Friedel and, partially, Laue symmetry. Anomalous dispersion is often modeled via the addition of *dispersion corrections* $\Delta f'_{a_k} + i\Delta f''_{a_k}$ to the Fourier transform of the electron density [29]. The dispersion factors for an individual atom can, for most commonly used energies, be modeled as constants, which depend only on the wavelength of the x-ray beam, and can be extended to the entire sample via superposition:

$$F(q,\lambda) = \sum_k (f_{o,a_k}(q) + \Delta f'_{a_k}(\lambda) + \Delta f''_{a_k}(\lambda)) e^{2\pi i \mathbf{x}_k \cdot q} \tag{2.33}$$

2.3.6 Noise Models in X-ray Nanocrystallography

Due to the quantum nature of light, only a discrete number of photons can be detected. In particular, one can think of the measured magnitudes in Equations (2.19) and (2.31) as representing the probability of a photon appearing at a pixel. Consequently, this discrete behavior induces a type of noise, known as *shot noise*, which can be described in terms of a Poisson distribution, see Figure 2.5b. More specifically, if the expected value at a pixel with position (x, y) is $v = I(x, y)$ then the probability of measuring p photons at that pixel $Pr(v_m = p)$ can be approximated by

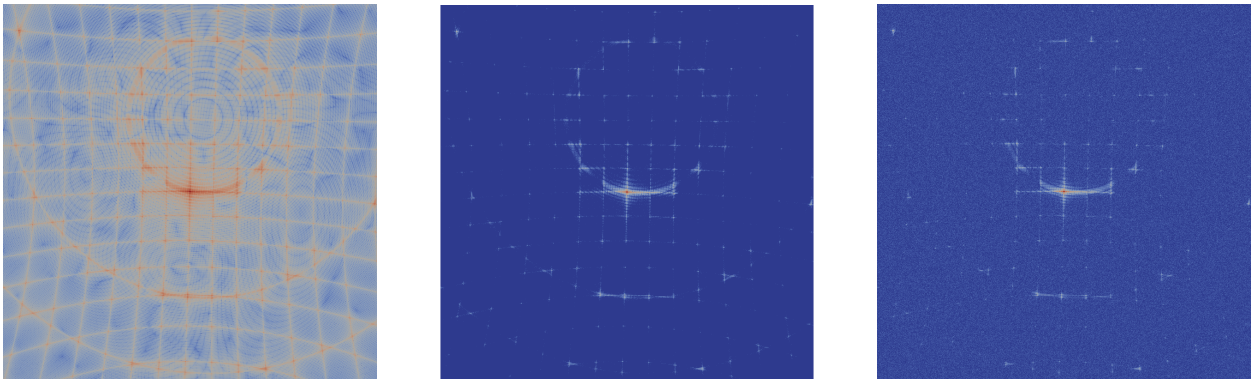
$$Pr(v_m = p) = \frac{v^p e^{-v}}{p!}. \tag{2.34}$$

In addition to the desired information from elastic scattering of the nanocrystal, the detectors collect signal due to various background sources. This includes scattering from the disordered liquid jet and solvent molecules, electronic noise from the detector, and inelastic scattering effects. Apart from detector noise, which is typically correlated with specific pixels, these sources induce a diffuse background on the detector. These background effects can, for the most part, be measured over several blanks shots, where the nanocrystal sample fails

to intersect the x-ray beam, and subsequently subtracted out from the desired nanocrystal diffraction images. However, fluctuations in the background levels lead to another source of noise, known as *background noise*, which is uncorrelated with the elastic scattering from the nanocrystal, see Figure 2.5c. Background noise is typically modeled as an additive Gaussian white noise term, in which the measured intensities I_m are given as

$$I_m(x, y) = I(x, y) + W(x, y), \text{ where } W(x, y) \sim \mathcal{N}(0, \sigma), \tag{2.35}$$

i.e., W is drawn from a normal distribution with mean 0 and standard deviation σ.



(a) Clean image. (b) Shot noise. (c) Shot and background noise.

Figure 2.5: Examples of the effects of noise in x-ray nanocrystallography diffraction images (log scale). Shot noise largely removes pixels with low intensity while background noise adds fuzziness uniformly throughout the image.

A major source of uncertainty is the large variation in the incident photon flux density J. In particular, the liquid jet delivery system does not allow precise control over the position of the nanocrystals with respect to the x-ray beam, which causes several partial hits along with complete misses. The profile of the beam is often approximated as a Gaussian, i.e., the photon flux density at a point $\mathbf{x}$ in the plane normal to the beam with peak photon flux density J_o at its center $\mathbf{x}_o$ is given by

$$J(\mathbf{x}) = J_o e^{-\frac{|\mathbf{x}-\mathbf{x}_o|^2}{2\sigma_J^2}}, \tag{2.36}$$

for some σ_J, which represents the width of the beam. If we assume that the liquid jet is accurately aligned with the beam, then the position of the nanocrystals only vary along the jet and can be modeled via a uniform distribution over some interval [-B,B], denoted by $U(-B, B)$, resulting in

$$J = J_o e^{-\frac{x^2}{2\sigma_J^2}}, \text{ where } x \sim U(-B, B). \tag{2.37}$$

More specifically, for a large ensemble of diffraction patterns, each image will be multiplied by a random J drawn from (2.37).

Other sources of noise and uncertainty include effects from detector artifacts, crystal imperfections, damage processes, vibrations, and limitations in beam tuning. However, the noise processes above tend to serve as the biggest obstacles to processing x-ray nanocrystallography data.

2.4 Autoindexing

In principle, one can use the location of Bragg peaks in a crystallographic diffraction image to determine the lattice structure of the crystal along with partial orientation information. In particular, the Fourier transform of the reciprocal lattice retrieves the lattice structure of the periodic crystal, which determines the Bravais lattice vectors $(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$, along with the orientation R up to symmetry of the lattice. This fact is used in a class of techniques, referred to as *autoindexing*, to retrieve the above information by analyzing the distribution of Bragg peaks in the images. However, note that the two-dimensional images only contain partial information about the three-dimensional lattice. While autoindexing has been used extensively to increase the orientation information for conventional crystallography, where diffraction images are collected through a rotational average to integrate out the shape transform, its application to nanocrystallography is a current area of study [47].

2.4.1 Autoindexing Techniques

One method of autoindexing is based on embedding the Bragg points detected on the 2D images into three-dimensional space and then taking the Fourier transform of this embedding [13, 55]. In particular, a mask $b : \mathbb{R}^2 \to \mathbb{R}$ with a tolerance τ is applied to the image to filter out everything but the Bragg peaks:

$$b(x, y) = \begin{cases} 1, & \text{if } I(x, y) > \tau \text{ and } I(x, y) \text{ is a local maximum,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.38}$$

This is then embedded onto a 3-D Cartesian grid by forming $B : \mathbb{R}^3 \to \mathbb{R}$ where

$$B(\boldsymbol{\xi}) = \begin{cases} b(q^{-1}(\boldsymbol{\xi})), & \text{if } \boldsymbol{\xi} \in q(\mathbb{R}), \\ 0, & \text{otherwise,} \end{cases} \tag{2.39}$$

where some type of interpolation is used when the Bragg points don't align with the grid. The function B, used as an approximation to the reciprocal lattice, is then Fourier transformed to retrieve $\hat{B}$, which gives a 3-D approximation to the crystal lattice. The Bravais vectors are then retrieved by analyzing the lattice structure of $\hat{B}$. However, the crystal lattice approximation, provided by $\hat{B}$, becomes blurry and difficult to analyze when an insufficient

number of peaks are used or if non-Bragg intensities are not filtered out, both of which are prominent issues in nanocrystallography. A recent version of the above approach utilizes compressed sensing techniques in order to reduce the number of required reflections and achieves a sharper image of the crystal lattice at the cost of performing several iterations of a nonlinear solver [47].

Another approach to autoindexing is based on the Fourier analysis of projections of the Bragg data onto a series of one-dimensional lines, which are used to search for the individual Bravais lattice vectors of the crystal [69]. For each line, represented by a unit vector ℓ coming from a uniform distrubtion of the unit sphere, one creates a set of frequency bins $f_\ell : \mathbb{Z} \to \mathbb{Z}$ of a set length L and computes the projections of the set of detected Bragg points Br:

$$f_\ell(n) = \big|\{\mathbf{x} \in Br : nL \leq \mathbf{x} \cdot \ell \leq (n+1)L\}\big|. \tag{2.40}$$

One can think of f as the x-ray projection operator being applied twice to B, i.e., it approximates

$$f_\ell(x) = \int_{\ell^\perp} B(x\ell + \mathbf{y})d\mathbf{y} = (P_{\ell_1^\perp} P_{\ell_2^\perp} B)(x), \tag{2.41}$$

where $\ell^\perp = \ell_1^\perp \oplus \ell_2^\perp$ is an orthogonal decomposition. One then proceeds by computing the Fourier transform of f, which by the Fourier projection-slice theorem, applied twice, equals the restriction of $\hat{B}$ to ℓ, i.e., $\hat{f}_\ell = \hat{B}|_\ell$. In particular, when ℓ is in the direction of a Bravais vector, $\hat{f}_\ell$ contains several peaks where ℓ intersects the crystal lattice points, which are separated by the length of the Bravais vector. Therefore, the lines which correspond to the function $\hat{f}_\ell$ with largest norm are then used as candidates for the directions of the Bravais lattice characteristic vectors and the associated Bravais vector lengths are computed as the distance between the peaks of the function. The advantage of this approach over the 3-D approach is that one has more direct control over the resolution of the search directions.

Another version of the 1-D projection approach focuses on directions which are orthogonal to planes formed by triplets of reciprocal lattice points [20]. In particular, the normal to such a reciprocal plane is of the form

$$\mathbf{n} = a_1\mathbf{h}_1 + a_2\mathbf{h}_2 + a_3\mathbf{h}_3, \text{ where } a_i \in \mathbb{Z}. \tag{2.42}$$

Instead of using a Fourier transform, the length of the corresponding lattice vectors are computed by looking for the largest distance between consecutive projected peaks. The directions are then analyzed and those which best fit the peak distribution are taken as the Bravais vectors. While effective at handling complexities due to crystal anomalies, it requires the presence of an appropriate set of measured peaks in order to ensure that true Bravais lattice vectors are present in the list of the triplet normals.

2.4.2 Lattice Orientations and the Twinning Problem

Once the Bravais lattice characteristic vectors $H = (\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$, represented in a reference frame, are known, one can proceed to compute orientation information for each image,

but only up to symmetry of the lattice. In particular, note that the Bravais vectors $V = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ retrieved by autoindexing are only determined up to multiplication of the $\mathbf{v}_i$ by -1. Each choice of sign, consistent with $\det(V) = \det(H)$, corresponds to a different rotation $\tilde{R}$ such that

$$V = \tilde{R}H. \tag{2.43}$$

Note that the set of possible rotations in (2.43) is given by multiplying $\tilde{R}$ by elements of the lattice rotational symmetry group, i.e., for every $R \in \mathcal{S}_R(\mathcal{L})$, $R\tilde{R}$ is a another possible rotation. Therefore, the orientation of the crystal, up to symmetry of the lattice, can be retrieved by computing

$$\tilde{R} = VH^{-1}. \tag{2.44}$$

The ambiguity in the calculation of $\tilde{R}$ is known as the *autoindexing ambiguity*. In conventional crystallography, the autoindexing ambiguity can be resolved since one has control over the crystal orientations up to a certain precision. However, due to the randomization of the delivery system, this is not possible in nanocrystallography. In particular, if the diffraction pattern does not have the same symmetry as the lattice, then autoindexing is unable to determine complete orientation information, which is known as the *twinning problem*, depicted in Figure 2.6. In such cases, there is an ambiguity in assigning the recorded intensities to their corresponding locations in reciprocal space. For most crystals, the twinning problem induces no more than a two-fold ambiguity in the Bragg data. However, the size of this ambiguity can potentially increase in the presence of anomalous dispersion or when observing non-Bragg data. We will refer to a quantity as being *twinned* when the twinning problem leads to an ambiguity in its determination, and will refer to it as being *detwinned* when its associated ambiguity is resolved.

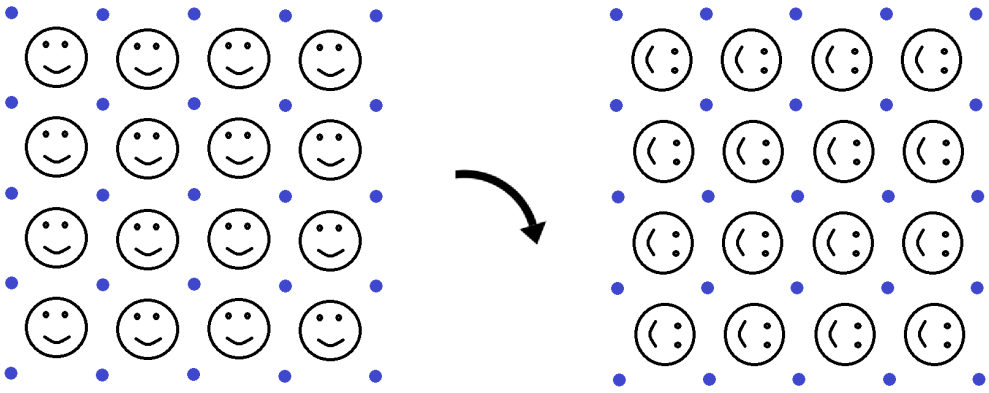


Figure 2.6: Example of the twinning problem: The lattice points (blue) are symmetric with respect to rotation by 90 degrees while the constituent molecule (happy face) is not. Autoindexing techniques only utilize the lattice points and, thus, cannot distinguish between these two orientations.

2.5 Phase Recovery

As described in Section 2.3, diffraction images do not contain phase information, which must be recovered in order to retrieve the electron density of the sample. Note that, due to Friedel symmetry, the Fourier magnitude information can, at most, determine a structure up to inversion through some point, i.e., chirality is lost. Moreover, in some cases, fundamentally different structures can have the same Fourier magnitudes, as was originally discovered in the x-ray crystallography study of the molecule bixbyite [58]. In particular, there exist many examples of structures which demonstrate such non-uniqueness, known as *homometric structures* [25, 57]. Fortunately, this non-uniqueness can be avoided by including additional measurements or constraints in the inversion process, e.g., via anomalous dispersion, addition of heavy atoms, and/or additional modeling requirements.

2.5.1 Techniques

One of the earliest developed phase recovery methods is *Patterson map analysis* [56], which analyzes the *Patterson map* $P : \mathbb{R}^3 \to \mathbb{C}$ of the structure factor magnitudes $|F|$:

$$P(\mathbf{x}) = \sum_{\boldsymbol{\xi} \in \hat{\mathcal{L}}_m} |F(\boldsymbol{\xi})|^2 e^{-2\pi i \mathbf{x} \cdot \boldsymbol{\xi}}, \tag{2.45}$$

where $\hat{\mathcal{L}}_m$ is the set of reciprocal lattice points whose intensities were measured. In particular, the Patterson map is an aliased version of the autocorrelation function, which reveals information about the relative displacements between the atoms in a molecule. This information can potentially be used, barring uniqueness issues, to determine the corresponding atomic positions directly. Unfortunately, the displacements become difficult to determine when a large molecule, with more than a few hundred atoms, is studied or if atomic resolution is not obtained. However, Patterson map analysis is often used to locate specific types of atoms in other phasing methods.

Another class of phasing techniques, known as *direct methods*, attempt to exploit known statistical relations between the phases of the structure factors [31]. For instance, if we have a collection of reciprocal coordinates that satisfy $\sum_{j=1}^{N} \boldsymbol{\xi}_j = 0$ then the sum of the phases of the corresponding structure factors

$$\sum_{j=1}^{N} -i \log \left(\frac{F(\boldsymbol{\xi}_j)}{|F(\boldsymbol{\xi}_j)|} \right), \tag{2.46}$$

called *structure invariants*, are invariant with respect to translation of the solution. These types of invariants, along with the known positivity of the solution, can be exploited via probabilistic techniques in order to determine a likely set of phases for the solution. However, this approach is typically only feasible for small molecules, containing no more than a few hundred atoms, but it can be used to augment other methods.

An additional way to determine phase information is through *anomalous dispersion* methods, where one uses x-ray wavelengths near an absorption edge of some of the atoms in the sample, in order to exploit the behavior of the dispersion correction terms, discussed in Section 2.3.5. In this case, assuming there is only one type of atom displaying anomalous scattering, with known scattering factors $f_o + \Delta f' + \Delta f''$ when placed at the origin, the magnitude of the modified structure factors F_m, which depend on the x-ray wavelength λ, can be expressed as

$$|F_m(\boldsymbol{\xi}, \lambda)|^2 = \left| F(\boldsymbol{\xi}) + \sum_{k=1}^{K} F_{A,k}(\boldsymbol{\xi}) \frac{\Delta f'(\lambda) + i\Delta f''(\lambda)}{f_o(\boldsymbol{\xi})} \right|^2, \tag{2.47}$$

where F is the unmodified structure factor, due only to elastic scattering, of the entire sample and the $F_{A,k}$ are the unmodified structure factors of just the anomalous scatterers, i.e., the atoms with absorption edges near λ. If we group wavelength-dependent terms, we can rewrite (2.47) as

$$\begin{aligned} |F_m(\boldsymbol{\xi}, \lambda)|^2 = |F(\boldsymbol{\xi})|^2 + a(\boldsymbol{\xi}, \lambda)|F_A(\boldsymbol{\xi})|^2 + b(\boldsymbol{\xi}, \lambda)|F(\boldsymbol{\xi})||F_A(\boldsymbol{\xi})| \cos(\phi - \phi_A) \\ + c(\boldsymbol{\xi}, \lambda)|F(\boldsymbol{\xi})||F_A(\boldsymbol{\xi})| \sin(\phi - \phi_A), \end{aligned} \tag{2.48}$$

where $F_A = \sum_{k=1}^{K} F_{A,k}$, $\phi = -i \log(\frac{F}{|F|})$ and $\phi_A = -i \log(\frac{F_A}{|F_A|})$ are the phases, and the a, b, and c terms are functions of the known structure factors, along with dispersion corrections, from the anomalously scattering atoms.

In *multi-wavelength anomalous dispersion* (MAD), one collects diffraction patterns at a few, typically 3-4, different wavelengths in order to compute $|F|$, $|F_A|$, $\cos(\phi - \phi_A)$, and $\sin(\phi - \phi_A)$ [34, 35, 40, 41]. The anomalous scattering factors F_A can then be deduced by performing a Patterson map analysis on $|F_A(\boldsymbol{\xi})|^2$ or $||F_m(\boldsymbol{\xi})| - |F_m(-\boldsymbol{\xi})||^2$. The corresponding Patterson map consist of a few peaks that correspond to the relative displacements of the anomalously scattering atoms and are simple enough that one can directly infer the locations of the anomalous scatterers and, subsequently, deduce F_A. The values of F_A, $|F|$, $\cos(\phi - \phi_A)$, and $\sin(\phi - \phi_A)$ can then be used to determine F. The MAD procedure is similar in the case of multiple anomalous scattering types, except that Equation (2.48) has more variables. An alternative to MAD is *single-wavelength anomalous dispersion* (SAD), which only measures the sample with a single wavelength and solves for the variables in (2.48) via other constraints, such as minimal support in solvent flattening or reinforcing certain structure factor statistics in histogram matching [36, 71].

Another approach to phase recovery is *isomorphous replacement*, where one collects diffraction patterns from a crystal along with additional versions of the crystal with heavy atoms added [17, 27, 59]. In particular, the crystals must be *isomorphic*, i.e., they must have the same crystal lattice structure and the same placement of the original sample in the unit cell. If the structure factors of the heavy atom are given by F_H, which are known *a priori*

up to translation of the molecule, then the squared magnitudes of the measured structure factors of the crystal with heavy atoms is given by

$$|F_m(\boldsymbol{\xi})|^2 = |F(\boldsymbol{\xi}) + F_H(\boldsymbol{\xi})|^2, \tag{2.49}$$

where $|F|^2$ is determined from the diffraction pattern of the unmodified crystal. Since $|F_m| - |F|$ will largely reflect the additional scattering from the heavy atoms, a Patterson map analysis on $||F_m| - |F||^2$ can be used to locate the positions of the heavy atoms and, thus, determine F_H. One can then compute

$$|F_m(\boldsymbol{\xi})|^2 - |F(\boldsymbol{\xi})|^2 - |F_H(\boldsymbol{\xi})|^2 = 2\Re(F(\boldsymbol{\xi})\overline{F_H(\boldsymbol{\xi})}). \tag{2.50}$$

However, (2.50) only narrows down the possible phases for F to two possibilities for each $\boldsymbol{\xi}$.

Multiple isomorphous replacement (MIR) resolves the phase ambiguity in (2.50) by using a third crystal with a different heavy atom [10]. An alternative to MIR is *single isomorphous replacement* (SIR), which only uses one crystal with heavy atoms and, similar to SAD, removes the remaining phase ambiguity through other constraints such as symmetry, solvent flattening, and histogram matching [11, 76]. Additionally, one can combine isomorphous replacement with anomalous scattering as is done in *single isomorphous replacement with anomalous scattering* (SIRAS) [60] and *multiple isomorphous replacement with anomalous scattering* (MIRAS) [42].

If one already has a good model for the structure of the sample, then the phases can often be deduced directly from the collected magnitude information with a technique known as *molecular replacement* (MR) [65, 64]. However, depending on the space group symmetry of the crystal, there are often several copies of the molecule sample placed at different locations and orientations in a unit cell. Therefore, one must determine the orientation and translation of the sample molecules with respect to the model. One can decouple the orientation and translation search by first analyzing intramolecular information contained in the Patterson map restricted to a sphere around the origin. Given a spherical shell S and the Patterson maps P of the sample and P_M of the model, the orientation step can be expressed as solving

$$\max_{R \in SO(3)} \int_S P(R\mathbf{x}) P_M(\mathbf{x}) d\mathbf{x}. \tag{2.51}$$

The optimization problem (2.51) can be solved efficiently in reciprocal space by expanding the structure factors and characteristic function of S onto a basis of spherical Bessel functions and spherical harmonics [18]. Once the orientation of the reference molecule in the unit cell is determined, the placement of molecules within the entire unit cell must be determined. In particular, the arrangement is described by the space group of the crystal, which one may have to first deduce, and is based on the position of the reference molecule within the unit cell. Therefore, the translation step can be formulated as maximizing the inner product of

the observed structure factor square magnitudes $|F(\boldsymbol{\xi})|^2$ with a model $|F_M(\boldsymbol{\xi}, \mathbf{x})|^2$ of the unit cell U with the origin of the reference placed at $\mathbf{x}$ [19]:

$$\max_{\mathbf{x}\in U} \sum_{\boldsymbol{\xi}\in\hat{\mathcal{L}}_m} |F(\boldsymbol{\xi})|^2 |F_M(\boldsymbol{\xi}, \mathbf{x})|^2, \tag{2.52}$$

where $\hat{\mathcal{L}}_m$ is the set of reciprocal lattice points whose intensities were measured. Once the orientation and translation information is known, one can proceed to refine the model to better match the observed intensities.

Alternatively one can, in principle, determine the phase information based solely on the collected Fourier magnitudes if certain sampling requirements are met, in a process known as *computational phase retrieval*. In particular, phase information can almost always be uniquely determined, up to certain trivial operations, from Fourier magnitude information, if one can sample the power spectrum of an object at twice the *Nyquist rate* of the object, i.e., the sampling rate used in the Shannon-Nyquist theorem. Unfortunately, this sampling requirement is infeasible in conventional crystallography, where diffraction data is only measured at reciprocal lattice points, which sample directly at the Nyquist rate.

Recall that, due to the noticeable effects of the shape transform, diffraction images from nanocrystals typically contain a significant amount of intensity information between Bragg peaks, e.g., see Figure 2.4. In particular, this inter-Bragg data may allow one to sample the power spectrum at a rate which is sufficient for the use of computational phase retrieval. In the following subsections, we explore this idea by discussing the theory of computational phase retrieval and current algorithmic approaches. In particular, we describe the effectiveness of computational phase retrieval when using different sampling strategies and discuss how the sampling requirement might be reduced by seeking a solution with minimal support.

2.5.2 Computational Phase Retrieval: Theory

Well-posedness of the computational phase retrieval problem typically requires certain assumptions about the solution's support. In fact, solutions to the phase retrieval problem with compact support have a very restrictive form [5, 63]:

Theorem 5. *Suppose $f \in L^2(\mathbb{R}^d)$ has compact support and define $F : \mathbb{C}^d \to \mathbb{C}$ to be the analytic extension of $\hat{f}$ to $\mathbb{C}^d$. If $g \in L^2(\mathbb{R}^d)$ satisfies $|\hat{g}| = |\hat{f}|$ then $\hat{g}(\boldsymbol{\xi}) = e^{i(\alpha+\boldsymbol{\beta}\cdot\boldsymbol{\xi})} F_1(\boldsymbol{\xi})\overline{F_2}(\boldsymbol{\xi})$, where $F = F_1F_2$ is an analytic factorization of F over $\mathbb{C}^d$, $\alpha \in \mathbb{R}$, and $\boldsymbol{\beta} \in \mathbb{R}^d$.*

In practice, we will have to approximate the continuous electron density with discrete functions, i.e., $f : \mathbb{Z}^d \to \mathbb{C}$. A result similar to Theorem 5 holds in the discrete case. The main idea is that the analytic extension of the power spectrum of a compactly supported function f, which can be assumed to have support only for nonnegative entries, to $\mathbb{C}^d \setminus \{0\}$

is a polynomial in $\frac{1}{\mathbf{z}}$, given by $\mathcal{Z}[f](\mathbf{z})\overline{\mathcal{Z}[f](\overline{\mathbf{z}}^{-1})}$, and, thus, the Z-transform of any other function with the same power spectrum must consist of a mix of factors from $\mathcal{Z}[f]$ and its conjugate inversion [32, 62]:

Theorem 6. *Suppose that $f, g : \mathbb{Z}^d \to \mathbb{C}$ have compact support. If $|\hat{f}| = |\hat{g}|$ then $\hat{g}(\boldsymbol{\xi}) = e^{2\pi i(\alpha+\boldsymbol{\tau})}\mathcal{Z}_1(e^{i\boldsymbol{\xi}})\overline{\mathcal{Z}_2(e^{i\boldsymbol{\xi}})}$, where $\mathcal{Z}[f] = \mathcal{Z}_1\mathcal{Z}_2$ is a polynomial factorization, $\alpha \in \mathbb{R}$, and $\boldsymbol{\tau} \in \mathbb{Z}$.*

Note that if the Z-transform of f has at most one nontrivial irreducible factor without conjugate inversion symmetry, then Theorem 6 implies that a compactly supported function is determined uniquely by its power spectrum up to translation, conjugate inversion, and multiplication by a constant phase factor. In one dimension this is very common since, by the fundamental theorem of algebra, all one-dimensional polynomials are reducible over $\mathbb{C}$ and almost never display conjugate inversion symmetry. Fortunately, reducibility is extremely rare in higher dimensions. While a few non-unique cases are known to exist in nature, such examples typically exhibit very specific types of symmetry and, thus, non-uniqueness tends to be rare for more complicated structures [38, 43]. Moreover, for $d \geq 2$ and $\mathbf{N} = (N_1, \ldots, N_d)$, if we define $P(\mathbf{N})$ to be the set of polynomials over $\mathbb{C}$ in d variables of the form

$$p(\mathbf{z}) = \sum_{\mathbf{n} \in \mathbb{Z}_{\mathbf{N}}} a_{\mathbf{n}} \mathbf{z}^{\mathbf{n}}, \tag{2.53}$$

where $a_{\mathbf{n}} \in \mathbb{C}$, then, when realizing $P(\mathbf{N})$ as $\mathbb{C}^{|\mathbf{N}|}$ and using the Lebesgue measure, the set of reducible polynomials in $P(\mathbf{N})$ forms a set of measure zero [33, 66]. Therefore, for a rectangular[1] bounded region T, phase recovery for almost all discrete functions with support contained in T has a unique solution up to these three trivial ambiguities, where we are realizing our function space as $\mathbb{C}^{|T|}$ and using the Lebesgue measure:

Corollary 1. *For $d \geq 2$, $T \subseteq \mathbb{Z}^d$ rectangular and bounded, and almost all $f : \mathbb{Z}^d \to \mathbb{C}$ with $\operatorname{supp}(f) \subseteq T$, if $g : \mathbb{Z}^d \to \mathbb{C}$ satisfies $|\hat{g}| = |\hat{f}|$ and $\operatorname{supp}(g) \subseteq T$ then for all $\mathbf{n} \in \mathbb{Z}^d$ either $g(\mathbf{n}) = e^{i\theta} f(\mathbf{n} + \boldsymbol{\tau})$ or $g(\mathbf{n}) = e^{i\theta}\overline{f(-\mathbf{n} + \boldsymbol{\tau})}$, where $\boldsymbol{\tau} \in \mathbb{Z}^d$ and $\theta \in \mathbb{R}$.*

Since the alternative solutions in Corollary 1 have the same basic structure we will consider them to be equally valid solutions:

Definition 10. *$f : \mathbb{Z}^d \to \mathbb{C}$ and $g : \mathbb{Z}^d \to \mathbb{C}$ are equal up to form if there exist $\theta \in \mathbb{R}$ and $\boldsymbol{\tau} \in \mathbb{Z}^d$ such that $g(\mathbf{x}) = e^{i\theta} f(\mathbf{x} + \boldsymbol{\tau})$ or $g(\mathbf{x}) = e^{i\theta}\overline{f(-\mathbf{x} + \boldsymbol{\tau})}$.*

Note that uniqueness of the phase retrieval problem requires knowledge of $|\hat{f}(\boldsymbol{\xi})|$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$, while, in practice, one is only able to sample $|\hat{f}|$ at a finite rate. Fortunately, full

[1] We take the convention that a rectangular region is not degenerate, i.e., it extends by at least two elements in every dimension.

recovery of $|\hat{f}|$ is possible if it is sampled at a sufficiently high rate. In the continuous case, due to the Wiener-Khinchin theorem, the Fourier transform of $|\hat{f}|^2$ is the autocorrelation A_f of f and has support twice as large as that of f. Therefore, by the Shannon-Nyquist theorem and finite support of f, $|\hat{f}|$ can be recovered exactly by sampling at the Nyquist rate for A_f, i.e., by obtaining $\{|\hat{f}(\frac{n_1}{2L_1}, \ldots, \frac{n_d}{2L_d})| : n_1, \ldots, n_d \in \mathbb{Z}\}$, where $\text{supp}(f) \subseteq [\tau_1, \tau_1 + L_1] \times \ldots \times [\tau_d, \tau_d + L_d]$ for some $\tau_1, \ldots, \tau_d \in \mathbb{R}$. In the discrete case, the Fourier transform of a signal with bounded support is a polynomial over $\mathbb{T}^d$ with finite degree and, thus, it can be retrieved with a finite number of samples:

Theorem 7. *For all $d \geq 2$, $T \subseteq \mathbb{Z}^d$ rectangular and bounded, and almost all $f : \mathbb{Z}^d \to \mathbb{C}$ that satisfy $\text{supp}(f) \subseteq [\tau_1, \tau_1 + L_1] \times \cdots \times [\tau_d, \tau_d + L_d] \subseteq T$, for $\tau_i, L_i \in \mathbb{Z}$, f is uniquely determined, up to form, by $\{|\hat{f}|(\frac{n_1}{2L_1}, \ldots, \frac{n_d}{2L_d}) : n_i \in \mathbb{Z}_{2L_i}\}$.*

The number of samples required in Theorem 7 is 4 times what is required to recover f from $\hat{f}$, known as the *Nyquist density* of f. More generally, the required sampling rate in d dimensions corresponds to sampling with 2^d times the Nyquist density. This suggests that the same sampling procedure in higher dimensions may be redundant. In fact, in [53], it was shown that it is possible to reduce the sampling requirement in higher dimensions. In particular, if another constraint, such as positivity, is used then we have:

Theorem 8. *For all $d \geq 3$, $T \subseteq \mathbb{R}^d$ rectangular and bounded, and almost all $f : \mathbb{Z}^d \to \mathbb{R}_{\geq 0}$ with $\text{supp}(f) \subseteq T$, there exists a set of sample points $\{x_i\}$ with density 4 times the Nyquist density of f such that f can be retrieved uniquely, up to form, from $\{|\hat{f}(x_i)|\}$.*

While Theorem 8 implies that the Fourier magnitudes require less sampling in higher dimensions, it has only been shown for specific sampling strategies. For example, in three dimensions, $\{x_i\} = \{(\frac{n_1}{2L_1}, \frac{n_2}{2L_2}, \frac{n_3}{L_3}) : n_1 \in \mathbb{Z}_{2L_1}, n_2 \in \mathbb{Z}_{2L_2}, n_3 \in \mathbb{Z}_{L_3}\}$, i.e., one of the dimensions is sampled at half of the rate of the others. Nevertheless, this suggests that phase retrieval may still be possible with a reduced sampling requirement.

Note that the Nyquist sampling rate for f is based on the smallest box that contains the support of f. However, it is possible that the support of f only takes up a small percentage of its containing box, which suggests that it may be possible to further reduce the sampling requirement needed for phase recovery. In fact, in [15] it was shown that one can recover a sparse discrete function from its Fourier values sampled far below the Nyquist rate, if the sampling is random:

Theorem 9. *Let $f : \mathbb{Z}_{\mathbf{N}} \to \mathbb{C}$ have support T and let Ω be a set of sampling points from a uniform random distribution on $\mathbb{Z}_{\mathbf{N}}$. For every $M > 0$, there exists C_M such that if $|T| < C_M(\log |\mathbf{N}|)^{-1}|\Omega|$ then with probability at least $1 - \mathcal{O}(|\mathbf{N}|^{-M})$ the solution to*

$$\min_{g:\mathbb{Z}_{\mathbf{N}} \to \mathbb{C}} ||g||_{\ell^1}, \text{ where } \hat{g}|_\Omega = \hat{f}|_\Omega \tag{2.54}$$

is unique and equal to f.

Note that for a discrete function f supported on T, its autocorrelation A_f has support of size at most $|T|^2$. Since uniqueness in the phase problem is tied with being able to determine the autocorrelation from a set of measurements of the power spectrum a function, Theorem 9 implies that a function with support T can be recovered with $\mathcal{O}(|T|^2 \log(|\mathbf{N}|))$ random measurements of its Fourier magnitudes [54]:

Theorem 10. *For $\mathbf{N} = (N_1, \ldots, N_d)$ with $d \geq 2$ and $M > 0$, almost all $f : \mathbb{Z}_{\mathbf{N}} \to \mathbb{C}$ can be uniquely determined up to form by $\mathcal{O}(|T|^2 \log(|\mathbf{N}|))$ random measurements of $|\hat{f}|$ with probability at least $1 - \mathcal{O}(|\mathbf{N}|^{-M})$, where T is the support of f.*

While the random sampling requirement of Theorem 10 may be difficult to realize in practice for diffractive imaging, without throwing away useful data, it suggests that it may be possible to reduce the sampling requirement by seeking a compressed solution. More specifically, note that the ℓ^1 minimization in Theorem 9 attempts to retrieve the sparsest solution that matches the random Fourier measurements, i.e., it is a convex relaxation of the minimization of the "ℓ^0 norm", which measures the support size of a function. In particular, this idea of reducing the required number of sample points by seeking a solution with minimal support has the potential to make computational phase retrieval feasible for nanocrystallography images, where sampling is limited by the scaling from the shape function.

2.5.3 Computational Phase Retrieval: Algorithms

Many phase retrieval algorithms are based on having an estimate of the solution support, or at least knowledge of a containing region. In this case, the phase retrieval problem may be formulated as follows: Given Fourier magnitude values $a : \mathbb{Z}_{\mathbf{N}} \to \mathbb{C}$ and a support $T \subseteq \mathbb{Z}_{\mathbf{N}}$, find $\rho \in M \cap S$, where $M = \{y \in \ell^2(\mathbb{Z}_{\mathbf{N}}) : |\hat{y}| = a\}$ and $S = \{y \in \ell^2(\mathbb{Z}_{\mathbf{N}}) : \text{supp}(y) \in T\}$. Given $\rho : \mathbb{Z}_{\mathbf{N}} \to \mathbb{C}$, we define the projector P_M onto M by

$$\begin{aligned} \tilde{P}_M \hat{\rho}(k) &= \begin{cases} a(k) \frac{\hat{\rho}(k)}{|\hat{\rho}(k)|}, & \text{if } \hat{\rho} \neq 0, \\ a(k), & \text{otherwise,} \end{cases} \\ P_M \rho &= \mathcal{F}^* \tilde{P}_M \mathcal{F}, \end{aligned} \tag{2.55}$$

and the projector P_S onto S by

$$P_S \rho(x) = \begin{cases} \rho(x), & \text{if } x \in S, \\ 0, & \text{if } x \notin S. \end{cases} \tag{2.56}$$

Note that while these projectors preserve real-valuedness if seeking a real-valued solution, one may have to, in practice, take the real part of the projection to remove complex terms

introduced by floating point arithmetic. We can represent the error in each set as $\varepsilon_M(\rho) = ||P_M\rho - \rho||_2$ and $\varepsilon_S(\rho) = ||P_S\rho - \rho||_2$. The projection operators can then be realized as a gradient descent step in minimizing its associated square error:

$$P_M\rho = \rho - \frac{1}{2}\nabla_\rho \varepsilon_M^2(\rho), \quad P_S\rho = \rho - \frac{1}{2}\nabla_\rho \varepsilon_S^2(\rho). \tag{2.57}$$

Since the support projector is a linear operator, while the modulus projector is nonlinear, a natural minimization algorithm is projected gradient descent on the modulus error:

$$\rho^{(n+1)} = P_S\rho^{(n)} - \frac{1}{2}\nabla_S \varepsilon_M^2(\rho^{(n)}) = P_S P_M \rho^{(n)}, \tag{2.58}$$

where $\nabla_S = P_S\nabla_\rho$ is the gradient projected onto S and $\rho^{(n)}$ is the n-th iterate, starting with some initial guess $\rho^{(0)}$. The update rule in (2.58) is known as the error reduction (ER) algorithm, alternating projection method, and the Gerchberg-Saxton algorithm [26]. One can show that every iteration of ER reduces or maintains the total error [23], i.e.,

$$\varepsilon_S^2(\rho^{(n+1)}) + \varepsilon_M^2(\rho^{(n+1)}) \leq \varepsilon_S^2(\rho^{(n)}) + \varepsilon_M^2(\rho^{(n)}). \tag{2.59}$$

However, ER tends to slow down and get trapped into local minimum as $\nabla\varepsilon_M^2$ becomes orthogonal to S.

An alternative phase retrieval technique, based on nonlinear feedback control theory, is the hybrid input-output (HIO) method, which is expressed as

$$\rho^{(n+1)} = \begin{cases} P_M\rho^{(n)}(x), & \text{if } x \in T, \\ \rho^{(n)}(x) - \beta P_M\rho^{(n)}(x), & \text{if } x \notin T, \end{cases} \tag{2.60}$$

where $\beta \in (0, 1]$ is a feedback parameter [22]. As shown in [49], HIO seeks the saddle point

$$\min_{\rho_S} \max_{\rho_{S^c}} (\varepsilon_m^2(\rho) - \varepsilon_S^2(\rho)), \text{ where } \rho_S = \rho|_S \text{ and } \rho_{S^c} = \rho|_{S^c}. \tag{2.61}$$

Consequently, HIO is able to escape most of the local minimum that plagues the ER algorithm. However, since HIO does not directly seek a minimum of the error, in practice, it is best to combine HIO and ER, e.g., by alternating between several HIO steps and several ER steps.

There are several other related phase retrieval algorithms, each with their own strengths and weaknesses, including difference maps [21], solvent flipping [2], average successive reflections [7], hybrid projection reflection [8], relaxed averaged alternating reflectors [46], saddle point optimization [48], and alternating direction methods [73].

A major drawback to the algorithms mentioned above is that they typically require a tight approximation of the support of the solution, such as from a low resolution image, in order

to be effective [52]. However, obtaining support information *a priori* is infeasible in many situations. A potential way to solve to this problem is based on searching for a solution that agrees with the diffraction the data and has minimal support, similar to the minimization in Theorem 9. One version of this approach, known as the *shrinkwrap algorithm*, periodically refines an estimate of the support after several phase retrieval iterations [51]:

Algorithm 1 (Shrinkwrap)

1. *Start with an initial guess T for the support, e.g., the support of the autocorrelation.*
2. *Apply several iterations of a phase retrieval algorithm, e.g., ER or HIO.*
3. *Convolve the current iterate with a Gaussian of width σ, $G(\mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|\mathbf{x}|^2}{2\sigma^2}}$.*
4. *Set T to be the set of all points where the current iterate is larger than τ times its maximum value.*
5. *Decrease σ and/or τ.*
6. *Repeat steps 2-5 until convergence.*

In the process of seeking a tight estimate of the support, the Shrinkwrap algorithm essentially produces, assuming convergence, a solution with minimal support. We refer to methods that attempt to find a solution with minimum support as *compressive phase retrieval* methods. In particular, these compressive phase retrieval methods may have the potential to reduce the sampling requirement for computational phase retrieval, similar to how the minimization in Theorem 9 is able to retrieve a function with sub-Nyquist sampling of its Fourier transform. Other strategies for compressive phase retrieval abandon the use of a strict support and instead use techniques to promote sparsity. Such approaches include charge flipping and Espresso [50], which use a sparsity promoting operator in place of the typical support projection, and methods which constrain the ℓ^1 norm of the solution [54].

Even though the phase retrieval problem, in general, has a unique solution up to form, the above algorithms are not guaranteed to converge to the correct solution. Common stagnation scenarios include image twinning, where an iterate gets stuck near the average of a solution and its conjugate inverse, $f(x) + \overline{f(-x)}$, which are both valid solutions by themselves, and development of phase vortices, where the vector field formed by interpreting $\hat{f}$ as a real vector valued function contains false regions of vorticity that must necessarily contain a point with 0 magnitude, which may be inconsistent with the true solution [24]. While a host of new techniques based on compressed sensing have had some success in removing such stagnation, they often require nontrivial changes to the imaging process, such as the use of random binary masks, which are, in particular, infeasible for nanocrystallography imaging [14, 70]. Consequently, practical phase retrieval algorithms may require some intervention from a user in order to detect and overcome such stagnation issues.

2.6 X-ray Nanocrystallography Reconstruction

The reconstruction problem in nanocrystallography can be stated as follows: Determine the unit cell electron density ρ from a given ensemble of diffraction images

$$I_m(x,y) = J_m r_e^2 P(x,y)\Delta\Omega(x,y)|\hat{\rho}(R_m q_\lambda(x,y))|^2 |S_m(R_m q_\lambda(x,y))|^2, \tag{2.62}$$

where the incident photon flux density J_m, crystal sizes which determine the shape transform S_m, and orientations R_m are unknown, and where the images are subjected to large amounts of shot noise, background noise, and other sources of error.

The *Monte Carlo* approach handles the unknown quantities and noise in (2.62) by averaging the intensities corresponding to each reciprocal lattice point $\boldsymbol{\xi} = q_\lambda(x,y)$ [74]. However, this requires a large number of images in order to effectively correct for the parameter variances. Also, autoindexing algorithms are complicated by the presence of partial and non-Bragg reflections and, furthermore, can only orient the images up to the symmetry of the crystal lattice. Unfortunately, in many cases, the symmetry of the crystal lattice is larger than that of the diffraction data, leading to the twinning problem, discussed in Section 2.4.2. If the twinning problem is not resolved, then one cannot retrieve the correct structure factor magnitudes via this averaging, which severely complicates the use of many of the phase recovery techniques mentioned in Section 2.5.

Current reconstruction methods deal with the twinning problem by *twinning* the data, i.e., they work with the average W of the computed structure factor magnitudes $|\hat{\rho}|$ over all of the rotational symmetries of the crystal lattice $\mathcal{S}(\mathcal{L})$:

$$W(\boldsymbol{\xi}) = \sum_{R\in\mathcal{S}_R(\mathcal{L})} |\hat{\rho}(R\boldsymbol{\xi})|. \tag{2.63}$$

However, the twinned data may not offer enough information to directly invert the image information to retrieve ρ. Consequently, reconstruction has mainly been limited to molecular replacement techniques, described in Section 2.5.1, which are able to test models of the structure against the twinned data. While many successful reconstructions from x-ray nanocrystallography images have been performed with molecular replacement, [4, 6, 12, 39, 44], fundamentally new structures cannot be determined this way, as this technique requires knowledge of a similar structure in order to function.

In order to determine the structure of fundamentally new objects with x-ray nanocrystallography, the twinning problem, if present, needs to be solved. This will be the goal of the methods presented in Chapter 3.

Chapter 3

Algorithms

3.1 Overview

Here we present a new algorithmic framework for x-ray nanocrystallography reconstruction, which aims to accurately determine the unknown parameters in Equation (2.62) and, in particular, solve the twinning problem, described in Section 2.4.2, in the presence of large amounts of noise and uncertainty. Our approach is based on the following steps:

1. We determine the Bravais characteristic vectors and orientations, up to lattice symmetry, with autoindexing techniques. While current autoindexing methods can accurately compute Bravais vector and indexing information, which is the twinned coordinate assignment of the observed Bragg peaks, they may not always calculate lattice orientation information to the precision that we will require in order to accurately deduce the crystal sizes and evaluate the shape function. In Section 3.2, we develop a new autoindexing technique, based on maximizing a cosine function, which enhances the precision of the twinned orientations by allowing us enough flexibility to utilize non-Bragg peak information.

2. Once the twinned orientations are determined, we infer the crystal sizes from a set of high resolution images of the low angle Bragg peaks, which can be obtained by placing a rear detector in the experimental setup, as in Figure 1.1. In particular, the Fourier transform of the data surrounding such a peak yields the x-ray projected autocorrelation of the crystal shape. In Section 3.3, we describe how to retrieve the associated crystal sizes by segmenting the image of this projected autocorrelated crystal.

3. The twinned orientation and crystal size information is then used to approximate the structure factor magnitudes for each reciprocal lattice point. However, the approximated structure factor magnitudes are only valid up to multiplication by the, unknown, incident photon flux densities, which differ for each image. Furthermore, due to the

twinning problem, the exact coordinates for the structure factor magnitudes are ambiguous, i.e., each reciprocal lattice point could correspond to several different structure factor magnitudes. In Section 3.4, we develop a procedure to determine the unknown incident photon flux densities and the possible structure factor magnitudes belonging to each reciprocal lattice point by using a multi-stage expectation maximization algorithm, which alternates between scaling the structure factor magnitudes and modeling their possible values via a multi-modal Gaussian distribution.

4. In Section 3.5, we develop a method for solving the twinning problem by utilizing the multi-modal models in step 3. In particular, we use these models to construct a graphical model of the structure factor magnitude concurrency, which describes how often pairs of magnitude values from different reciprocal lattice points occur within the same image. The solution to the twinning problem is then formulated as finding the maximum edge weight clique in this graph. While the maximal edge weight clique problem is, in general, NP-hard, we develop a greedy approach which is exact for the twinning problem in the absence of noise and still highly accurate in the face of large amounts of variation in the computed structure factor magnitudes. This algorithm yields detwinned structure factor magnitudes at each of the reciprocal lattice points, which we then use to compute the full orientations for each image. We can then assemble the three-dimensional volume of structure factor magnitudes by performing a weighted average over all of the images at each reciprocal lattice point and, optionally, for non-lattice points as well.

5. Now the phases can be recovered by using one of the techniques in Section 2.5.1. In particular, in Section 3.6, we develop a sampling strategy, which utilizes non-Bragg data, that allows computational phasing with only Fourier magnitude information if the orientations are determined up to symmetry of the utilized non-Bragg data. However, if the Laue symmetry of the Bragg data is greater than the symmetry of this non-Bragg data, then our computational phase retrieval approach may still be viable if one also solves the twinning problem data on the utilized non-Bragg data.

We now give a detailed description of the above steps.

3.2 Autoindexing

While the autoindexing techniques discussed in Section 2.4.1 have been used successfully in conventional crystallography, their robustness in nanocrystallography is complicated by the presence of partial reflections, non-Bragg reflections, and low peak counts. In particular, as the name suggests, autoindexing was originally intended to simply index the Bragg reflections, i.e., determine what reciprocal lattice points they correspond to, instead of directly computing the twinned orientation associated to each image. However, we will require

highly precise orientation information in order to determine the crystal sizes and evaluate the shape function. Unfortunately, the images consist mainly of partial reflections, which smear out the locations of the Bragg peaks and, thus, make it difficult to calculate precise orientations unless one uses a large number of these reflections, which may not be present in such quantities in nanocrystal diffraction images.

We now introduce an alternative approach for accurately determining nanocrystal diffraction image orientations, up to lattice symmetry, which is based on incorporating non-Bragg reflections that occur on lines between adjacent reciprocal lattice points. In order to utilize this approach we require a flexible autoindexing algorithm, which we present below.

3.2.1 Bravais Characteristic Vector Calculation

Our autoindexing approach starts by using one of the commonly used autoindexing methods, mentioned in Section 2.4.1, to determine the crystal lattice system, consisting of the lengths of and angles between the Bravais vectors, i.e., the vectors are determined for some reference configuration. In particular, this reference Bravais vector information is very robust as it takes into account the entire ensemble of images, as opposed to the orientation information, which is calculated on a per image basis. We then use this information to orient the images, up to lattice symmetry, with high precision.

Recall that for a crystal lattice $\mathcal{L}$ with Bravais characteristic vectors $(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3)$, the Bragg peaks occur at reciprocal lattice points $\boldsymbol{\xi} \in \hat{\mathcal{L}}$, which can be represented in terms of the Bravais vectors $(\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \hat{\mathbf{h}}_3)$:

$$\boldsymbol{\xi} = \sum_{j=1}^{3} n_j \hat{\mathbf{h}}_j, \text{ where } n_j \in \mathbb{Z}. \tag{3.1}$$

By Equation (2.23), we have the following property:

$$\mathbf{h}_i \cdot \hat{\mathbf{h}}_j = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

Therefore, by combining Equations (3.1) and (3.2), for every reciprocal lattice point $\boldsymbol{\xi} \in \hat{\mathcal{L}}$, we have that

$$\mathbf{h}_j \cdot \boldsymbol{\xi} = n_j. \tag{3.3}$$

Note that we currently only know the lengths L_j and relative angles of the Bravais vectors, but not their directions. Therefore, if we express the Bravais vectors as $\mathbf{h}_j = L_j \mathbf{d}_j$, where $|\mathbf{d}_j| = 1$, our goal for each image is to find $\mathbf{d}_j$. In particular, by Equation (3.3), we have that

$$\cos(2\pi L_j \mathbf{d}_j \cdot \boldsymbol{\xi}) = 1. \tag{3.4}$$

Assume that for a given image we can determine the set of reciprocal space coordinates B_r for each measured Bragg peak, e.g., by applying a threshold to the image intensities and then applying the q map in Equation 2.17 to the associated detector coordinates. Then, by Equation (3.4), the Bravais vector directions $\mathbf{d}_j$ can be recovered, up to multiplication by -1, by seeking the maximizer of the sum of (3.4) over B_r:

$$\max_{|\mathbf{d}|=1} \sum_{\boldsymbol{\xi} \in Br} \cos(2\pi L_j \mathbf{d} \cdot \boldsymbol{\xi}). \tag{3.5}$$

This Bravais direction recovery process is illustrated in Figures 3.1 - 3.3.

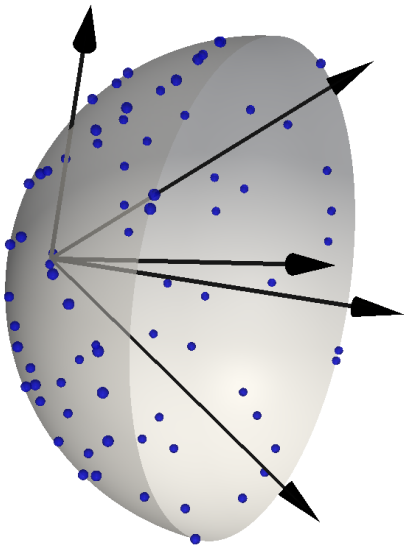


Figure 3.1: The Bragg peaks (blue) recorded in an image are mapped to the Ewald sphere (gray) via the q map. The periodicity of the peaks, when projected onto the test directions (black), is used to determine the directions of the Bravais characteristic vectors.

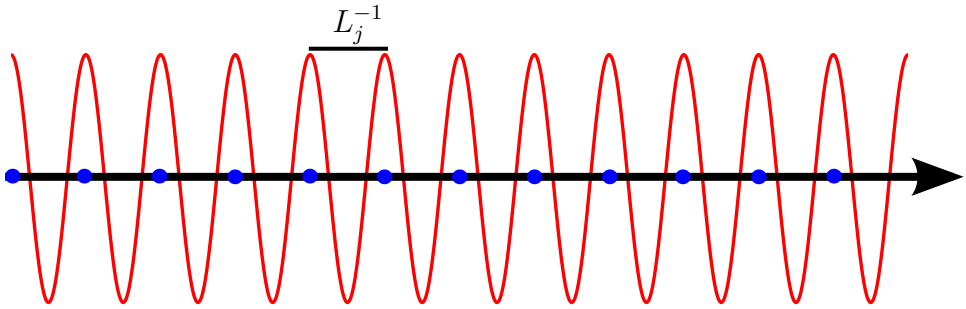


Figure 3.2: Example of Bragg peaks (blue) being projected onto a Bravais characteristic vector (black). The projected peaks line up exactly with the local maxima of $\cos(2\pi L_j x)$ (red).

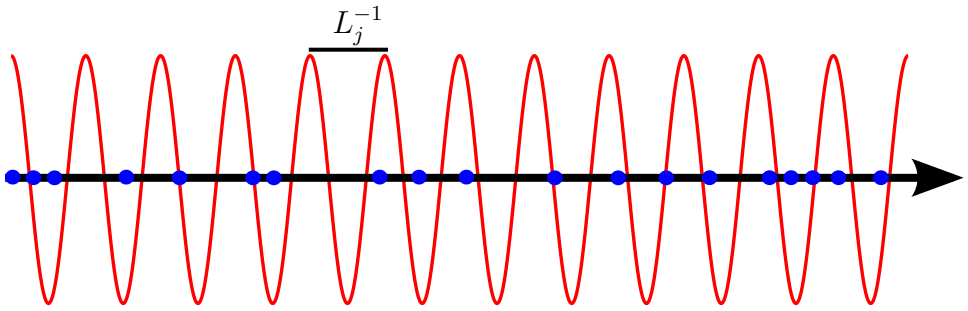


Figure 3.3: Example of Bragg peaks (blue) being projected onto a non-Bravais vector (black). The projected peaks do not line up with the local maxima of $\cos(2\pi L_j x)$ (red).

As stated, the optimization problem in (3.5) is mainly a reformulation of autoindexing. In fact, the methods in Section 2.4.1 can be seen as an efficient way to evaluate the sum in (3.5) over several lengths and/or directions, but they are limited by the resolution of their associated discretizations. However, we will directly solve this optimization problem, which allows us to precisely control the lengths and directions we search over and gives us a way to incorporate non-Bragg peaks, which we now discuss.

Consider a reflection located at a point $\boldsymbol{\xi}$ on the line between two reciprocal lattice points, i.e.,

$$\boldsymbol{\xi} = \sum_{j=1}^{2} n_{\sigma(j)} \hat{\mathbf{h}}_{\sigma(j)} + w \hat{\mathbf{h}}_{\sigma(k)}, \;\; w \in \mathbb{R} \setminus \mathbb{Z}, \tag{3.6}$$

where σ is some permutation of the indices. We will refer to a point satisfying (3.1) as a *primary reciprocal lattice point* and one satisfying (3.6) as a *secondary reciprocal lattice point*.

For every secondary reciprocal lattice point $\boldsymbol{\xi}$, we have that

$$\mathbf{h}_j \cdot \boldsymbol{\xi} = \begin{cases} n_j, & \text{if } j \in \{\sigma(1), \sigma(2)\}, \\ w, & \text{otherwise.} \end{cases} \tag{3.7}$$

Therefore, if incorporated into the sum in (3.5), each of these secondary lattice points can be used to help determine the directions for two of the Bravais vectors, but not the third. Note that for a Bravais direction $\mathbf{d}_j$ we can separate the set B_{rs} of reciprocal coordinates of the measured primary and secondary Bragg peaks into the sets $B_g = \{\boldsymbol{\xi} \in B_{rs} : L_j \mathbf{d}_j \cdot \boldsymbol{\xi} \in \mathbb{Z}\}$ and $B_b = \{\boldsymbol{\xi} \in B_{rs} : L_j \mathbf{d}_j \cdot \boldsymbol{\xi} \in \mathbb{R} \setminus \mathbb{Z}\}$. The sum in (3.5) now becomes

$$\sum_{\boldsymbol{\xi} \in B_{rs}} \cos(2\pi L_j \mathbf{d}_j \cdot \boldsymbol{\xi}) = \sum_{\boldsymbol{\xi} \in B_g} 1 + \sum_{\boldsymbol{\xi} \in B_b} \omega_{\boldsymbol{\xi}}, \;\; \text{where } \omega_{\boldsymbol{\xi}} \in [-1, 1). \tag{3.8}$$

In most cases, for a given Bravais direction, only a fixed percentage of the primary and secondary reciprocal lattice points will be in B_b. Even though the sum over B_b averages to

zero if enough points are used, one can obtain better results by attempting to remove terms belonging to B_b. In particular, we can filter out these terms by sorting the cosine values and only considering the set B_p consisting of the p points with the largest cosine value. With this framework, we now search for the Bravais directions by solving

$$\max_{|\mathbf{d}|=1} \sum_{\boldsymbol{\xi} \in B_p} \cos(2\pi L_j \mathbf{d} \cdot \boldsymbol{\xi}). \tag{3.9}$$

We illustrate the recovery of a Bravais direction with both primary and secondary reciprocal lattice points in Figure 3.4.

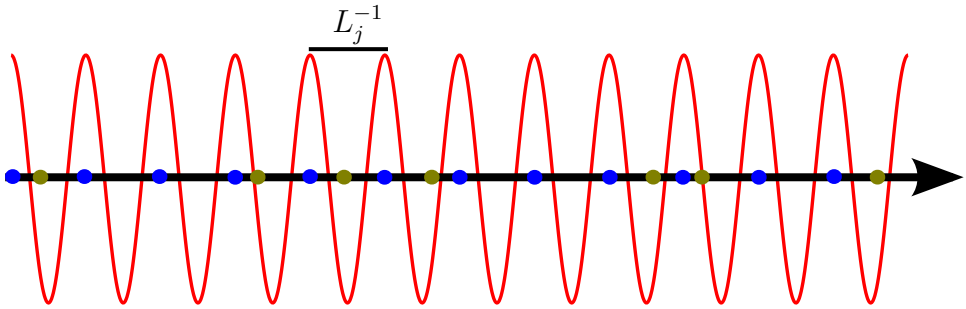


Figure 3.4: Example of primary and secondary Bragg peaks being projected onto a Bravais characteristic vector (black). The set of projected peaks in B_g (blue), consisting of primary reciprocal lattice points and secondary reciprocal lattice points with an integer $\hat{\mathbf{h}}_j$ component, line up exactly with the local maxima of $\cos(2\pi L_j x)$ (red), while those in B_b (yellow), consisting of secondary reciprocal lattice points that have a non-integer $\hat{\mathbf{h}}_j$ component, do not. Removal of the yellow points allows one to still detect the Bravais vector.

In practice, most of the recorded reflections are partial reflections and, thus, will not lie directly on a primary or secondary reciprocal lattice point. In this case, (3.9) seeks the direction that best fits the reflections in the image. If multiple Bravais vectors have the same length, one will have to search for multiple solutions which are separated approximately by the known relative angles between these vectors.

3.2.2 Direction Sampling

In order to search for the Bravais directions, we will need to sample the half unit sphere. For optimal efficiency we would like this sampling to be as uniformly distributed as possible. In particular, we utilize the method in [45] to generate an approximately uniform distribution:

Algorithm 2
for $i = 1 : N_t$ **do**
$dt \leftarrow \frac{\pi}{N_t}$

$\theta \leftarrow (i - \frac{1}{2})dt$
$dp \leftarrow \frac{dt}{\sin\theta}$
for $j = 1 : \left\lfloor \frac{\pi}{dp} + \frac{1}{2} \right\rfloor$ **do**
$\psi \leftarrow (j - \frac{1}{2})dp$
$\mathbf{d}_{i,j} \leftarrow (\cos(\theta), \sin(\theta)\cos(\psi), \sin(\theta)\sin(\psi))$
end for
end for

After candidate Bravais directions are located, we generate a finer set of sample directions around them, by restricting the angular range in Algorithm 2, and repeat the search process on this finer sampling. This resampling strategy can be repeated several times, until the desired precision is achieved. Also, note that if one of the Bravais directions is not found, we can use the known relative angles between the other two Bravais directions to narrow down the search for the missing one. For example, if the Bravais vectors are known to be orthogonal, the missing directions can be retrieved as the cross product of the other two.

3.2.3 Computing the Lattice Orientations

Once the Bravais directions $D = (\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3)$ for an image are located, we can use Equation (2.44) to retrieve an approximation $\tilde{R}_a$ to the orientation matrix by utilizing the known reference configuration of the Bravais directions $B = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$:

$$\tilde{R}_a = DB^{-1}. \tag{3.10}$$

If $|\det(\tilde{R}_a)|$ is far from unity, this indicates that the autoindexing procedure failed to compute accurate Bravais directions and, thus, we reject $\tilde{R}_a$. In practice, $\tilde{R}_a$ might not be an exact rotation matrix, i.e., it might not satisfy $\tilde{R}_a^T \tilde{R}_a = I$ and $\det(\tilde{R}_a) = 1$. Therefore, we first enforce the determinant to be positive, by multiplying a column vector by -1 if necessary, and then find the closest rotation matrix $\tilde{R}$ by using the singular value decomposition of $\tilde{R}_a$,

$$\tilde{R}_a = U\Sigma V^T, \text{ where } U^T U = V^T V = I \text{ and } \Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3) \text{ with } \sigma_j \geq 0, \tag{3.11}$$

to calculate $\tilde{R}$ via

$$\tilde{R} = UV^T. \tag{3.12}$$

The matrix $\tilde{R}$ is then used as the approximation to the image orientation, up to lattice symmetry.

3.2.4 Summary

We summarize our autoindexing approach as follows:

Algorithm 3

1. *Compute the Bravais vectors for a reference configuration from the image ensemble with one of the commonly used autoindexing algorithms in Section 2.4.1. In particular, this yields the lengths* (L_1, L_2, L_3) *and directions* $B = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ *in a reference frame.*

2. *Determine the set* B_{rs} *of the reflections on primary or secondary reciprocal lattice points by locating reflections which are local maximums and whose measured intensities are a set tolerance* τ_1 *above the background.*

3. *Create an approximately uniform sampling of search directions on the half unit sphere with Algorithm 2.*

4. *For each sample direction* $\mathbf{d}$ *perform the following:*

   4.1. *For each lattice length* L_j*, compute* $C_j = \{\cos(2\pi L_j \mathbf{d} \cdot \boldsymbol{\xi}) : \boldsymbol{\xi} \in B_{rs}\}$.

   4.2. *Sort each* C_j.

   4.3. *For each* j*, set* $s_j(\mathbf{d})$ *to be the sum of the top* p *elements of* C_j.

5. *The direction* $\mathbf{d}$ *with the largest value of* $s_j(\mathbf{d})$ *is taken to be a candidate for the Bravais direction* $\mathbf{d}_j$*. However, if* $s_j(\mathbf{d}) < \tau_2 p$*, for some fixed tolerance* τ_2*, then reject* $\mathbf{d}$*. If two Bravais vector lengths are equal then, in order to avoid duplication, enforce their corresponding candidate directions to have angles which differ by some fixed tolerance.*

6. *If a candidate for one of the Bravais directions is not found then approximate it with the set of directions which have the correct angles in relation to the other candidates.*

7. *Repeat steps 3-6 by using a finer sampling supported around the candidate Bravais directions and continue until the desired precision is reached.*

8. *Set the approximate Bravais directions* $D = (\mathbf{d}_1, \mathbf{d}_2, \mathbf{d_3})$ *to be the associated candidate directions.*

9. *Form the candidate orientation* $\tilde{R}_a = DB^{-1}$.

10. *If* $|\det(\tilde{R}_a)| < \tau_3$*, for some fixed tolerance* τ_3*, then report that the image has failed to be autoindexed.*

11. *Ensure that* $\det(\tilde{R}_a) > 0$ *by multiplying a column by* -1 *if the determinant is negative.*

12. *Compute the singular value decomposition* $\tilde{R}_a = U\Sigma V^T$ *and return the rotation matrix* $\tilde{R} = UV^T$ *and associated Bravais characteristic vectors:*

$$(\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3) = \tilde{R}(L_1\mathbf{b}_1, L_2\mathbf{b}_2, L_3\mathbf{b}_3).$$

3.3 Crystal Size Determination

In order to compute the structure factor magnitudes from the measured intensities, the squared magnitude of the shape function must be divided out of the intensity measurements in Equation (2.62). Note that near a Bragg peak, the shape function grows quadratically with the crystal size, which can vary by several orders of magnitude in the nanocrystal ensemble. Therefore, in order to obtain highly accurate structure factors, the crystal sizes need to be determined. We accomplish this by analyzing the intensities around low angle Bragg peaks in a high resolution image, such as from a rear detector image, depicted in Figure 3.5. The intensities around these peaks reveal the shape of the shape function, whose Fourier transform determines the crystal sizes. We note that a similar idea was used in [16] to gather statistics about the crystal size distribution by applying several iterations of a phase retrieval algorithm on a few select images, but this information was not used in the structure factor calculation. In comparison, our approach only requires one Fourier transform calculation and is applied to every recorded low angle image.

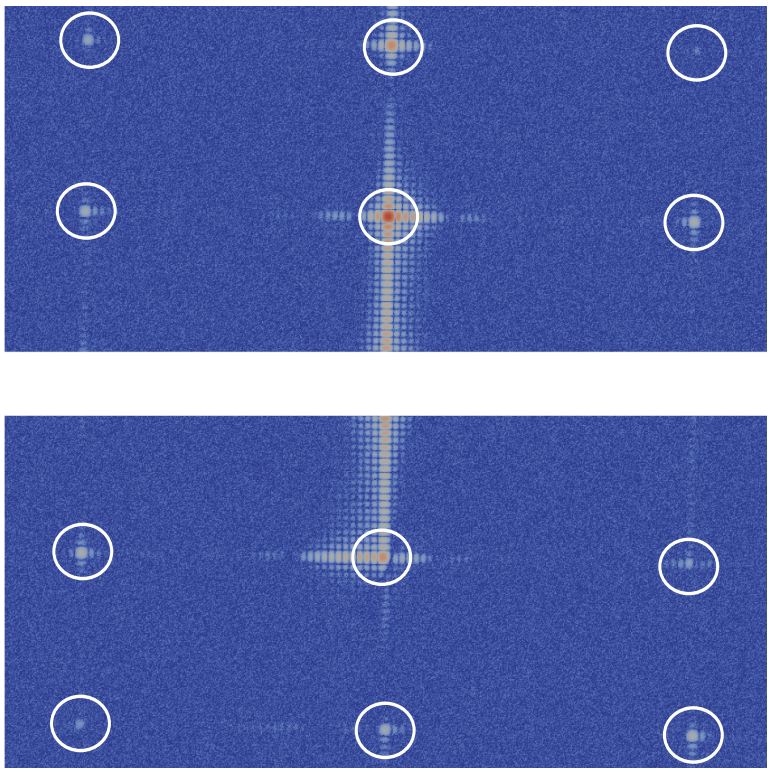


Figure 3.5: Example of a low angle image. The shape of the shape function around the Bragg peaks (circled) can be used to determine the crystal sizes.

3.3.1 Fourier Analysis of the Shape Function

For an image I with orientation R, consider its restriction I_r to a small neighborhood $\mathcal{N}_b$ centered at a low angle Bragg peak with detector coordinates $\mathbf{x}_o \in \mathbb{R}^2$ corresponding to the reciprocal lattice point $\boldsymbol{\xi} \in \hat{\mathcal{L}}$ where $|\boldsymbol{\xi}|$ is small. In $\mathcal{N}_b$ the q map in Equation (2.17) can be approximated as a linear map by Taylor expanding $q(\mathbf{x})$:

$$q(\mathbf{x}) \approx R^T \boldsymbol{\xi} + \frac{1}{\lambda D} \begin{pmatrix} \mathbf{x} - \mathbf{x}_o \\ 0 \end{pmatrix}. \tag{3.13}$$

Furthermore, in the neighborhood $\mathcal{N}_b$, the structure factors, polarization factor, and angle subtended by a pixel are approximately constant. Therefore, by Equation (2.31), the measured intensities in N_b are, up to a constant factor C, approximately equal to the squared norm of the shape function S on a plane:

$$I_r(\mathbf{x}) \approx C|S(\boldsymbol{\xi} + K\, R(\tilde{\mathbf{x}}))|^2, \tag{3.14}$$

where $K = (\lambda D)^{-1}$ and $\tilde{\mathbf{x}} = (\mathbf{x}, 0)$. Since S is invariant to translation on the lattice, Equation (3.14) can be rewritten as

$$I_r(\mathbf{x}) \approx C|S(K\, R(\tilde{\mathbf{x}}))|^2. \tag{3.15}$$

Note that since S is symmetric with respect to rotation by elements of the lattice rotational symmetry group $\mathcal{S}_R(\mathcal{L})$, we only need to know the orientation modulo $\mathcal{S}_R(\mathcal{L})$, i.e., we can use the twinned orientations from the autoindexing process here. If we denote $G(\mathbf{x}) = C|S(K\mathbf{x})|^2$, then we can realize (3.15) as the restriction of G to the plane rotated by R:

$$I_r \approx G|_{R(\mathbb{R}^2)}. \tag{3.16}$$

Due to the Fourier projection slice theorem, the Fourier transform of (3.16) is given by

$$\hat{I}_r(\gamma) \approx (P_{R^{(3)}} \hat{G})(\gamma), \tag{3.17}$$

where $R^{(3)}$ is the third column vector of R, which is the normal vector to $R(\mathbb{R}^2)$. Now, by applying the Wiener-Khinchin theorem to $\hat{G}$ in (3.17), we can represent the right hand side as the x-ray projected autocorrelation of the Dirac comb $\Delta_{\mathcal{L}_\mathcal{C}}$ of the finite crystal lattice $\mathcal{L}_\mathcal{C}$ in an unrotated reference frame:

$$\hat{I}_r(\gamma) \approx K^{-1}(P_{R^{(3)}} A \Delta_{\mathcal{L}_\mathcal{C}})(K^{-1}\gamma). \tag{3.18}$$

An example of this Fourier transform is depicted in Figure 3.6. Note that the support of this projected autocorrelation is given by the *Minkowski sum* of the rotated projected crystal, i.e.,

$$\operatorname{supp}(P_{R^{(3)}} A \Delta_{\mathcal{L}_\mathcal{C}}) = \{x_1 + x_2, y_1 + y_2 : (x_1, y_1, z_1), (x_2, y_2, z_2) \in R(\mathcal{L}_\mathcal{C})\}. \tag{3.19}$$

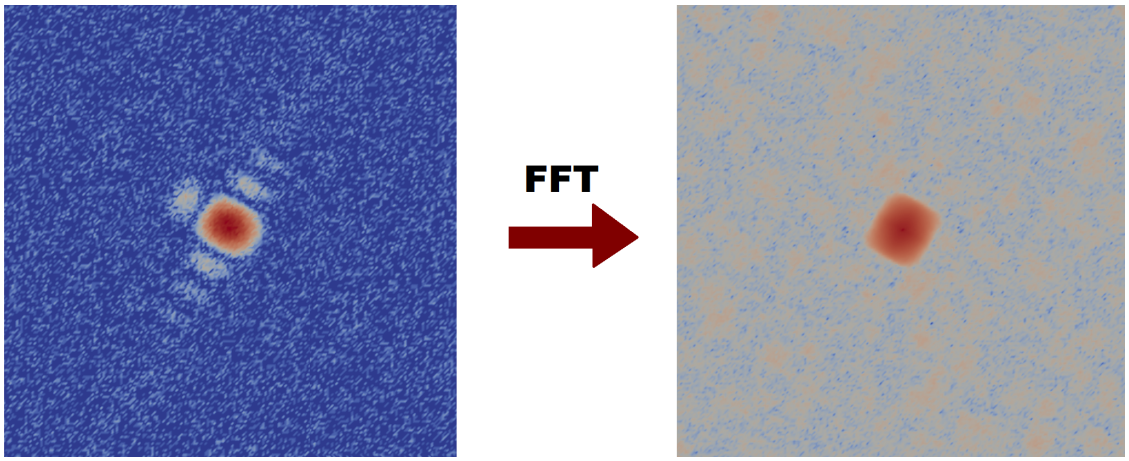


Figure 3.6: The Fourier transform of the shape function (left) around a low angle peak reveals the projected autocorrelated crystal (right).

Now recall that the *convex hull* $\mathrm{Conv}(X)$ of a set X is defined by:

$$\mathrm{Conv}(X) = \left\{ \sum_{x \in X} \alpha_x x : \alpha_x \geq 0 \text{ and } \sum_{x \in X} \alpha_x = 1 \right\}. \tag{3.20}$$

If we approximate the crystal lattice as $\mathcal{L}_\mathcal{C} = \{\sum_{j=1}^3 n_j \mathbf{h}_j : n_j \in \mathbb{Z}_{N_j}\}$, where $\mathbf{N} = (N_1, N_2, N_3)$ are the crystal sizes for each Bravais direction, then the convex hull of the projected autocorrelation can be expressed as

$$\mathrm{Conv}(\mathrm{supp}(P_{R^{(3)}} A \Delta_{\mathcal{L}_\mathcal{C}})) = \mathrm{Conv} \left\{ (x, y) : (x, y, z) = R \sum_{j=1}^3 \pm (N_j - 1) \mathbf{h}_j \right\}. \tag{3.21}$$

Therefore, by computing $\hat{I}_r$, we can almost always deduce the size of the crystal by analyzing this convex hull. The one exception to this rule is when one of the rotated Bravais vectors $R\mathbf{h}_j$ is orthogonal to the detector plane, in which case any value of N_j produces the same convex hull. In general, the boundary of this convex hull consists of a series of line segments, with three normals $(\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3)$ along with their three anti-parallel directions, which can be found by computing the convex hull of the rotated projected autocorrelated reciprocal unit cell, i.e., the right hand side of (3.21) with $N_j = 2$ for each j. Note that the extent of the convex hull in the direction of $\mathbf{n}_j$ must also be equal to the extent of the unprojected crystal in this direction. In particular, if we set

$$b_i = \max\big\{ |\mathbf{n}_i \cdot \mathbf{x}| : \mathbf{x} \in \mathrm{Conv}(\mathrm{supp}(P_{R^{(3)}} A \Delta_{\mathcal{L}_\mathcal{C}})) \big\}, \tag{3.22}$$

then we have that

$$b_i = \max\left\{ |\mathbf{n}_i \cdot \mathbf{x}| : \mathbf{x} \in \text{Conv}\left\{ (x,y) : (x,y,z) = R\sum_{j=1}^{3} \pm(N_j - 1)\mathbf{h}_j \right\} \right\} \tag{3.23}$$

$$= \max \sum_{j=1}^{3} |\mathbf{n}_i^T R \mathbf{h}_j|(N_j - 1). \tag{3.24}$$

Therefore, if we define a matrix A by

$$A_{ij} = |\mathbf{n}_i^T R \mathbf{h}_j|, \tag{3.25}$$

then the crystal sizes $\mathbf{N}$ can be retrieved by solving

$$A(\mathbf{N} - \mathbf{1}) = b, \tag{3.26}$$

where $\mathbf{1} = (1,1,1)$.

In the above analysis, we assumed that image passed directly through a reciprocal lattice point $\boldsymbol{\xi}$, which may not always happen in practice. Now, suppose that $\boldsymbol{\xi} + \boldsymbol{\nu}$ is the closest point to $\boldsymbol{\xi}$, which the image samples. Define $H(\mathbf{x}) = C|S(K\mathbf{x} + \boldsymbol{\nu})|^2$. In this case, Equation (3.16) becomes

$$I_r \approx H|_{R(\mathbb{R}^2)}. \tag{3.27}$$

Then, by the translational property of the Fourier transform, (3.18) becomes

$$\hat{I}_r(\gamma) \approx K^{-1}(P_{R^{(3)}}(e_{\boldsymbol{\nu}} A \Delta_{\mathcal{L}_\mathcal{C}}))(K^{-1}\gamma), \text{ where } e_{\boldsymbol{\nu}}(\mathbf{x}) = e^{2\pi i \boldsymbol{\nu}\cdot\mathbf{x}}. \tag{3.28}$$

In almost every case, in the sense of measure theory, the inclusion of the exponential function, $e_{\boldsymbol{\nu}}$, in (3.28) does not affect the support of the convex hull. Therefore, the methods described above can still use I_r to retrieve the crystal sizes. However, if $\boldsymbol{\nu}$ is too big, then there may be large oscillations in $\hat{I}_r(\gamma)$, which can make detecting the convex hull difficult.

3.3.2 Image Segmentation

While the convex hull in Equation (3.21) is typically very pronounced, see Figure 3.6, we require an accurate way of automating its retrieval. In particular, we would like to segment the projected autocorrelated crystal, represented by $|\hat{I}_r|$, from the background, which consists mainly of small oscillations due to noise and approximation errors. In theory, for a grid aligned image with sufficient resolution and signal, one could potentially see several sharp spikes in the image corresponding to the peaks of the associated Delta comb. However, in most practical cases, the Fourier transform of the intensities consists of a large peak, which gradually fades into the background as one approaches the end of the convex hull, see Figure 3.7. Therefore, the main difficulty is finding a cutoff value, which separates the crystal from

the background noise. We then approximate the support of the projected autocorrelated crystal by the set C, which consists of all pixels with values above this cutoff.

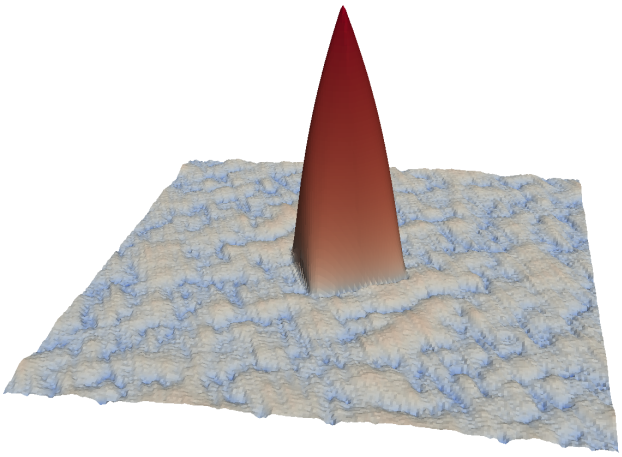


Figure 3.7: Fourier transform of the shape function around a low angle peak, visualized as a height function which maps the pixels to the Fourier magnitudes.

Note that if we start near the top of the peak and work our way down, the values decrease monotonically until one reaches the bottom, where the oscillations begin. This idea motivates the following approach. We initialize C to consist of all pixels whose value is greater than a fixed percentage τ_1 of the largest value $v_{\max}$ in the image, ignoring the origin. Here we ignore the origin value since it tends to be uninformative as it picks up all of the noise in the image. We then sort the remaining values and traverse the sorted list, adding each traversed pixel to C, until we reach a pixel which is more than some threshold τ_2, typically a few pixels, away from all of the current pixels in C. Such a jump suggests that one has reached the bottom of the peak and has begun to see the oscillations. This approach is summarized in Algorithm 4, where we utilize a queue Q, formed by pairs $(|\hat{I}_r(i,j)|, (i,j))$, which we sort by the first element of each pair.

Algorithm 4

$v_{\max} \leftarrow \max_{i,j} |\hat{I}_r(i,j)|$
$C \leftarrow \{(i,j) : |\hat{I}_r(i,j)| \geq \tau_1 v_{\max}\}$
for all $i,j \notin C$ **do**
    Q.push($(|\hat{I}_r(i,j)|, (i,j))$)
end for
Sort(Q)
loop
    $(|\hat{I}_r(m,n)|, (m,n)) \leftarrow Q$.pop()
    $dist \leftarrow \min_{(i,j)\in C} |(i,j) - (m,n)|$

```
    if dist > τ then
        return C
    end if
    C ← C ∪ {(m, n)}
end loop
```

The result of the above segmentation is illustrated for an example in Figure 3.8. Algorithm 4 can be made to run in $\mathcal{O}(n\log(n))$ time, where n is the total number of pixels, with various geometric data structures for determining if *dist* surpasses τ, e.g., by maintaining back pointers from the grid to the elements of C. If $|\hat{I}_r|$ happens to contain several sharp peaks from the delta comb then one can still use this segmentation technique by first applying a low pass filter to $|\hat{I}_r|$.

Note that C gives the set of unitless pixel coordinates which approximate the convex hull of the projected autocorrelated crystal. Therefore, in order to determine the crystal sizes, these coordinates should first be scaled by their corresponding units, which, for an image with $N_p \times N_p$ pixels of size $dx \times dx$, are given by $\frac{\lambda D}{N_p dx}$.

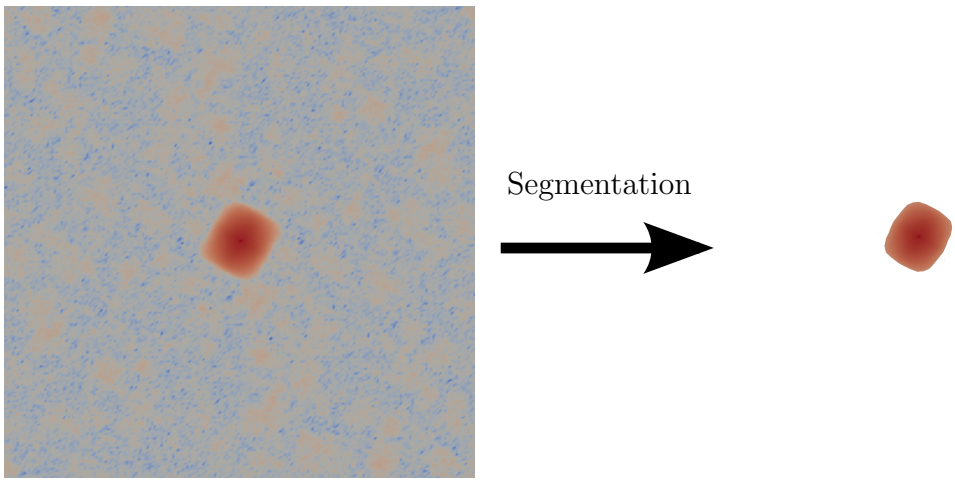


Figure 3.8: Segmentation of a projected autocorrelated crystal.

3.3.3 Summary

We now summarize our approach for determining the crystal sizes from a low angle high resolution diffraction pattern I with pixel size $dx \times dx$:

Algorithm 5

1. *Search for the peak, with location* $\mathbf{x}_o$*, in* I *with the largest intensity.*

2. *Set* I_r *to be* I *restricted to* $N_p \times N_p$ *pixels around* $\mathbf{x}_o$*.*

3. *Compute the Fourier transform* $\hat{I}_r$.

4. *Calculate the convex hull of the rotated projected autocorrelated reciprocal unit cell:*

$$\text{Conv}\{(x,y) : (x,y,z) = R(\pm\mathbf{h}_1 \pm \mathbf{h}_2 \pm \mathbf{h}_3)\}.$$

5. *Determine the three non-collinear normals* $(\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3)$ *of the convex hull edges.*

6. *Segment* $|\hat{I}_r|$ *via Algorithm 4 and denote the result as* C.

7. *Scale every element of* C *by* $\frac{\lambda D}{N_p dx}$.

8. *Compute the extents* (b_1, b_2, b_3) *for each normal direction, i.e.,*

$$b_i = \max_{\mathbf{x}\in C}(|\mathbf{n}_i \cdot \mathbf{x}|).$$

9. *Compute the matrix*

$$A_{ij} = |\mathbf{n}_i^T R \mathbf{h}_j|.$$

10. *Solve the linear system* $A(\mathbf{N} - \mathbf{1}) = b$ *and return the crystal sizes* $\mathbf{N}$.

3.4 Structure Factor Magnitude Modeling

Once the twinned orientations $\tilde{R}_m$ and the crystal sizes $\mathbf{N}_m$ are known, we can proceed to compute an approximation $|\tilde{F}_m|^2$ to the structure factor square magnitudes $|F_m|^2$ from the image I_m by:

$$|\tilde{F}_m(\tilde{R}_m q(\mathbf{x}))|^2 = \frac{I_m(\mathbf{x})}{r_e^2 P(\mathbf{x}) \Delta\Omega(\mathbf{x}) |S_{\mathbf{N}_m}(\tilde{R}_m q(\mathbf{x}))|^2}. \tag{3.29}$$

In particular, note that since we do not know the incident photon flux density J_m, which varies between images, $|\tilde{F}_m|$ only determines $|F_m|$ up to a constant factor, with a different constant for each image. Furthermore, due to the twinning problem, one only knows the corresponding reciprocal space coordinates associated to the values of $|\tilde{F}_m|$ up to symmetry of the crystal lattice, i.e., the possible structure factor magnitudes for each reciprocal lattice point take the form of a multi-modal distribution. Moreover, these two problems are strongly coupled together since we cannot perform the scaling correction unless we know what modes to scale to and the modes are indistinguishable in the unscaled data set. Hence, we will need a method to simultaneously determine the scaling and the parameters in the multi-modal distribution.

3.4.1 Processing the Data

Since the majority of the signal is spread out in a small region around the Bragg peaks, it is prudent to average over this region in order to obtain an approximation of the structure factor magnitudes. In particular, for each image I_m and for each reciprocal lattice point $\boldsymbol{\xi}_i \in \hat{\mathcal{L}}$, that I_m potentially measures, we compute the approximate structure factor magnitudes by averaging the numerator and denominator of (3.29) in the neighboring ball $B(\boldsymbol{\xi}_i, r)$ with radius r:

$$v_{i,m} = \frac{\sum_{\tilde{R}_m q(\mathbf{x}) \in B(\boldsymbol{\xi}_i, r)} I_m(\mathbf{x})}{\sum_{\tilde{R}_m q(\mathbf{x}) \in B(\boldsymbol{\xi}_i, r)} r_e^2 P(\mathbf{x}) \Delta\Omega(\mathbf{x}) |S_{\mathbf{N}_m}(\tilde{R}_m q(\mathbf{x}))|^2}. \tag{3.30}$$

However, if the intensity $I_m(\mathbf{x})$ is below some threshold, then its signal is most likely dominated by noise, which will cause large errors due to the scaling provided by the shape function. Hence, such intensities are dropped from the sum in (3.30). Note that since we only know the orientation up to the twinning ambiguity, we also set $v_{t,m} = v_{i,m}$ for every t such that for some $R \in \mathcal{S}_R(\mathcal{L})$, $R\boldsymbol{\xi}_t = \boldsymbol{\xi}_i$. In practice, $\{v_{i,m}\}$ can be stored efficiently through a map data structure for each image, i.e., by using ordered pairs $(\boldsymbol{\xi}_{i,m}, v_{i,m})$ with an efficient lookup construct on $\boldsymbol{\xi}_{i,m}$. In order to simplify our notation, we will assume that any unmeasured values and their corresponding indices are removed from all of the remaining sets and summations.

One major difficulty in analyzing the multi-modal distribution of (3.30) is that the variance of the data scales with the size of the collected magnitudes. In particular, data generated from a Poisson distribution with mean v, described in Section 2.3.6, is approximately equal to a Gaussian distribution with a standard deviation of $\sqrt{v}$. Fortunately, we can modify the data to have a standard deviation that is largely independent of the mean, in a technique known as *variance stabilization*. In particular, if a set of data values $\{y_i\}$ is generated from a Poisson distribution then $\{\sqrt{y_i}\}$ approximately has the distribution of a Gaussian with standard deviation $\frac{1}{4}$ [28]. However, in addition to Poisson noise, the data $\{v_{i,m}\}$ is also contaminated with errors in the calculation of the shape transform, which again scales with the mean. Therefore, we apply two steps of variance stabilization to the data and, instead, work with

$$w_{i,m} = v_{i,m}^{\frac{1}{4}}. \tag{3.31}$$

3.4.2 Multi-Modal Analysis

For the moment, assume that the structure factor magnitudes are already properly scaled. Due to the twinning problem, we currently only know the orientations up to the lattice symmetry and, thus, for each reciprocal lattice point $\boldsymbol{\xi}_i$, the values of $w_{i,m}$ could correspond to K different structure factor magnitudes. In particular, for elastic scattering, K is the order of the lattice symmetry group modulo the Laue symmetry group, i.e., $K = |\mathcal{S}(\mathcal{L})|/|\mathcal{S}_L(\mathcal{C})|$, and, in most cases, $K \leq 2$. However, in the presence of strong anomalous dispersion from certain crystals or when considering non-reciprocal lattice points, K can be larger, up to

the order of the crystal rotational symmetry group $|S_R(\mathcal{C})|$. Therefore, if one were to plot a histogram of $\{w_{i,m}\}$ for $\boldsymbol{\xi}_i$, one will see K different peaks, which are smeared out as noise and uncertainty in the parameters is increased, see Figure 3.9. Our goal here will be to detect these peaks and model the associated multi-modal distribution.

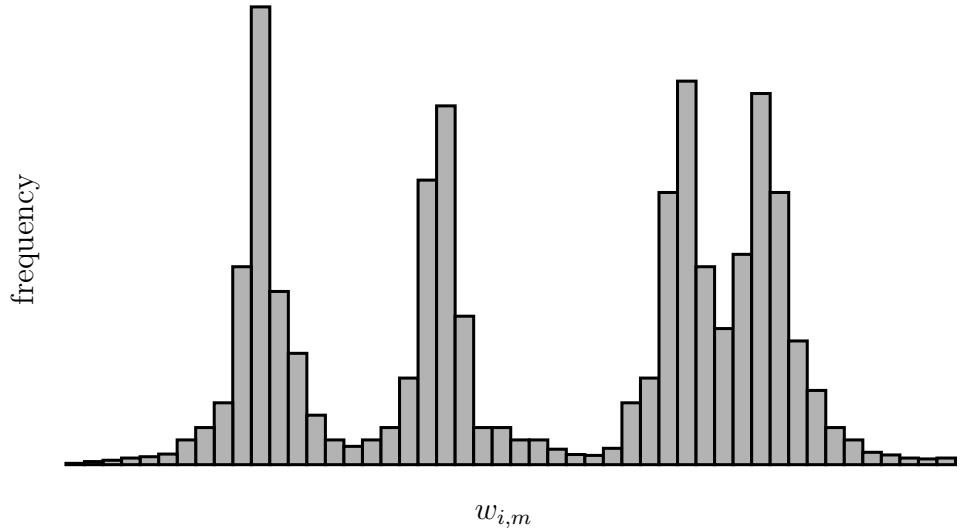


Figure 3.9: Histogram of the possible scaled variance stabilized structure factor magnitudes at a reciprocal lattice point, corresponding to a four-fold twinning problem.

In order to retrieve the set of possible structure factor magnitudes, we will model the computed values $\{w_{i,m}\}$ from each reciprocal lattice point $\boldsymbol{\xi}_i$ with a multi-modal Gaussian distribution, see Figure 3.10. Specifically, the associated probability density functions can be expressed in terms of multiple Gaussian distributions with means $\boldsymbol{\mu}_i = (\mu_{i,1}, \ldots, \mu_{i,K})$, ordered so that $\mu_{i,1} \leq \mu_{i,2} \leq \ldots \leq \mu_{i,K}$, and standard deviations $\boldsymbol{\sigma} = (\sigma_{i,1}, \ldots, \sigma_{i,K})$ by

$$p(w_{i,m}, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}, \sigma_{i,k}), \tag{3.32}$$

where

$$\mathcal{G}(w_{i,m}, \mu_{i,k}, \sigma_{i,k}) = \frac{1}{\sigma_{i,k}\sqrt{2\pi}} e^{\frac{(w_{i,m}-\mu_{i,k})^2}{2\sigma_{i,k}^2}}. \tag{3.33}$$

Given a set of data values $\{w_{i,m}\}_{m=1}^{M}$, which we model via (3.32), we can determine its associated means and standard deviations through an expectation maximization algorithm. In more detail, given some initial guess for the model parameters $\mu_{i,j}^{(0)}$ and $\sigma_{i,j}^{(0)}$, we perform several iterations of the following:

$$
\begin{aligned}
T_{i,j,m}^{(n)} &= \frac{\mathcal{G}(w_{i,m}, \mu_{i,j}^{(n-1)}, \sigma_{i,j}^{(n-1)})}{\sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}^{(n-1)}, \sigma_{i,k}^{(n-1)})} \\
\mu_{i,j}^{(n)} &= \frac{\sum_{m=1}^{M} T_{i,j,m}^{(n)} w_{i,m}}{\sum_{m=1}^{M} T_{i,j,m}^{(n)}} \\
\sigma_{i,j}^{(n)} &= \sqrt{\frac{\sum_{m=1}^{M} T_{i,j,m}^{(n)} \left(w_{i,m} - \mu_{i,j}^{(n)}\right)^2}{\sum_{m=1}^{M} T_{i,j,m}^{(n)}}}
\end{aligned}
\tag{3.34}
$$

The initialization of (3.34) can be a very delicate issue, as poor initial conditions can lead to the iterations getting stuck in local minimum far from the desired solution. In general, each orientation will have approximately the same number of samples in the distribution. Taking this into account, we initialize by separating the data into K equal bins, set $\mu_{i,j}$ to be the location in the j-th bin with the greatest sampling density, and set $\sigma_{i,j}$ to initially be less than the typical size of a bin. Furthermore, some sort of outlier rejection is typically required to make expectation maximization algorithms robust. We perform this outlier rejection by removing any $w_{i,m}$ in which $\frac{\sqrt{2\pi}}{K^2} \sum_{j=1}^{K} \sigma_{i,j}^{(n)} \sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}^{(n)}, \sigma_{i,k}^{(n)})$ is below some fixed threshold.

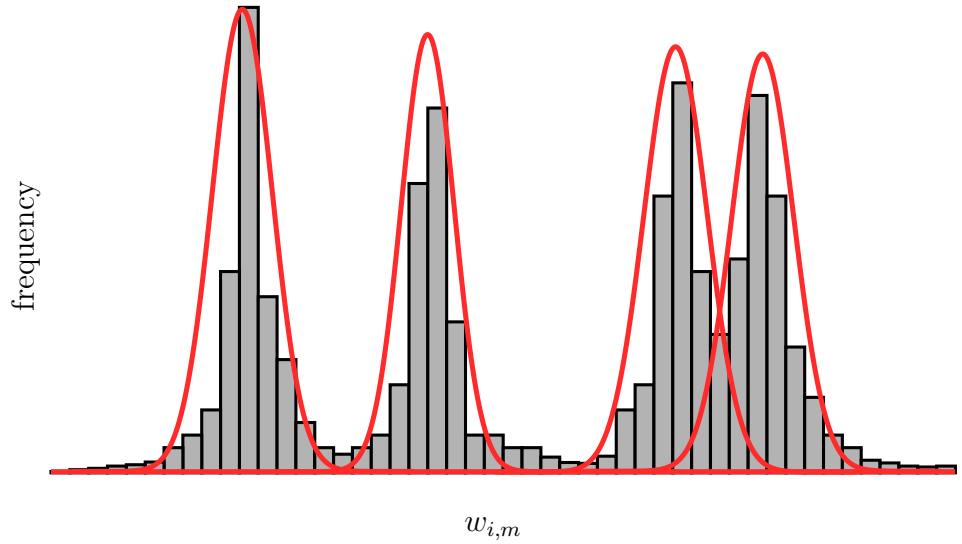


Figure 3.10: Modeling a histogram with a multi-modal Gaussian distribution (red).

3.4.3 Scaling Correction

In practice, the variance in the incident photon flux density, noise, and errors in autoindexing and crystal size determination smear out the peaks in the histogram, which makes them difficult to locate via expectation maximization, see Figure 3.11. Hence, the data must

be scaled in order to properly model the structure factor magnitudes. We will perform this scaling by seeking a scaling factor which minimizes the variance in the histograms and alternate this procedure with the expectation maximization step in (3.34).

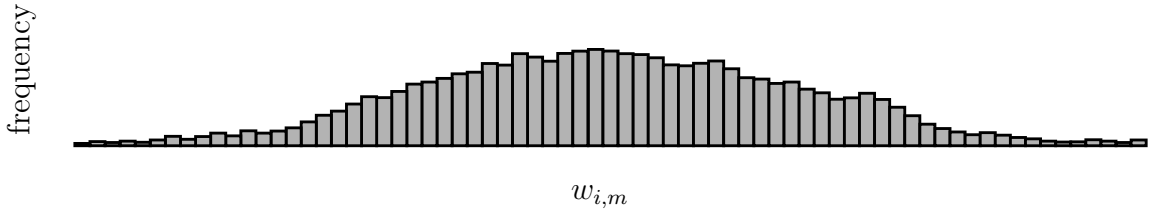


Figure 3.11: Histogram of the possible unscaled variance stabilized structure factor magnitudes for a reciprocal lattice point, corresponding to a four-fold twinning problem. The four different peaks are too smeared out to distinguish.

The scaling problem can be formulated as follows. For every image I_m we seek the scaling factor c_m which solves

$$\min_{c_m} \sum_i |c_m w_{i,m} - \mu_i^*|^2, \tag{3.35}$$

where μ_i^* is the mean of the Gaussian distribution closest to $w_{i,m}$ in the multi-modal model. However, until we have accurate scaling factors, the closest Gaussian distribution may be the incorrect one. Therefore, we weight (3.35) by the probability that each associated sample belongs to any given peak in the current multi-modal model:

$$\min_{c_m} \sum_{i,j} |(c_m w_{i,m} - \mu_{i,j}) p_{i,j,m}|^2, \text{ where } p_{i,j,m} = \frac{\mathcal{G}(w_{i,m}, \mu_{i,j}, \sigma_{i,j})}{\sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}, \sigma_{i,k})}. \tag{3.36}$$

The solution to (3.36) is given by:

$$c_m = \frac{\sum_{i,j} w_{i,m} \mu_{i,j} p_{i,j,m}^2}{\sum_{i,j} w_{i,m}^2 p_{i,j,m}^2}. \tag{3.37}$$

Once c_m is computed we use it to scale the image, i.e., we replace every $w_{i,m}$ with $c_m w_{i,m}$. This scaling step is alternated with several iterations of (3.34) until the model parameters and the scaling factors converge. In order to prevent numerical overflow/underflow, we normalize each c_m value after every step, e.g., so that $\sum_m c_m = \text{constant}$.

3.4.4 Summary

We calculate the multi-modal model parameters $\{\mu_{i,j}\}$ and $\{\sigma_{i,j}\}$ and scaling factors $\{c_m\}$ with the following steps:

Algorithm 6

1. *For each image I_m, compute the approximate unscaled structure factor square magnitudes around the recorded Bragg peaks, located at each reciprocal lattice point $\boldsymbol{\xi}_i$:*

$$v_{i,m} = \frac{\sum_{\tilde{R}_m q(\mathbf{x}) \in B(\boldsymbol{\xi}_i, r)} I_m(\mathbf{x})}{\sum_{\tilde{R}_m q(\mathbf{x}) \in B(\boldsymbol{\xi}_i, r)} r_e^2 P(\mathbf{x}) \Delta\Omega(\mathbf{x}) |S_{\mathbf{N}_m}(\tilde{R}_m q(\mathbf{x}))|^2}.$$

2. *Set $v_{t,m} = v_{i,m}$ for every t such that for some $R \in \mathcal{S}_R(\mathcal{L})$, $R\boldsymbol{\xi}_t = \boldsymbol{\xi}_i$.*

3. *Perform variance stabilization:*

$$w_{i,m} = v_{i,m}^{\frac{1}{4}}.$$

4. *For each $\boldsymbol{\xi}_i$, compute the model parameters $\mu_{i,j}$ and $\sigma_{i,j}$ via several iterations of expectation maximization, starting with some initial guess $\mu_{i,j}^{(0)}$ and $\sigma_{i,j}^{(0)}$:*

$$T_{i,j,m}^{(n)} = \frac{\mathcal{G}(w_{i,m}, \mu_{i,j}^{(n-1)}, \sigma_{i,j}^{(n-1)})}{\sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}^{(n-1)}, \sigma_{i,k}^{(n-1)})}$$

$$\mu_{i,j}^{(n)} = \frac{\sum_{m=1}^{M} T_{i,j,m}^{(n)} w_{i,m}}{\sum_{m=1}^{M} T_{i,j,m}^{(n)}}, \quad \sigma_{i,j}^{(n)} = \sqrt{\frac{\sum_{m=1}^{M} T_{i,j,m}^{(n)} \left(w_{i,m} - \mu_{i,j}^{(n)}\right)^2}{\sum_{m=1}^{M} T_{i,j,m}^{(n)}}}.$$

5. *Scale each $w_{i,m}$ with c_m, which is given by*

$$c_m = \frac{\sum_{i,j} w_{i,m} \mu_{i,j} p_{i,j,m}^2}{\sum_{i,j} w_{i,m}^2 p_{i,j,m}^2}, \text{ where } p_{i,j,m} = \frac{\mathcal{G}(w_{i,m}, \mu_{i,j}, \sigma_{i,j})}{\sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}, \sigma_{i,k})}.$$

6. *Repeat steps 4-5 until convergence and then return the values of $w_{i,m}$, $\mu_{i,j}$, $\sigma_{i,j}$, and c_m.*

If the number of modes is initially unknown, then it may be discovered by performing Algorithm 6 with different values of K until the number of modes in the histogram matches K. Additionally, we note that, if desired, non-Bragg peak data can also be modeled with the above methods.

3.5 Solving the Twinning Problem

After performing the structure factor magnitude modeling in Section 3.4, we have knowledge of up to K possible structure factor magnitudes at each reciprocal lattice point. Solving the twinning problem now amounts to deciding which of these K values belongs at each point.

In particular, note that there are K equally valid solutions, which are related to each other by applying a global rotation from the lattice rotational symmetry group. We solve this by first using the multi-modal model parameters, computed in Section 3.4, to construct a graphical model of the structure factor magnitude concurrency, i.e., the probability that two structure factor magnitudes from different points occur within the same image. Then, the solution to the twinning problem can be formulated as finding the maximum edge weight clique of this graph, which we compute efficiently with a greedy approach.

3.5.1 Graphical Modeling of Structure Factor Magnitude Concurrency

Given the scaled variance stabilized structure factor magnitudes $\{w_{i,m}\}$, means $\{\mu_{i,j}\}$ with $\mu_{i,1} \leq \mu_{i,2} \leq \ldots \leq \mu_{i,K}$, and standard deviations $\{\sigma_{i,j}\}$ for the m-th image and i-th reciprocal lattice point, we define the *occurrence probability* of $\mu_{i,j}$ in image I_m by

$$p(\mu_{i,j}|I_m) = \frac{\mathcal{G}(w_{i,m}, \mu_{i,j}, \sigma_{i,j})}{\sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}, \sigma_{i,k})} \tag{3.38}$$

and the *concurrence probability* of μ_{i_1,j_1} and μ_{i_2,j_2} in I_m by

$$p(\mu_{i_1,j_1}, \mu_{i_2,j_2}|I_m) = p(\mu_{i_1,j_1}|I_m)p(\mu_{i_2,j_2}|I_m). \tag{3.39}$$

Now construct a graph $G = (V, E)$ with vertices $V = \{(i, j)\}$ and edges E given by

$$E = \big\{((i_1, j_1), (i_2, j_2)) : (R\boldsymbol{\xi}_{i_1} = \boldsymbol{\xi}_{i_2},\ R \in \mathcal{S}_R(\mathcal{L})) \Longrightarrow (i_1 \neq i_2 \text{ and } j_1 \neq j_2)\big\}, \tag{3.40}$$

i.e., only one value of j can be selected at each reciprocal lattice point $\boldsymbol{\xi}_i$ and each j can only be selected once among its twin coordinates $\boldsymbol{\xi}_t = R\boldsymbol{\xi}_i$, where $R \in \mathcal{S}_R(\mathcal{L})$. If there is any known symmetry in the structure factor magnitudes, such as Friedel or Laue symmetry, we can simplify the structure of G by merging the corresponding symmetric nodes. Consequently, choosing a consistent set of structure factor magnitudes, where each possible value, apart from symmetry, appears exactly once, is equivalent to finding a maximal clique in this graph, i.e., a set $C \subset V$ that satisfies for all $v_1, v_2 \in C$, $(v_1, v_2) \in E$ and has no proper superset $\tilde{C} \supsetneq C$ satisfying the same property.

We define a directed weight $W : E \to \mathbb{R}$ on G by summing the concurrence probabilities over the sets $\mathcal{I}_{i_1,i_2}$, consisting of all the images which potentially measure $\boldsymbol{\xi}_{i_1}$ and $\boldsymbol{\xi}_{i_2}$:

$$W((i_1, j_1), (i_2, j_2)) = \frac{\sum_{m \in \mathcal{I}_{i_1,i_2}} p(\mu_{i_1,j_1}, \mu_{i_2,j_2}|I_m)}{\sum_{m \in \mathcal{I}_{i_1,i_2}} p(\mu_{i_1,j_1}|I_m)}. \tag{3.41}$$

This weight describes the likelihood that the values of μ_{i_1,j_1} and μ_{i_2,j_2} occur together in one of the solutions. In fact, if for a fixed i_2 all of the μ_{i_2,j_2} are distinct, then, in the absence of

noise and errors, W is exactly 1 when μ_{i_1,j_1} and μ_{i_2,j_2} are simultaneously part of one of the detwinned solutions and is 0 otherwise. However, if for a fixed i_2 there are B values of k such that $\mu_{i_2,k} = \mu_{i_2,j_2}$, then $W(\mu_{i_1,j_1}, \mu_{i_2,j_2}) = \frac{1}{B}$. Furthermore, in the presence of noise, the asymmetry of the weight function serves to favor structure factor magnitudes which have a strong signal and are well modeled in the expectation maximization step, as they provide the greatest amount of orientation information.

We now formulate the solution to the twinning problem as follows. We seek the clique in G with maximal edge weight, i.e., we solve

$$\max_{C} \sum_{v_1,v_2 \in C} W(v_1, v_2), \text{ where for all } v_1, v_2 \in C, \ (v_1, v_2) \in E. \tag{3.42}$$

If a sufficient number of images are used then, in the absence of noise and error, the solution to (3.42) retrieves one of the exact solutions to the twinning problem, i.e., the maximum edge weight clique C assigns the correct structure factor magnitudes, up to a global rotation, see Figures 3.12 - 3.14. In fact, there will be several maximal cliques which maximize (3.42), each corresponding to a different valid solution. This remains true in the presence of noise and error if the variations in the data can be sufficiently controlled in the calculation of the multi-modal model parameters.

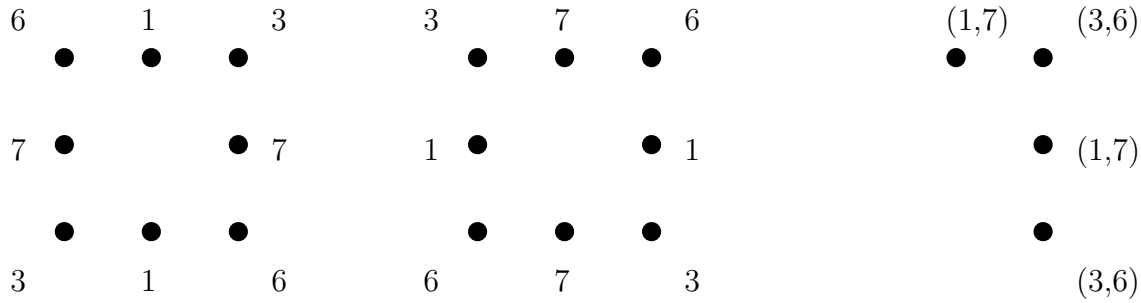


Figure 3.12: The left and middle images are two valid arrangements of the variance stabilized structure values $w_{i,m}$ on a square lattice, with the middle entry missing. Due to the twinning problem, autoindexing is unable to distinguish the orientation of the lattice when it is rotated by 90 degrees. Therefore, in this case, one will see two possible values, $\mu_{i,1}$ and $\mu_{i,2}$, for each structure factor magnitude when plotting the histograms of $w_{i,m}$. This is represented in the right image, where the dots represent the reciprocal lattice points $\boldsymbol{\xi}_i$, after merging Friedel pairs, and the number pairs are the associated possible values $(\mu_{i,1}, \mu_{i,2})$. In order to resolve the twinning problem, we must choose one of these numbers at each of the reciprocal lattice points and each number must be chosen at least once.

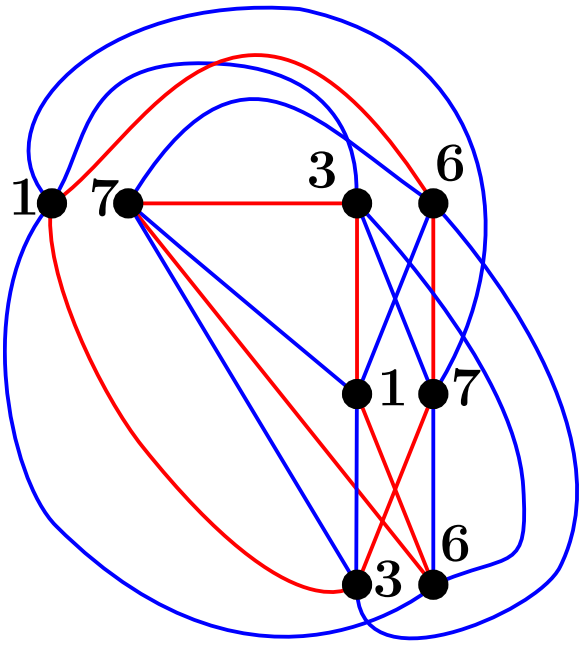


Figure 3.13: Weighted graph associated to the arrangement in Figure 3.12. Each reciprocal lattice point is split into 2 nodes, one for each of its possible values, and two nodes are connected if they represent a consistent choice of values. The blue edges correspond to large weight values and signifies high concurrency, which implies that the connected nodes likely correspond to the same orientation, and the red edges correspond to small weight values and signifies low concurrency, which implies that the connected nodes likely correspond to different orientations.

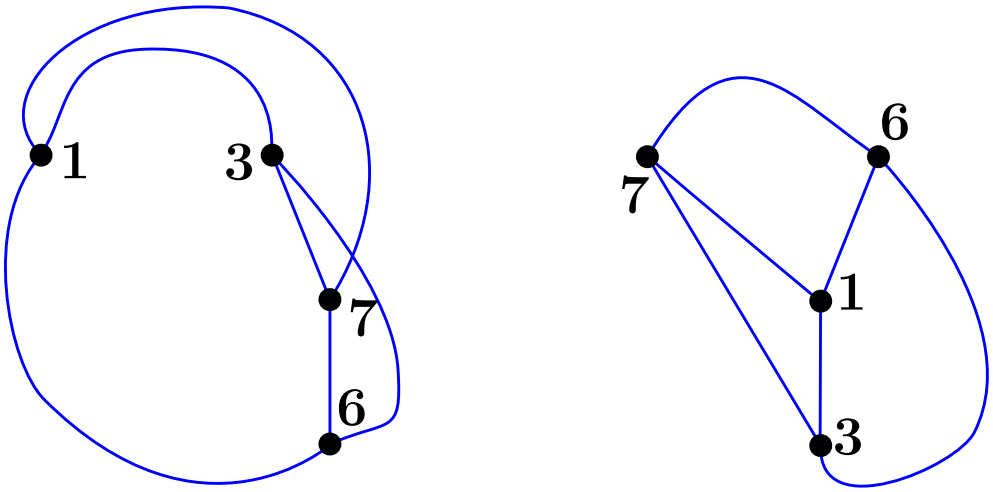


Figure 3.14: The two maximal cliques in Figure 3.13. These each correspond to one of the valid arrangements in Figure 3.12.

3.5.2 Greedy Approach to the Maximum Weight Clique Problem

In general, the maximum weight clique problem is NP-hard [3]. However, when this problem is constructed from the twinning problem via (3.42), we can solve it in quadratic time with a greedy approach. In particular, we accomplish this by starting at a node $v_s \in V$ and progressively add nodes which maximize the weight sum of the current clique. Also, in practice, we remove any nodes which, due to symmetry, are known to have less than the maximum amount of ambiguity, as they do not determine as much orientation information. For convenience, here we use a single index for the nodes $V = \{v_i\}_{i=1}^N$ and we set $W(v_1, v_2) = -\infty$ if $(v_1, v_2) \notin E$.

Algorithm 7
for $j = 1 : N$ **do**
$Y_j = 0$
end for
$C \leftarrow \{v_s\}$
$n \leftarrow s$
while C is not maximal **do**
for all $j \notin C$ **do**
$Y_j \leftarrow Y_j + W(v_n, v_j)$
end for
$n \leftarrow \arg\max_{j \notin C} Y_j$
$C \leftarrow C \cup \{v_n\}$
end while
return C

The elements of the set C, returned by Algorithm 7, are pairs of the form (i, j), which corresponds to choosing the j-th modeled variance stabilized structure factor magnitude $\mu_{i,j}$ at the reciprocal lattice point $\boldsymbol{\xi}_i$. In particular, this induces the map $\tilde{w} : \hat{\mathcal{L}} \to \mathbb{R}$ where $\tilde{w}(\boldsymbol{\xi}_i) = \mu_{i,j}$. In the absence of noise and error, if a sufficient number of images are collected then Algorithm 7 retrieves an exact solution to the twinning problem. In this case, the algorithm will always prefer a node v_n which has nonzero weighted edges connecting it with all of the elements of the current clique, i.e., v_n is only chosen if it is consistent with the current choice of structure factor magnitudes in the clique. Furthermore, when using imperfect data, this approach remains very robust as it takes into account all pairs of measured intensities over all of the images in order to choose the structure factor magnitudes at any single reciprocal lattice point.

3.5.3 Orientation Determination

Even though Algorithm 7 retrieves a good approximation of the detwinned structure factor magnitudes, its accuracy can be improved by first using this information to directly orient

each image, and then averaging the computed structure factor magnitudes for each of the corresponding reciprocal lattice points. More specifically, for every image I_m we compute its full orientation R_m by minimizing

$$\min_{R_m \in \mathcal{S}_R(\mathcal{L})} \frac{\sum_{i \in NZ} |w_{i,m} - \tilde{w}(R_m \boldsymbol{\xi}_i)|}{\sum_{i \in NZ} \tilde{w}(R_m \boldsymbol{\xi}_i)}, \text{ where } NZ = \{i : w_{i,m} > 0\}. \tag{3.43}$$

If there are at least two orientations close to the minimum value, then this indicates that the image might not have enough diverse information to be properly oriented and, thus, we reject the computed orientation. Once the orientations for each image are known, we can compute the structure factor magnitudes by averaging the data over the images. While there are many possible ways to compute this average, we found that it is best to average the numerator and denominator of Equation (3.30) over the images, with one level of variance stabilization, and with the scaling factors applied to the numerator:

$$|F(\boldsymbol{\xi}_i)| = \frac{\sum_m c_m^2 \sqrt{\sum_{R_m q(\mathbf{x}) \in B(\boldsymbol{\xi}_i, r)} I_m(\mathbf{x})}}{\sum_m \sqrt{\sum_{R_m q(\mathbf{x}) \in B(\boldsymbol{\xi}_i, r)} r_e^2 P(\mathbf{x}) \Delta\Omega(\mathbf{x}) |S_{\mathbf{N}_m}(R_m q(\mathbf{x}))|^2}}, \tag{3.44}$$

where we throw away terms in the sum that correspond to images that do not pass within a distance of r from $\boldsymbol{\xi}_i$.

3.5.4 Summary

We now summarize our approach to solving the twinning problem. We note that, if desired, this approach can be extended to utilize non-reciprocal lattice points as well, if their corresponding magnitudes are also modeled in the expectation maximization and scaling steps. Given the variance stabilized structure factor magnitude calculations $\{w_{i,m}\}$, the multi-modal model parameters $\{\mu_{i,j}\}$ and $\{\sigma_{i,j}\}$, and the scaling corrections c_m, which are all described in Section 3.4.2, we perform the following.

Algorithm 8

1. *Compute occurrence and concurrence probabilities*

$$p(\mu_{i,j}|I_m) = \frac{\mathcal{G}(w_{i,m}, \mu_{i,j}, \sigma_{i,j})}{\sum_{k=1}^{K} \mathcal{G}(w_{i,m}, \mu_{i,k}, \sigma_{i,k})}, \quad p(\mu_{i_1,j_1}, \mu_{i_2,j_2}|I_m) = p(\mu_{i_1,j_1}|I_m) p(\mu_{i_2,j_2}|I_m).$$

2. *Construct the graphical model G = (V,E) via*

$$V = \{(i,j)\},$$
$$E = \{((i_1,j_1),(i_2,j_2)) : (R\boldsymbol{\xi}_{i_1} = \boldsymbol{\xi}_{i_2},\ R \in \mathcal{S}_R(\mathcal{L})) \Longrightarrow i_1 \neq i_2 \text{ and } j_1 \neq j_2\}.$$

3. *Merge any nodes which are known a priori to have the same structure factor magnitudes.*

4. *Compute the directed weights*

$$W((i_1,j_1),(i_2,j_2)) = \frac{\sum_{m\in\mathcal{I}_{i_1,i_2}} p(\mu_{i_1,j_1},\mu_{i_2,j_2}|I_m)}{\sum_{m\in\mathcal{I}_{i_1,i_2}} p(\mu_{i_1,j_1}|I_m)}.$$

5. *Run Algorithm 7 on (G,W) to determine the placement of the approximate variance stabilized structure factor magnitudes* $\tilde{w}:\hat{\mathcal{L}}\to\mathbb{R}$.

6. *Compute the orientations* R_m *for each image* I_m *by solving:*

$$\min_{R_m\in\mathcal{S}_R(\mathcal{L})} \frac{\sum_{i\in NZ}|w_{i,m}-\tilde{w}(R_m\boldsymbol{\xi}_i)|}{\sum_{i\in NZ}\tilde{w}(R_m\boldsymbol{\xi}_i)},\ \textit{where } NZ=\{i:w_{i,m}>0\}.$$

7. *Compute the detwinned structure factor magnitudes*

$$|F(\boldsymbol{\xi}_i)| = \frac{\sum_m c_m^2\sqrt{\sum_{R_m q(\mathbf{x})\in B(\boldsymbol{\xi}_i,r)} I_m(\mathbf{x})}}{\sum_m\sqrt{\sum_{R_m q(\mathbf{x})\in B(\boldsymbol{\xi}_i,r)} r_e^2 P(\mathbf{x})\Delta\Omega(\mathbf{x})|S_{\mathbf{N}_m}(R_m q(\mathbf{x}))|^2}}.$$

3.6 Computational Phase Retrieval for X-ray Nanocrystallography

Once the detwinned structure factor magnitudes are computed, one can then proceed to find the missing phases and, thus, determine the electron density of the molecules in a unit cell, with any of the phasing techniques presented in Section 2.5.1. In particular, the methods of the previous section can be used to enhance molecular replacement techniques, which no longer have to work with a reduced data set, and, furthermore, allow for the use of techniques involving anomalous dispersion and isomorphous replacement, which typically require detwinned data in order to function effectively.

While these classical phasing techniques have been used extensively to reconstruct molecular structure in x-ray crystallography, they each have limitations or introduce extra difficulties into the experimental setup. For instance, molecular replacement requires one to already know a structure similar to the sample, which may not available for fundamentally new objects. Anomalous dispersion requires the presence of a sufficient number of anomalous scatterers and restricts one to use x-ray wavelengths which are near the absorption edge of these scatterers, which may not be possible to do at the desired brightness for nanocrystallography. Isomorphous replacement requires one to create an isomorphic crystal with the inclusion of heavy atoms, which may be particularly difficult to achieve for the types of samples one wishes to study with nanocrystallography, i.e., samples which are already difficult to crystallize on their own.

Alternatively, one can, in principle, use computational phase retrieval techniques to compute phase information from the Fourier magnitudes alone, if they are sampled at a sufficiently high rate, e.g., at least twice the Nyquist rate of the unit cell's electron density. While this approach has been infeasible for most conventional crystallography experiments, as they only sample the Fourier magnitudes directly at this Nyquist rate, the signal from nanocrystals contain a significant amount of information between Bragg peaks, and may allow sampling at the rate required for computational phase retrieval.

3.6.1 Sampling Strategies

We now discuss how to make computational phase retrieval viable for x-ray nanocrystallography by sampling non-Bragg data. In particular, the main obstacle in making computational phase retrieval feasible is its required sampling density. Recall from Section 2.5.3 that the well-posedness of the phase retrieval problem is largely based on being able to retrieve the autocorrelation $A\rho$ of the solution ρ from its power spectrum $|\hat{\rho}|^2$. By the Shannon-Nyquist theorem, this is possible if one samples the power spectrum at twice the Nyquist rate of ρ, since the support of $A\rho$ is twice as large as the support of ρ. Unfortunately, the strongest signal in the images occurs at Bragg peaks, which sample the power spectrum directly at the Nyquist rate of ρ. Therefore, in order to make computational phase retrieval feasible in diffraction images from crystals, one must also sample non-Bragg data, which have a considerably smaller amount of signal. Specifically, in order to satisfy the Nyquist rate for a crystal lattice with reciprocal Bravais characteristic vectors $(\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \hat{\mathbf{h}}_3)$, one must sample the power spectrum at points of the form

$$\boldsymbol{\xi} = \sum_{i=1}^{3} \frac{n_i}{2} \hat{\mathbf{h}}_i, \text{ where } n_i \in \mathbb{Z}. \tag{3.45}$$

If the crystal sizes in each Bravais direction are given by (N_1, N_2, N_3), then the shape function applied to these sample points will have the following orders of magnitude:

$$\left| S\left(\sum_{i=1}^{3} n_i \hat{\mathbf{h}}_i \right) \right|^2 = \mathcal{O}(N_1^2 N_2^2 N_3^2), \tag{3.46}$$

$$\left| S\left(\sum_{i=1}^{2} n_{\sigma(i)} \hat{\mathbf{h}}_{\sigma(i)} + \left(n_{\sigma(3)} + \tfrac{1}{2}\right) \hat{\mathbf{h}}_{\sigma(3)} \right) \right|^2 = \mathcal{O}(N_{\sigma(1)}^2 N_{\sigma(2)}^2), \tag{3.47}$$

$$\left| S\left(n_{\sigma(1)} \hat{\mathbf{h}}_{\sigma(1)} + \sum_{i=2}^{3} \left(n_{\sigma(i)} + \tfrac{1}{2}\right) \hat{\mathbf{h}}_{\sigma(i)} \right) \right|^2 = \mathcal{O}(N_{\sigma(1)}^2), \tag{3.48}$$

$$\left| S\left(\sum_{i=1}^{3} \left(n_i + \tfrac{1}{2}\right) \hat{\mathbf{h}}_i \right) \right|^2 = \mathcal{O}(1), \tag{3.49}$$

where σ is any permutation of the indices. Note that the amount of signal collected in (3.48) and (3.49) is several orders of magnitude smaller than what is seen at a Bragg peak.

We note that in order to compute the Fourier magnitudes at non-reciprocal lattice points, needed in (3.45), one must obtain the orientations up to their associated symmetry, which may be less than that of the Bragg peaks. For example, for a crystal with a space group which contains screw axis operations, the Laue symmetry tends to be greater than the symmetry of the non-Bragg points. In such cases, one may need to perform the detwinning techniques discussed in Section 3.5 on the non-Bragg peaks, in addition to the Bragg peaks.

The feasibility of using the sampling points in (3.45) to sample with at least twice the Nyquist rate was studied in [67], assuming knowledge of the orientations, small variations in the crystal sizes, and constant incident photon flux densities. However, they found that they required at least 10^6 images when using a beam with 10^{13} photons/pulse focused to a full-width at half-maximum of 0.5 µm, which corresponds to a total collected signal which is four orders of magnitude larger than currently , e.g., compared to what is collected in [16].

Alternatively, we would like to only consider sample points that satisfy (3.46) or (3.47), depicted in Figure 3.15, which have a signal several orders of magnitude larger than those in (3.48) and (3.49). While such a sampling no longer satisfies the hypothesis of the Shannon-Nyquist theorem for retrieving the autocorrelation, it does have a sampling density which is four times the Nyquist density for ρ. Note that this is the exact sampling density required to solve the phase retrieval problem in Theorem 8. Even though this result was proven for a different sampling strategy, it may still be valid in our case, especially if the molecular arrangement in the unit cell has sufficiently small support. By using this sampling strategy, we can potentially perform computational phase retrieval with far fewer images and beam power than when sampling directly at the Nyquist rate.

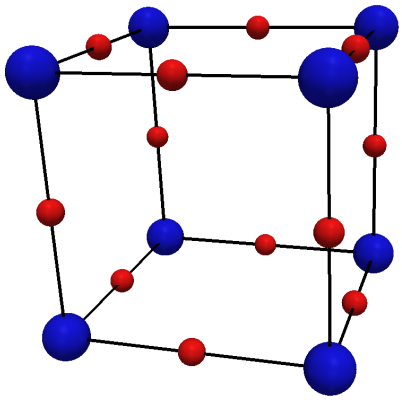


Figure 3.15: Sampling strategy on a reciprocal unit cell: We sample at reciprocal lattice points (blue) and halfway between two reciprocal lattice points (red).

3.6.2 Compressive Phase Retrieval

Recall that for Fourier magnitude values $a : \mathbb{Z}_{\mathbf{N}} \to \mathbb{C}$ and some support $T \subseteq \mathbb{Z}_{\mathbf{N}}$, phase retrieval algorithms seek a function $\rho \in M \cap S$, where $M = \{y \in \ell^2(\mathbb{Z}_{\mathbf{N}}) : |\hat{y}| = a\}$ and $S = \{y \in \ell^2(\mathbb{Z}_{\mathbf{N}}) : \text{supp}(y) \in T\}$. These algorithms typically make use of the projectors P_M and P_S onto M and S, respectively, via (2.55) and (2.56). In order to accurately and efficiently evaluate the Fourier transform in the definition of P_M via the Fast Fourier Transform, one requires information on a Cartesian grid. However, our sampling strategy in Section 3.6.1 does not record values for a at all locations on this grid. Additionally, information at the zero frequency is missing in experiments, as a hole in the detector is placed here to allow the x-ray beam to pass through. Furthermore, sample points with small structure factors may not have enough recorded signal to be used in the phase retrieval process. Therefore, we need to augment the magnitude projection operator to take this lack of information into account. In particular, if Ω is the set of points in reciprocal space where a has a recorded value, then we use the augmented projection operator $P_{M,\Omega}$, defined by

$$\begin{aligned} \tilde{P}_{M,\Omega}\hat{\rho}(k) &= \begin{cases} a(k)\frac{\hat{\rho}(k)}{|\hat{\rho}(k)|}, & \text{if } \hat{\rho} \neq 0 \text{ and } k \in \Omega, \\ a(k), & \text{if } \hat{\rho} = 0 \text{ and } k \in \Omega, \\ \hat{\rho}(k), & \text{if } k \notin \Omega, \end{cases} \\ P_{M\Omega}\rho &= \mathcal{F}^* \tilde{P}_{M,\Omega} \mathcal{F}. \end{aligned} \tag{3.50}$$

Since $P_{M,\Omega}$ does not alter information outside of Ω, in a phase retrieval algorithm such points serve as extra degrees of freedom in minimizing the support error $\varepsilon_S(\rho)$. In particular, when using $P_{M,\Omega}$, the ER algorithm, (2.58), can be expressed as

$$\rho^{(n+1)} = P_S P_{M,\Omega} \rho^{(n)}, \tag{3.51}$$

and the HIO algorithm, (2.60), can be expressed as

$$\rho^{(n+1)} = \begin{cases} P_{M,\Omega}\rho^{(n)}(x), & \text{if } x \in T, \\ \rho^{(n)}(x) - \beta P_{M,\Omega}\rho^{(n)}(x), & \text{if } x \notin T. \end{cases} \tag{3.52}$$

Recall, from Section 2.5.3, that the ER and HIO algorithms typically require a tight estimate for the support of the solution. Therefore, we utilize the shrinkwrap method, Algorithm 1, combined with the modified ER and HIO algorithms, (3.51) and (3.52), which we alternate after several iterations, to iteratively seek out the unknown solution support and retrieve the phase information. Here, we initialize the support to be the unit cell. By seeking the solution with smallest support that is consistent with the data, this method has the potential to further reduce the sampling requirement for computational phase retrieval, beyond the sampling strategy presented in Section 3.6.1.

3.6.3 Summary

Here we summarize our sampling and phase retrieval approach. We note that one can extend this approach by also using sample points which lie anywhere on the line between two reciprocal lattice points, instead of just the midpoint.

Algorithm 9

1. *Create the set of sample points,*

$$\Omega = \left\{ \sum_{i=1}^{3} n_i \hat{\mathbf{h}}_i : n_i \in \mathbb{Z} \right\} \bigcup \left\{ \sum_{i=1}^{2} n_{\sigma(i)} \hat{\mathbf{h}}_{\sigma(i)} + \left(n_{\sigma(3)} + \tfrac{1}{2} \right) \hat{\mathbf{h}}_{\sigma(3)} : n_i \in \mathbb{Z} \right\},$$

   where σ is any permutation of three elements.

2. *Compute the Fourier magnitude values $a(\boldsymbol{\xi}) = |F(\boldsymbol{\xi})|$ for all $\boldsymbol{\xi} \in \Omega$.*

3. *Remove any elements $\boldsymbol{\xi} \in \Omega$ where $a(\boldsymbol{\xi})$ could not be computed.*

4. *Initialize the initial support T to be the size of the unit cell.*

5. *Retrieve the electron density ρ by using Algorithm 1 with the modified ER and HIO algorithms, which use $P_{M,\Omega}$ in place of P_M. In particular, alternate between several iterations of HIO and ER.*

Chapter 4

Results

4.1 Overview

Here we demonstrate our x-ray nanocrystallography reconstruction methodology on realistic simulated diffraction data for three different crystal structures, described in Section 4.2, and vary the peak incident photon flux density in each case, which can, for example, be achieved by widening or narrowing the beam. Each data set consists of 33,856 diffraction images. Here we assume knowledge of the Bravais vector lengths and the space groups, which, in practice, may be deduced from autoindexing information and reflection conditions [30]. We compute the structure factors of the unit cells with the atomic positions of real molecules from the Protein Data Bank [9]. In particular, the intensity of every pixel is computed via Equation (2.33), by using the Cromer-Mann coefficients for each atom listed in [75] and tabulated in [1], along with dispersion factors listed in [37].

The orientation of each image is generated from a random distribution of unit quaternions (w, x, y, z) where the components are sampled from a normal distribution and then normalized, so that $w^2 + x^2 + y^2 + z^2 = 1$. The associated orientation matrices R are calculated from the unit quaternions via

$$R = \begin{pmatrix} 1 - 2y^2 - 2z^2 & 2xy - 2zw & 2xz + 2yw \\ 2xy + 2zw & 1 - 2x^2 - 2z^2 & 2yz - 2xw \\ 2xz - 2yw & 2yz + 2xw & 1 - 2x^2 - 2y^2 \end{pmatrix}. \tag{4.1}$$

The crystal sizes (N_1, N_2, N_3) associated to the Bravais characteristic vectors, with lengths (L_1, L_2, L_3), are generated by first sampling random average crystal widths W from a normal distribution and then randomizing the size along each dimension:

$$W \sim \mathcal{N}(\mu_C, \sigma_C), \quad W_j \sim \mathcal{N}\left(W, \frac{W}{10}\right), \quad N_j = \left\lfloor \frac{W_j}{L_j} \right\rfloor, \tag{4.2}$$

with different μ_C and σ_C chosen for each test case. For each image, we generate random incident photon flux densities J, measured in photons per square Angstrom per pulse, from

a peak density J_o, via

$$x \sim U(-1,1), \quad J = J_o e^{-\frac{x^2}{2(.25)^2}}. \tag{4.3}$$

This corresponds to interactions which occur in a region whose width is up to 3.4 times the full width at half maximum of the beam, resulting in incident photon flux densities as low as about $\frac{J_o}{3000}$. The image values $I(x,y)$ are then computed via Equation (2.62) along with shot and background noise. More specifically, we add Poisson noise and additive Gaussian noise, with a standard deviation of 1.3 photons, which is similar to the noise levels observed in [16]. We also use experimental parameters similar to [16]: 6.9 Å photon wavelength, $75\times75\,\mu m^2$ pixel size, a front detector with 1024×1024 pixels placed at a distance of 68 mm from the interaction point, and incident photon flux densities which vary around 218.4 $Å^{-2}$. For each of our test cases, the rear detector, with 1024×1024 pixels, was placed so that it could record at least 3 Bragg peaks along any given Bravais direction. We remove a small region of the detectors around the center, with size a fourth of the smallest reciprocal Bravais characteristic vector length, in order to allow the incoming beam to pass through. For convenience, we replicate the rear detector data on the front detector images at the appropriate resolution.

In Sections 4.3 - 4.7, we present results for each of the main steps in our algorithmic framework, where the output of each step is used to initialize the next. Our main runs, which each consist of autoindexing, described in Section 3.2; crystal size determination, described in Section 3.3; structure factor magnitude modeling, described in Section 3.4; and solving the twinning problem, described in Section 3.5, were performed on the Hopper Cray XE6 supercomputer, where each compute node consists of two twelve-core AMD 'MagnyCours' 2.1-GHz processors. Each run took between 15-30 minutes when using 529 cores. In general, this procedure scales linearly in the number of images and quadratically in the number of reciprocal points used to solve the twinning problem. The main bottleneck is the amount of memory required, which grows quadratically in the number of reciprocal points. The phase retrieval step was performed in serial on a Dell Optiplex 755 with an Intel Core 2 Duo 3-GHz processor and typically took between 30-180 minutes to converge, depending on the number of Bragg peaks that are measured.

4.2 Description of Test Cases

4.2.1 Test Case 1: PuuE Allantoinase

In our first test case, we determine the structure of PuuE Allantoinase using the atomic coordinates and crystal symmetry recorded in [61]. The associated crystal displays P4 space group symmetry. More specifically, we place the unit cell at the vertices of a tetragonal crystal lattice, with Bravais characteristic directions $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ with associated lengths 98.299 Å, 98.299 Å, and 62.393 Å, and apply the unit cell symmetry operators

(x, y, z), $(-x, -y, z)$, $(-y, x, z)$, and $(y, -x, z)$ to the atomic positions in a reference configuration. Consequently, the diffraction pattern is symmetric with respect to 90 degree rotation about the z-axis and inversion, due to Friedel symmetry. This leaves a two-fold twinning problem corresponding to 180 degree rotation about the x-axis, which, due to the symmetry, is equivalent to 180 degree rotation about the y-axis. In this case, the Laue symmetry of the Bragg reflections matches the symmetry of the non-Bragg reflections, and, thus, the orientations that detwin the Bragg data also detwin the non-Bragg data.

Here we place the rear detector at a distance of 141 mm from the interaction point. The crystal sizes were generated from (4.2) with $\mu_C = 2948.97$ Å and $\sigma_C = 982.99$ Å. We tested peak incident photon flux densities J_o of 2.18, 4.36, 10.9, 21.8, 43.6, 218, 1009, 2180, 4360, 10900, and 21800 Å^{-2}.

In Figures 4.1 - 4.5 we present typical diffraction images, colored by the logarithm of the intensity, from the front and rear detector for test case 1. Note that we replicate the rear detector data in the front detector image.

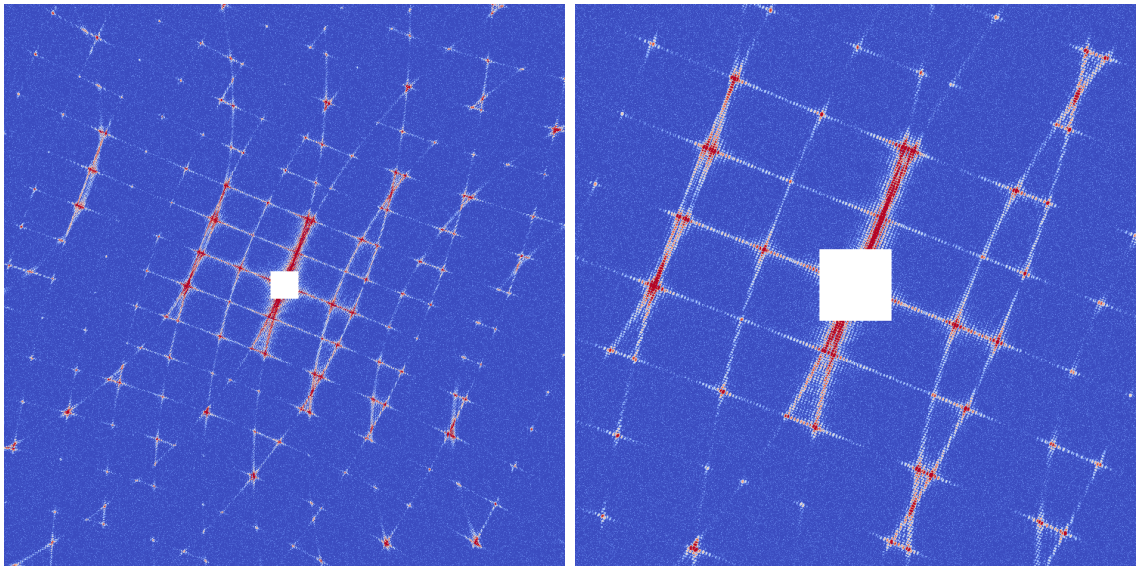


Figure 4.1: Simulated diffraction images for test case 1 with $J_o = 21800$. Left: Front detector. Right: Rear detector.

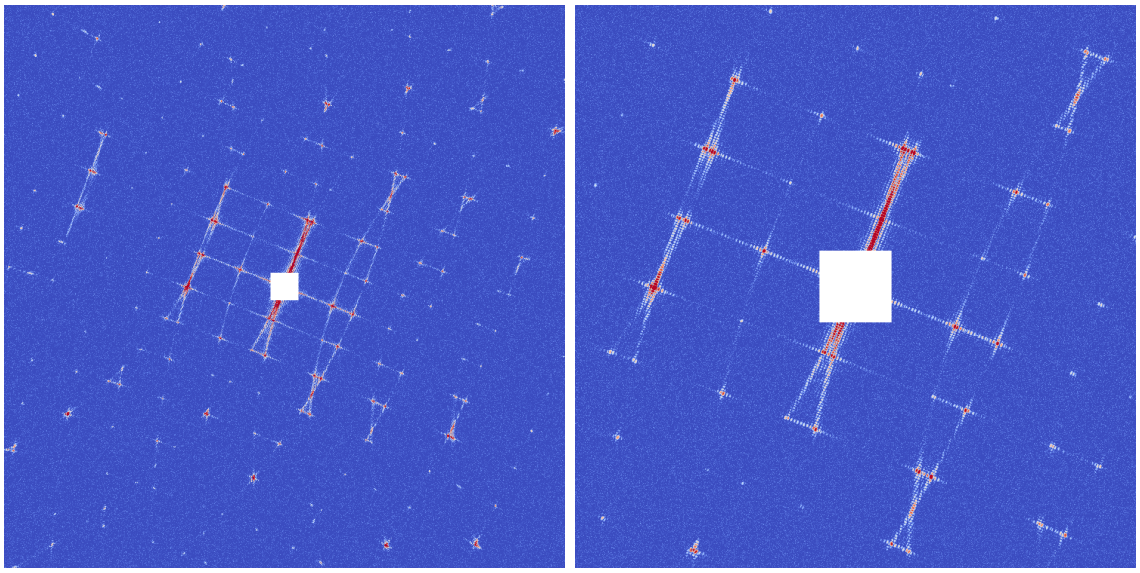


Figure 4.2: Simulated diffraction images for test case 1 with $J_o = 2180$. Left: Front detector. Right: Rear detector.

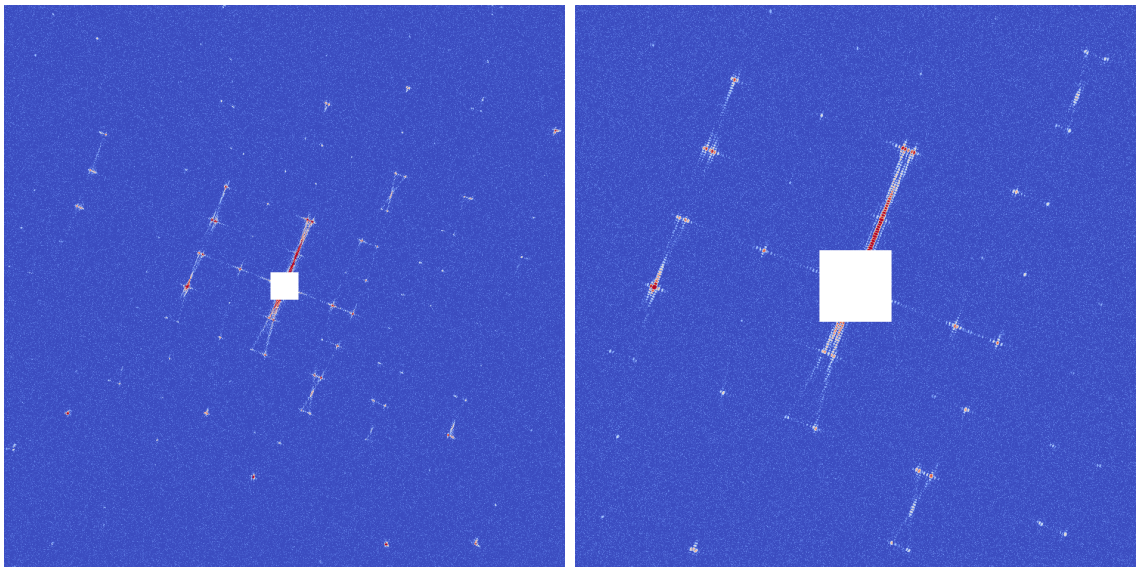


Figure 4.3: Simulated diffraction images for test case 1 with $J_o = 218$. Left: Front detector. Right: Rear detector.

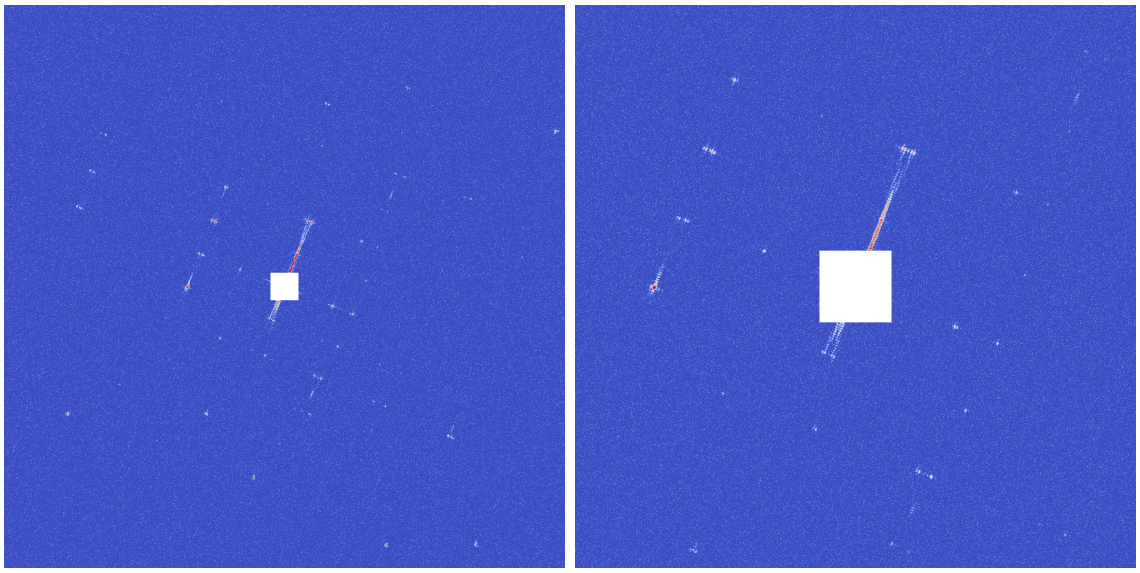


Figure 4.4: Simulated diffraction images for test case 1 with $J_o = 21.8$. Left: Front detector. Right: Rear detector.

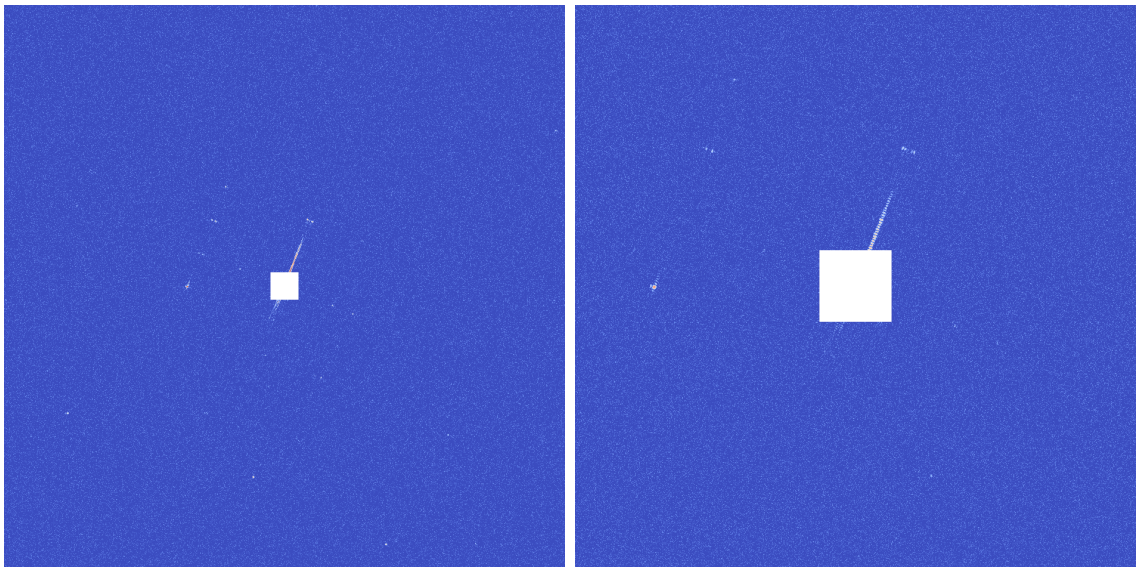


Figure 4.5: Simulated diffraction images for test case 1 with $J_o = 2.18$. Left: Front detector. Right: Rear detector.

4.2.2 Test Case 2: Photosystem II from Synechococcus Elongatus Without Unit Cell Symmetry

In our second test case, we determine the structure of Photosystem II from Synechococcus Elongatus using the atomic coordinates recorded in [77]. We place the molecule at the vertices in a orthorhombic crystal lattice, with Bravais characteristic directions $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ with associated lengths 130.01 Å, 226.72 Å, and 308.29 Å. In order to induce a four-fold twinning problem, we do not apply any unit cell symmetry operations, which results in a twinning problem corresponding to 180 degree rotations about the x, y, and z axes. Consequently, the measured signal is reduced by a factor of 4 and the associated phase retrieval problem simplifies a bit since the molecule has a smaller support within the unit cell. However, the main point of this example is test the robustness of our algorithmic framework for solving the twinning problem in the face of four-fold twinning for structure sizes similar to what is currently being studied in nanocrystallography experiments. In this example, the orientations that detwin the Bragg data also detwin the non-Bragg data.

Here we place the rear detector at a distance of 564 mm from the interaction point. The crystal sizes were generated from (4.2) with $\mu_C = 2600.2$ Å and $\sigma_C = 910.07$ Å. We tested peak incident photon flux densities J_o of 2.18, 4.36, 10.9, 21.8, 43.6, 218, 1009, 2180, 4360, 10900, and 21800 Å^{-2}.

In Figures 4.6 - 4.10 we present typical diffraction images, colored by the logarithm of the intensity, from the front and rear detector for test case 2. Note that we replicate the rear detector data in the front detector image.

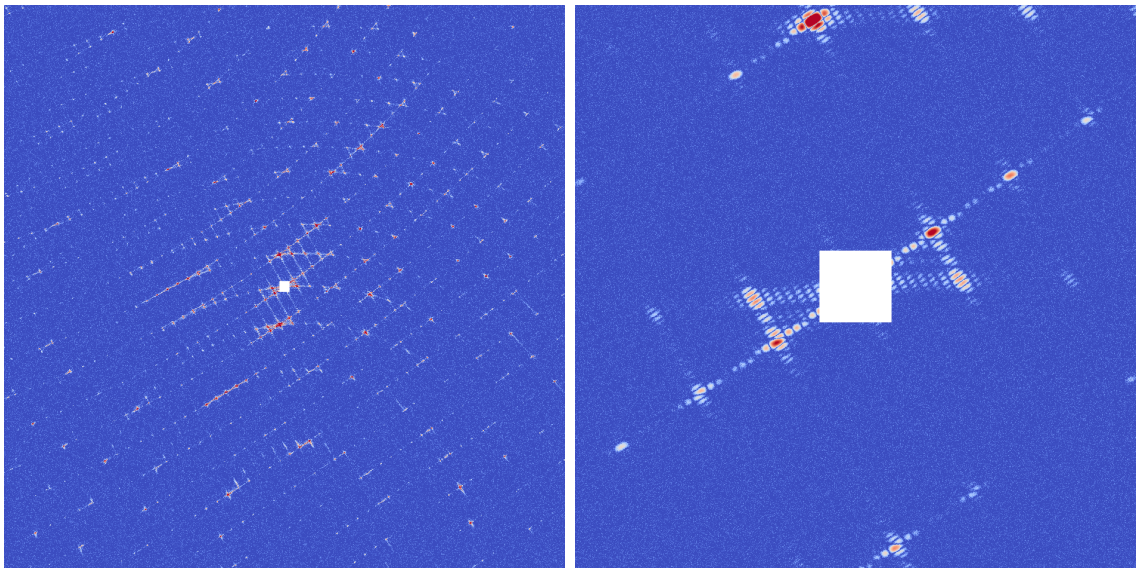


Figure 4.6: Simulated diffraction images for test case 2 with $J_o = 21800$. Left: Front detector. Right: Rear detector.

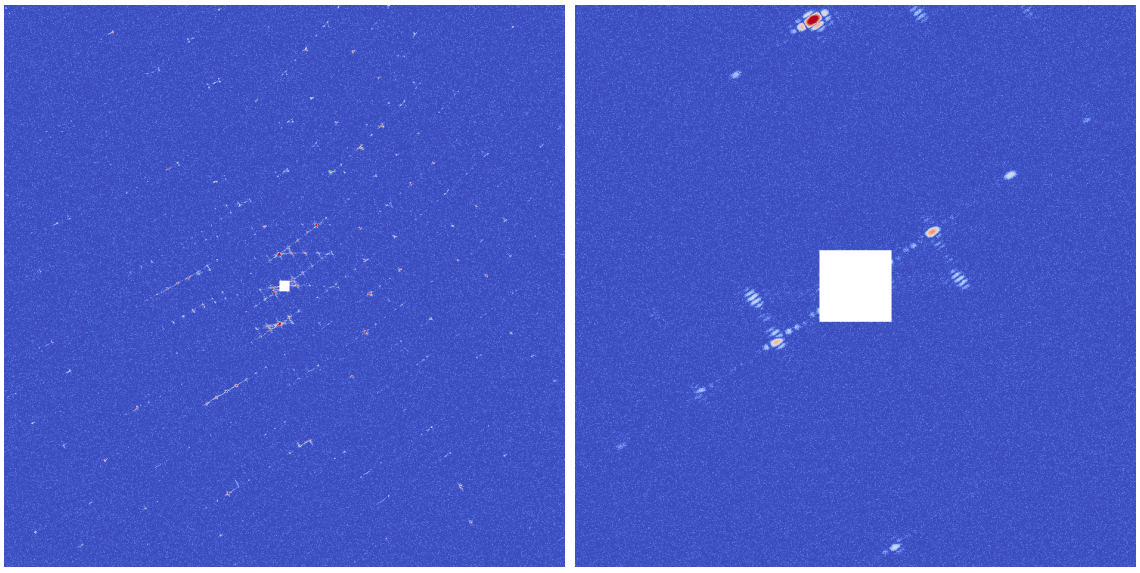


Figure 4.7: Simulated diffraction images for test case 2 with $J_o = 2180$. Left: Front detector. Right: Rear detector.

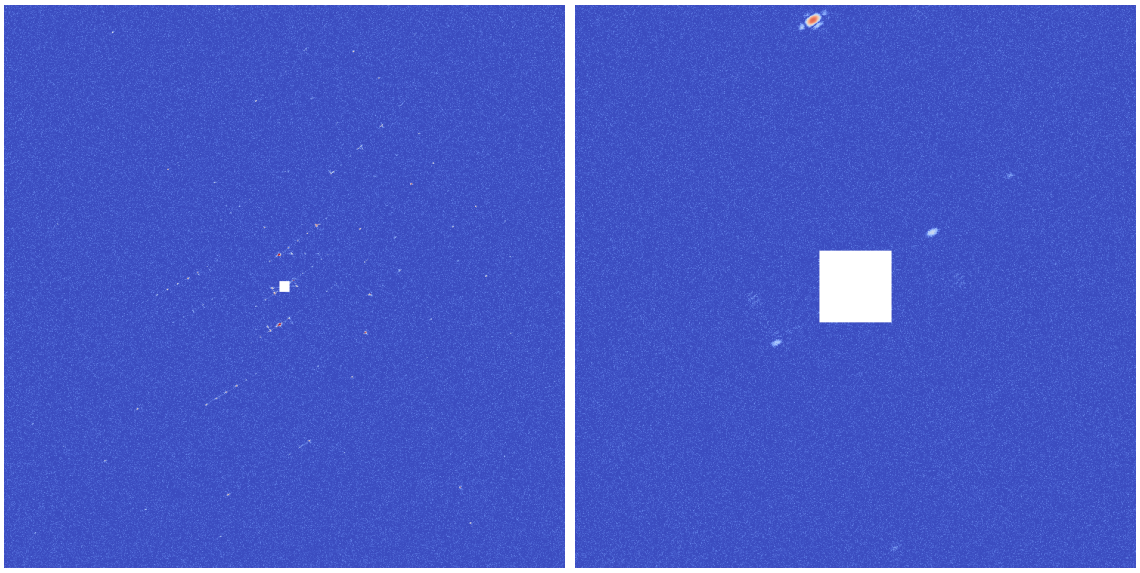


Figure 4.8: Simulated diffraction images for test case 2 with $J_o = 218$. Left: Front detector. Right: Rear detector.

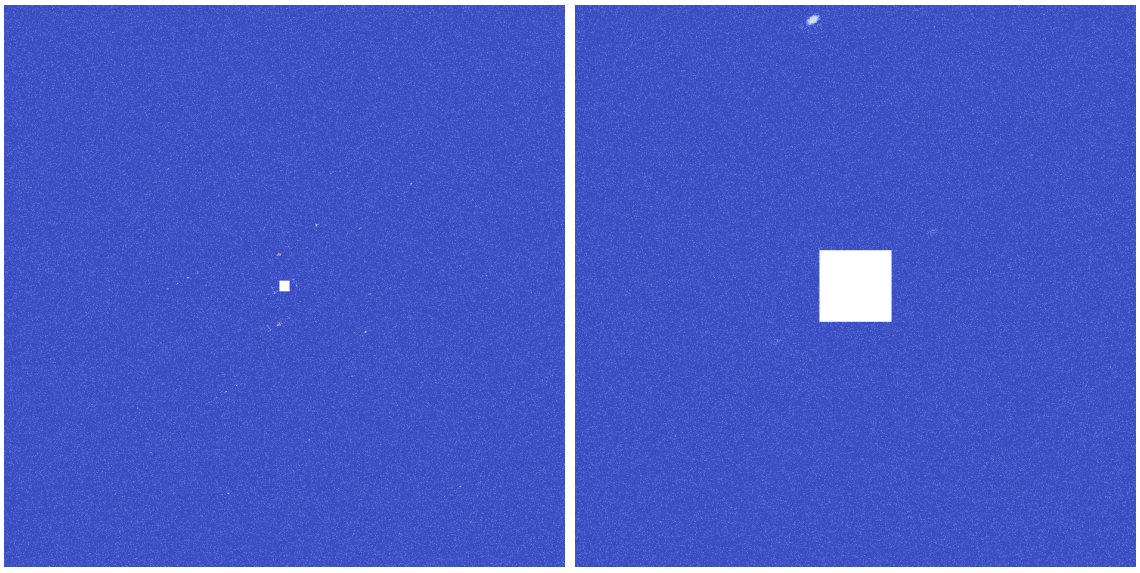


Figure 4.9: Simulated diffraction images for test case 2 with $J_o = 21.8$. Left: Front detector. Right: Rear detector.

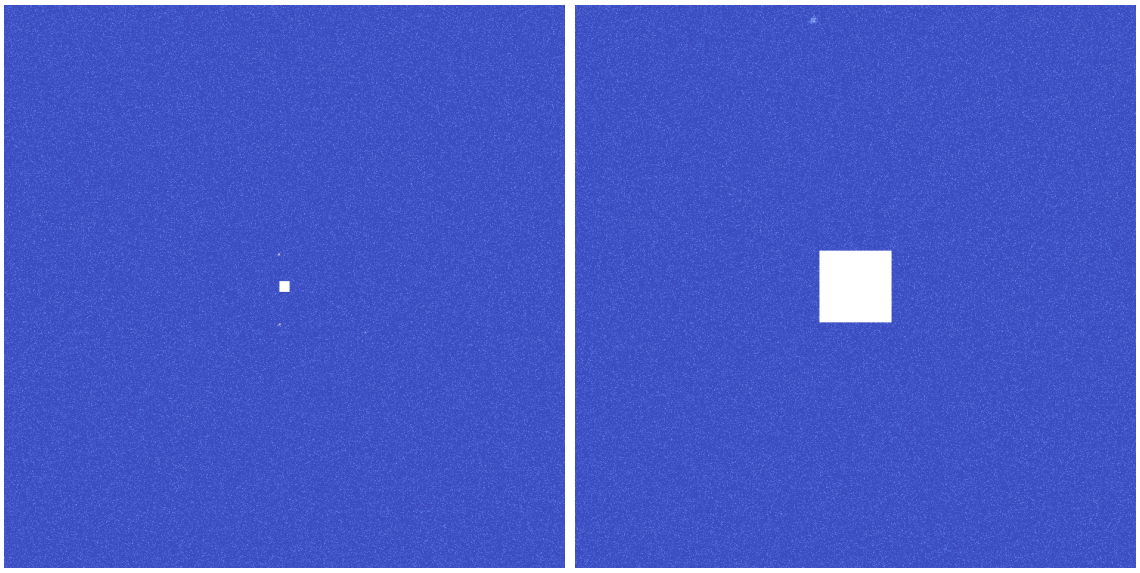


Figure 4.10: Simulated diffraction images for test case 2 with $J_o = 2.18$. Left: Front detector. Right: Rear detector.

4.2.3 Test Case 3: Photosystem II from Synechococcus Elongatus With Unit Cell Symmetry - Detwinning Non-Bragg Data

In our third test case, we determine the structure of Photosystem II from Synechococcus Elongatus with the unit cell symmetry observed in [77]. In this case, the associated crystal displays $P2_12_12_1$ space group symmetry. More specifically, we place the molecule at the vertices in a orthorhombic crystal lattice, with Bravais characteristic directions $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ with associated lengths 130.01 Å, 226.72 Å, and 308.29 Å, and apply the unit cell symmetry operators (x,y,z), $(-x+\frac{L_1}{2},-y,z+\frac{L_3}{2})$, $(-x,y+\frac{L_2}{2},-z+\frac{L_3}{2})$, and $(x+\frac{L_1}{2},-y+\frac{L_2}{2},-z)$ to the atomic positions in a reference configuration. While the Laue symmetry of the associated diffraction pattern is the same as the lattice symmetry, the non-Bragg data has less symmetry, i.e., the Bragg data does not display twinning but the non-Bragg data does. However, in order to perform computational phase retrieval, we require detwinned non-Bragg data at the midpoint between adjacent reciprocal lattice points, given by Equation (3.47). Consequently, we have a four-fold twinning problem when considering these non-Bragg points, corresponding to 180 degree rotations about the x, y, and z axes. However, the structure factor magnitudes which we are considering each only have two possible values at each of the non-Bragg points. In particular, lattice points with Miller indices of the form $(n_1+\frac{1}{2},n_2,n_3)$ have magnitudes which are symmetric with respect to 180 degree rotation about the z-axis, lattice points with Miller indices of the form $(n_1,n_2+\frac{1}{2},n_3)$ have magnitudes which are symmetric with respect to 180 degree rotation about the x-axis,

and lattice points with Miller indices of the form $(n_1, n_2, n_3 + \frac{1}{2})$ have magnitudes which are symmetric with respect to 180 degree rotation about the y-axis, where $n_1, n_2, n_3 \in \mathbb{Z}$. In order to record sufficient information from the weaker non-Bragg reflections, this example requires a larger amount of signal, but this can potentially be obtained by using larger crystals or decreasing the beam width in experiments, albeit at the cost of possibly missing more crystals with the beam.

Here we place the rear detector at a distance of 564 mm from the interaction point. The crystal sizes were generated from (4.2) with $\mu_C = 3900.3$ Å and $\sigma_C = 1040.08$ Å. We tested peak incident photon flux densities J_o of 218, 1009, 2180, 4360, 10900, and 21800 Å^{-2}.

In Figures 4.11 - 4.13 we present typical diffraction images, colored by the logarithm of the intensity, from the front and rear detector for test case 3. Note that we replicate the rear detector data in the front detector image.

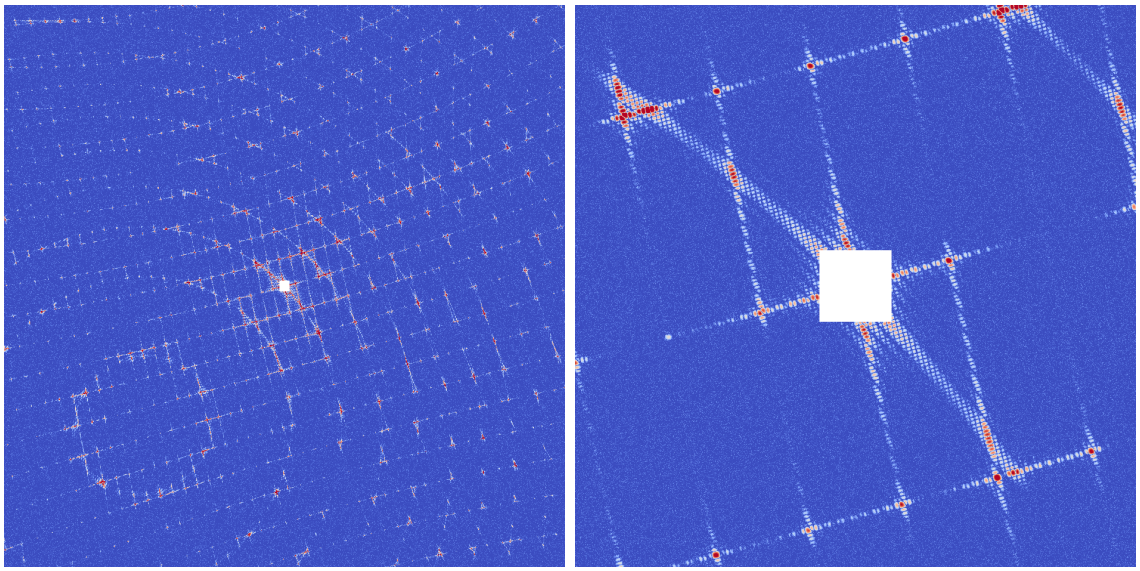


Figure 4.11: Simulated diffraction images for test case 3 with $J_o = 21800$. Left: Front detector. Right: Rear detector.

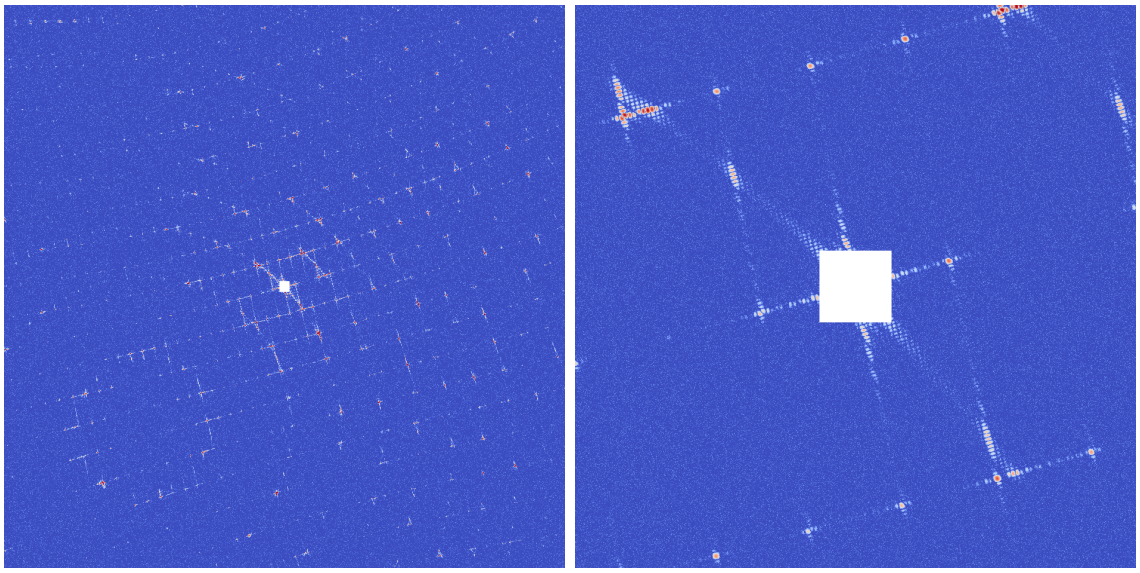


Figure 4.12: Simulated diffraction images for test case 3 with $J_o = 2180$. Left: Front detector. Right: Rear detector.

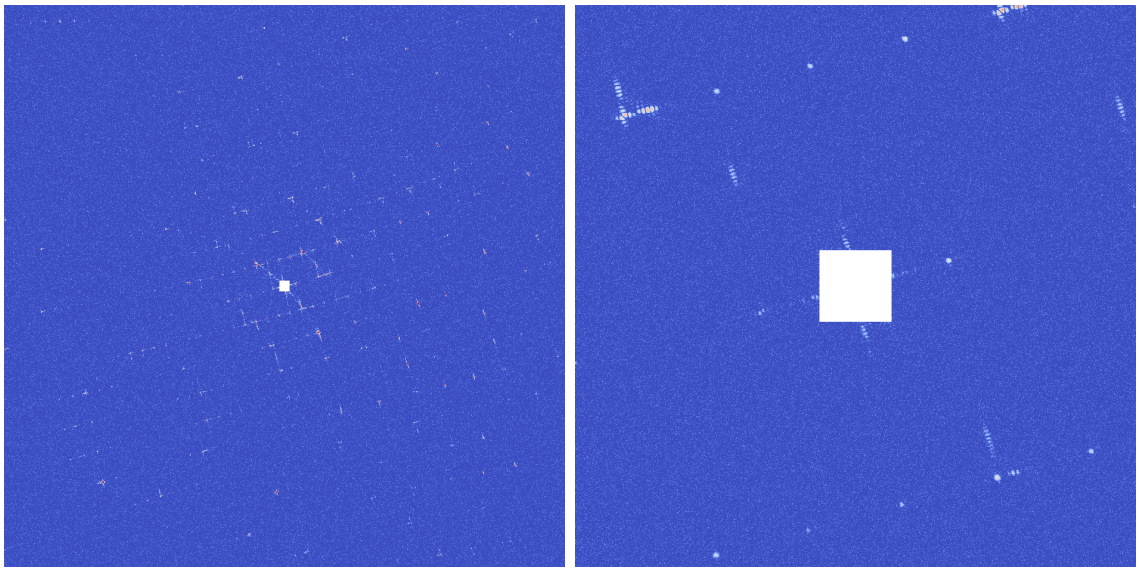


Figure 4.13: Simulated diffraction images for test case 3 with $J_o = 218$. Left: Front detector. Right: Rear detector.

4.3 Autoindexing

4.3.1 Test Description

Here we test the performance of our autoindexing strategy from Section 3.2 as we vary the incident photon flux density for each test case. In these tests, for every image, we consider the largest 100 peaks which are maxima in a neighborhood of 5 pixels and have a measured intensity which is greater than $\tau_1 = 10$ photons, and use half of these peaks for the calculation in (3.9). We set the rejection tolerances in Algorithm 3 to $\tau_2 = .7$ and $\tau_3 = .95$.

Recall that autoindexing only determines the orientations of the images up to symmetry of the lattice, which is represented by the lattice rotational symmetry group $\mathcal{S}_R(\mathcal{L})$. Hence, given the correct orientation R_c we calculate the error of the computed twinned orientation $\tilde{R}$ in the Frobenius norm, modulo $\mathcal{S}_R(\mathcal{L})$:

$$\min_{R \in \mathcal{S}_R(\mathcal{L})} ||\tilde{R} - RR_c||_F, \text{ where } ||A||_F = \sqrt{\sum_{i,j} A_{i,j}^2}. \tag{4.4}$$

In the following subsections, we tabulate the number of images that were able to be autoindexed, the number rejected, and the error computed by (4.4). We also present frequency plots of the autoindexing errors.

4.3.2 Test Case 1

			Error				
J_o	Accepted	Rejected	<.001	.001-.004	.004-.016	.016-.064	>.064
21800	32678	1178	7912	21485	3196	56	29
10900	31649	2207	7686	20514	3351	75	23
4360	29833	4023	6993	19380	3327	109	24
2180	28251	5605	6583	17993	3496	149	30
1090	26683	7173	6002	16840	3566	239	36
218	22701	11155	4733	13588	3947	383	50
43.6	16640	17216	2868	8897	4233	599	43
21.8	12859	20997	1786	6224	4180	622	47
10.9	8942	24914	852	3621	3707	709	53
4.36	4884	28972	140	1268	2710	693	73
2.18	3926	29930	39	798	2290	736	63

Table 4.1: Autoindexing performance for test case 1.

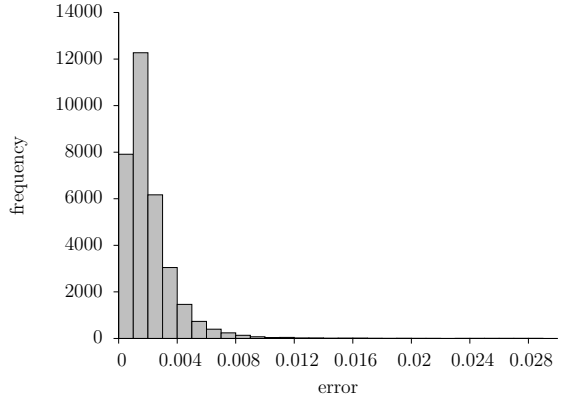


Figure 4.14: Autoindexing error for test case 1 with $J_o = 21800$.

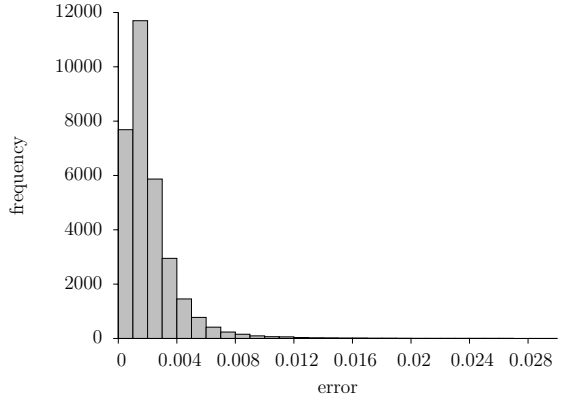


Figure 4.15: Autoindexing error for test case 1 with $J_o = 10900$.

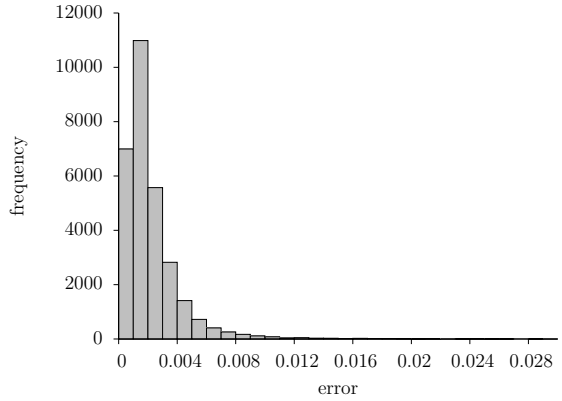


Figure 4.16: Autoindexing error for test case 1 with $J_o = 4360$.

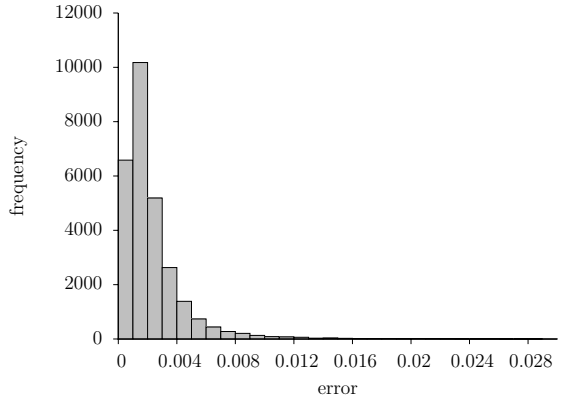


Figure 4.17: Autoindexing error for test case 1 with $J_o = 2180$.

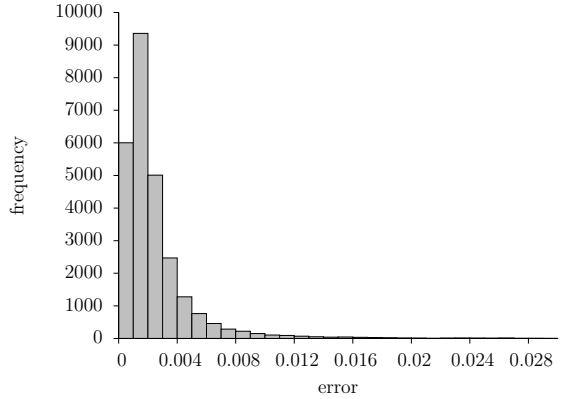


Figure 4.18: Autoindexing error for test case 1 with $J_o = 1090$.

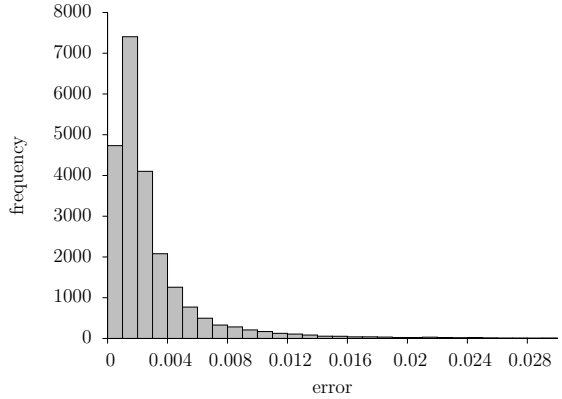


Figure 4.19: Autoindexing error for test case 1 with $J_o = 218$.

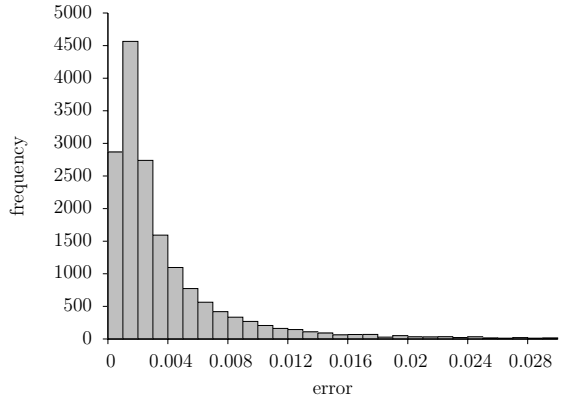


Figure 4.20: Autoindexing error for test case 1 with $J_o = 43.6$.

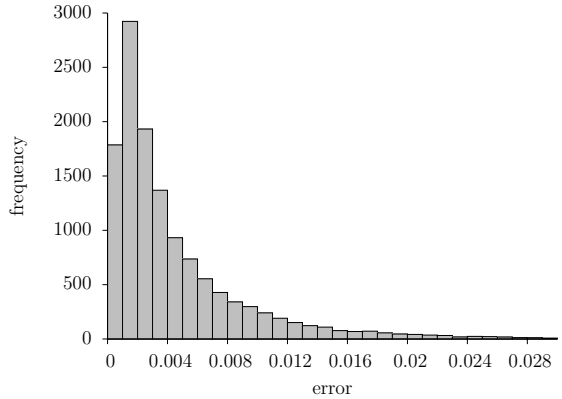


Figure 4.21: Autoindexing error for test case 1 with $J_o = 21.8$.

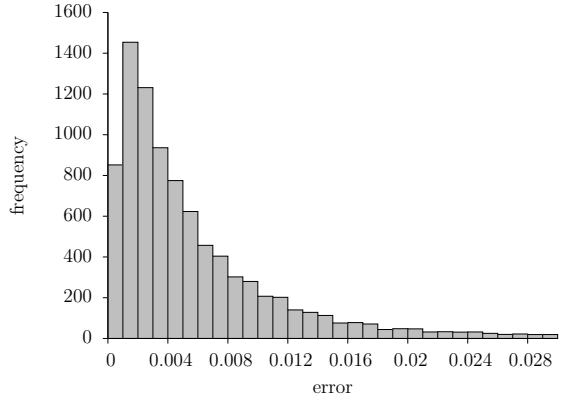


Figure 4.22: Autoindexing error for test case 1 with $J_o = 10.9$.

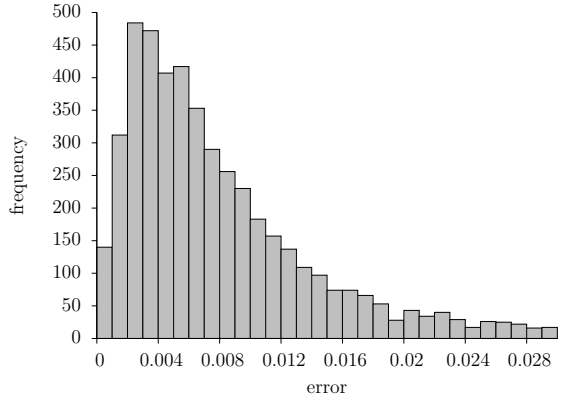


Figure 4.23: Autoindexing error for test case 1 with $J_o = 4.36$.

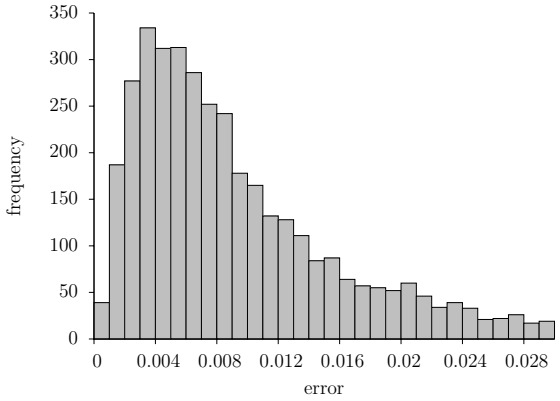


Figure 4.24: Autoindexing error for test case 1 with $J_o = 2.18$.

4.3.3 Test Case 2

			Error				
J_o	Accepted	Rejected	<.001	.001-.004	.004-.016	.016-.064	>.064
21800	31883	1973	9814	22225	2662	151	31
10900	30988	2868	6724	21504	2529	189	42
4360	29466	4390	6563	20027	2584	238	54
2180	28000	5856	6293	18925	2469	256	57
1090	26045	7811	6018	17407	2247	308	65
218	21099	12757	5525	13018	2100	350	106
43.6	15227	18629	4641	8172	1992	328	94
21.8	12520	21336	3922	6358	1863	292	85
10.9	9871	23985	1999	4898	1636	249	89
4.36	6478	27378	1658	3248	1300	197	75
2.18	4277	29579	976	2200	905	145	51

Table 4.2: Autoindexing performance for test case 2.

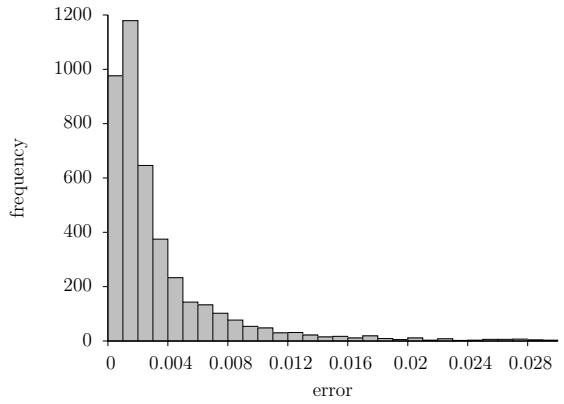


Figure 4.25: Autoindexing error for test case 2 with $J_o = 21800$.

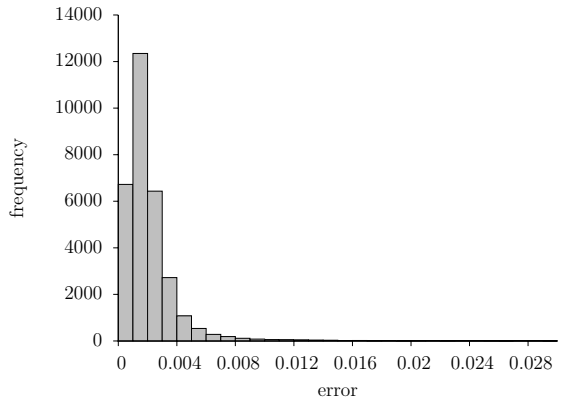


Figure 4.26: Autoindexing error for test case 2 with $J_o = 10900$.

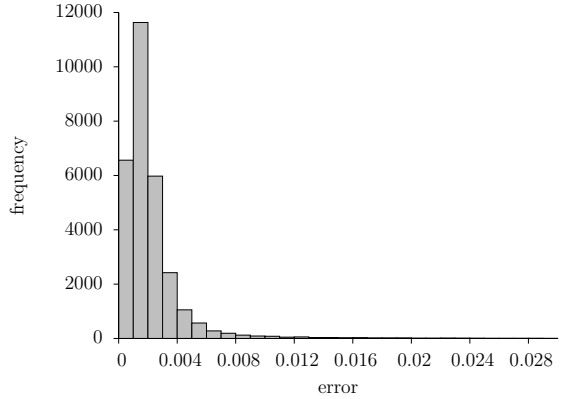


Figure 4.27: Autoindexing error for test case 2 with $J_o = 4360$.

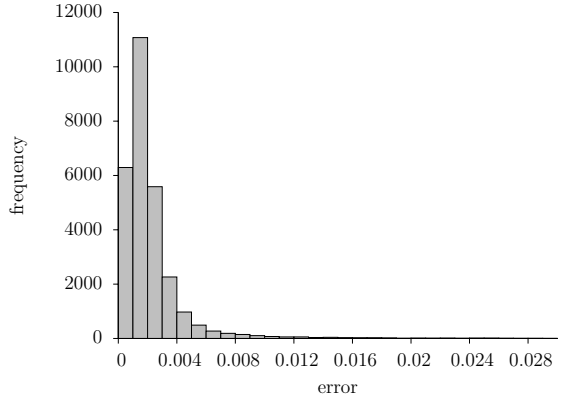


Figure 4.28: Autoindexing error for test case 2 with $J_o = 2180$.

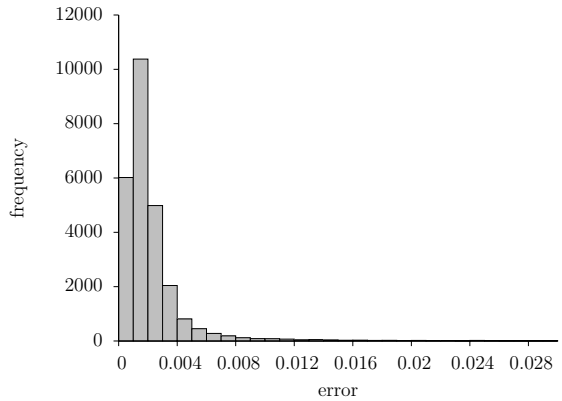


Figure 4.29: Autoindexing error for test case 2 with $J_o = 1090$.

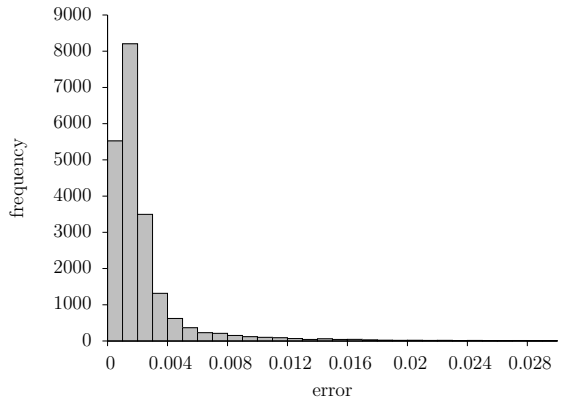


Figure 4.30: Autoindexing error for test case 2 with $J_o = 218$.

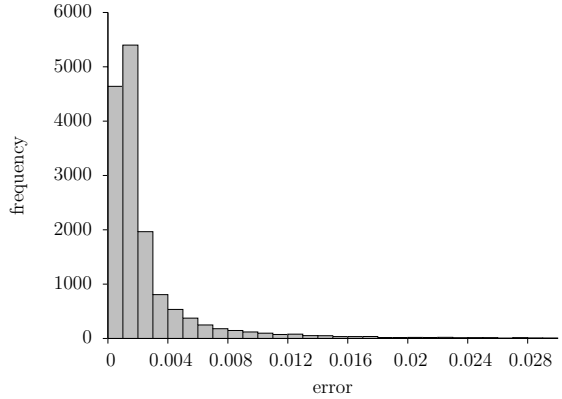


Figure 4.31: Autoindexing error for test case 2 with $J_o = 43.6$.

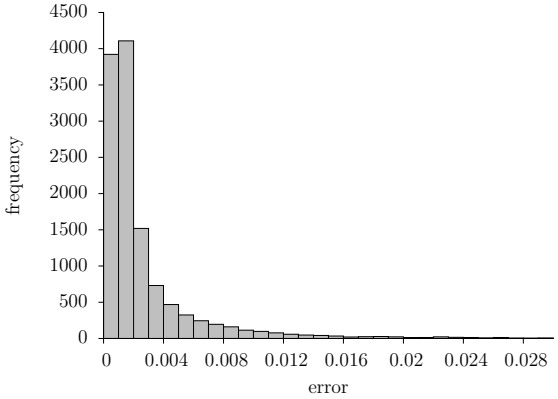


Figure 4.32: Autoindexing error for test case 2 with $J_o = 21.8$.

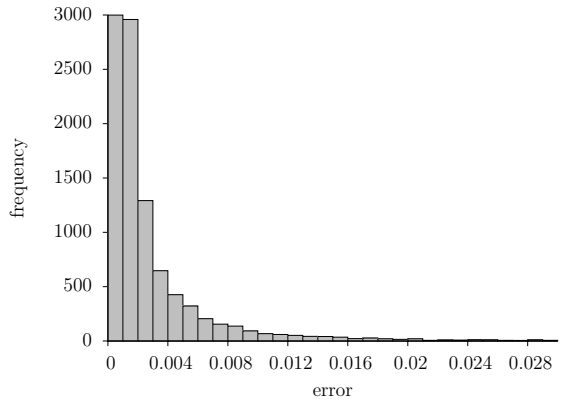


Figure 4.33: Autoindexing error for test case 2 with $J_o = 10.9$.

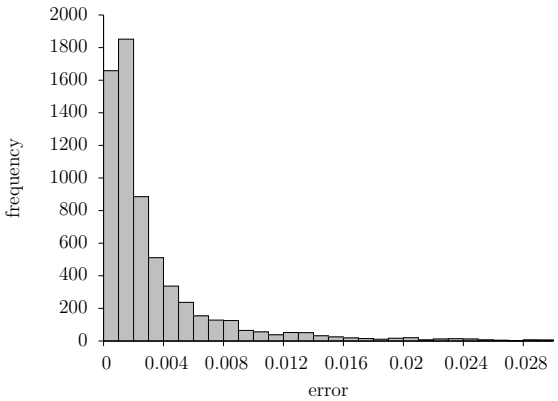


Figure 4.34: Autoindexing error for test case 2 with $J_o = 4.36$.

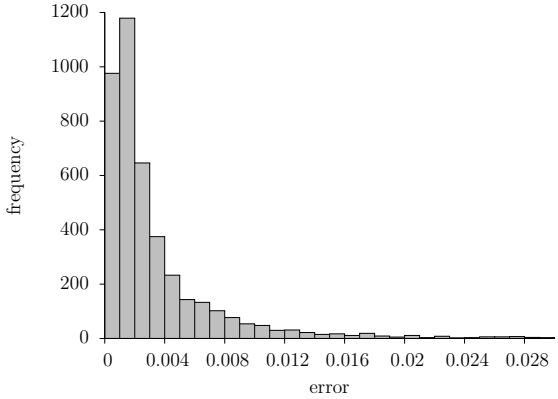


Figure 4.35: Autoindexing error for test case 2 with $J_o = 2.18$.

4.3.4 Test Case 3

			Error				
J_o	Accepted	Rejected	<.001	.001-.004	.004-.016	.016-.064	>.064
21800	33791	65	12859	20703	226	2	1
10900	33732	124	13038	20422	269	3	0
4360	33541	315	13291	19931	310	6	3
2180	33325	531	13312	19667	329	15	2
1090	33014	842	13482	19089	424	17	2
218	31152	2704	12703	17902	517	29	1

Table 4.3: Autoindexing performance for test case 3.

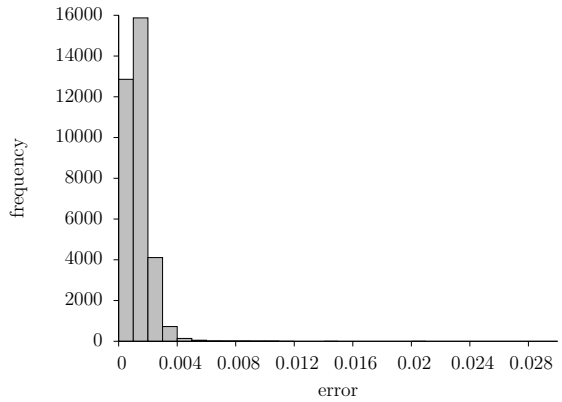


Figure 4.36: Autoindexing error for test case 3 with $J_o = 21800$.

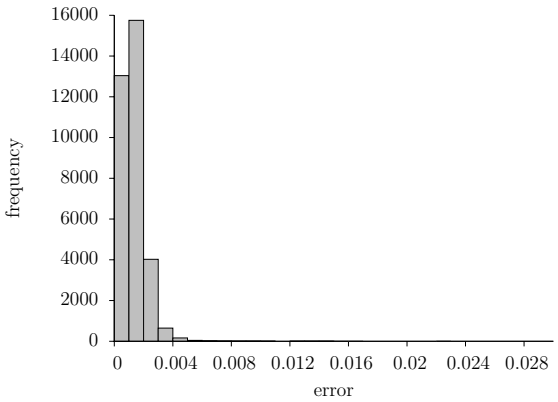


Figure 4.37: Autoindexing error for test case 3 with $J_o = 10900$.

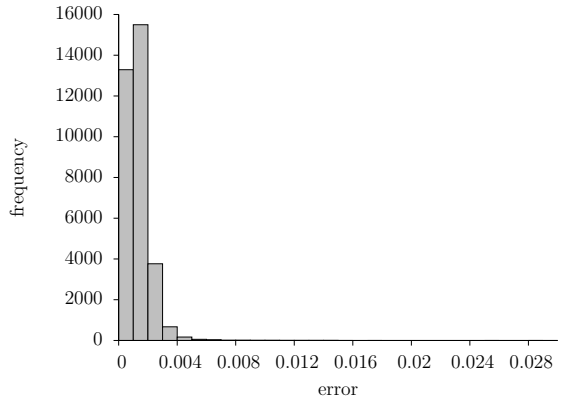


Figure 4.38: Autoindexing error for test case 3 with $J_o = 4360$.

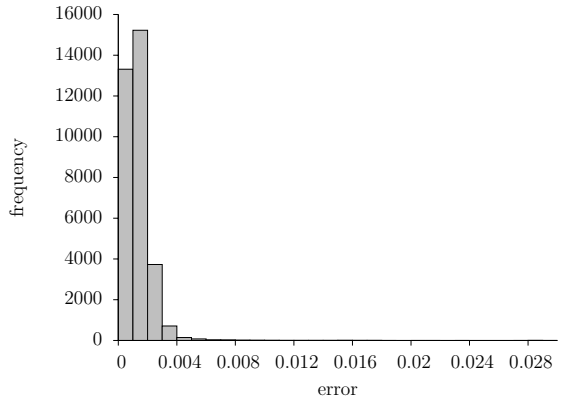


Figure 4.39: Autoindexing error for test case 3 with $J_o = 2180$.

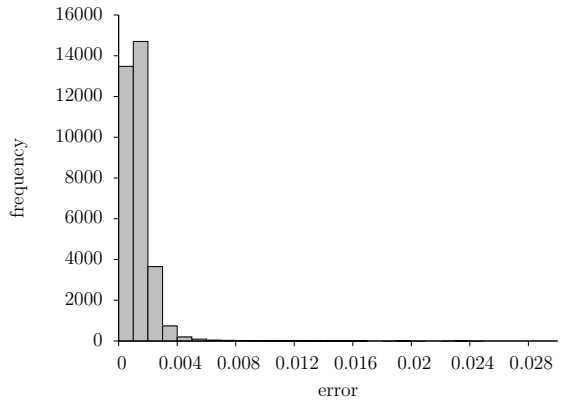


Figure 4.40: Autoindexing error for test case 3 with $J_o = 1090$.

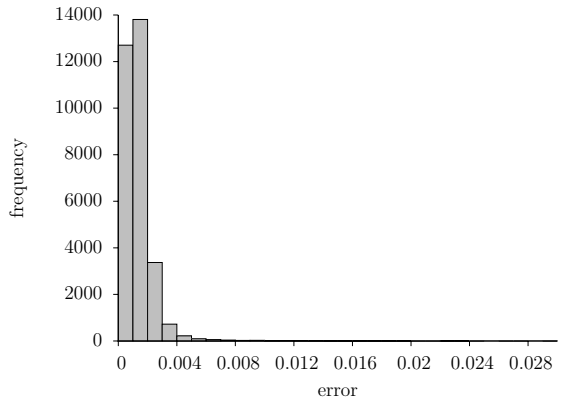


Figure 4.41: Autoindexing error for test case 3 with $J_o = 218$.

4.4 Crystal Size Determination

4.4.1 Test Description

Here we test the performance of our crystal size determination technique from Section 3.3. We set $N_p = 204$, $\tau_1 = .1$, and $\tau_2 = 1.5$ in Algorithm 4 and reject any computed sizes which correspond to lengths less than 700 Å or greater than 9000 Å. If the crystal sizes are rejected, then we repeat the process at the next brightest Bragg peak, and repeat this process for up to 5 peaks.

Since the computed structure factors scale as a power of the reciprocal product of the crystal sizes, a natural metric is the relative error in the geometric average. In particular, we compute the error as the relative difference between the geometric averages of the computed sizes (N_1, N_2, N_3) and the correct sizes $(N_{c,1}, N_{c,2}, N_{c,3})$:

$$\frac{|(N_1N_2N_3)^{\frac{1}{3}} - (N_{c,1}N_{c,2}N_{c,3})^{\frac{1}{3}}|}{(N_{c,1}N_{c,2}N_{c,3})^{\frac{1}{3}}}. \tag{4.5}$$

In the following subsections, we tabulate the number of images whose computed crystal sizes were accepted, the number rejected, and the error computed by (4.5). We also present frequency plots of the crystal size errors.

4.4.2 Test Case 1

			Error					
J_o	Accepted	Rejected	<.1	.1-.2	.2-.3	.3-.4	.4-.5	>.5
21800	30506	2172	19652	9667	903	98	26	160
10900	29568	2081	18829	9546	924	97	26	146
4360	27750	2083	17376	9183	983	85	30	93
2180	26260	1991	16080	8996	992	71	21	100
1090	24711	1972	14867	8575	1029	74	16	150
218	20195	2506	10941	7681	1166	126	33	248
43.6	14056	2584	6150	6074	1343	171	45	273
21.8	10639	2220	4071	4626	1388	184	50	320
10.9	6699	2243	2255	2751	1113	184	64	332
4.36	2883	1991	828	906	645	139	50	325
2.18	1851	2075	458	458	387	115	54	382

Table 4.4: Crystal size determination performance for test case 1.

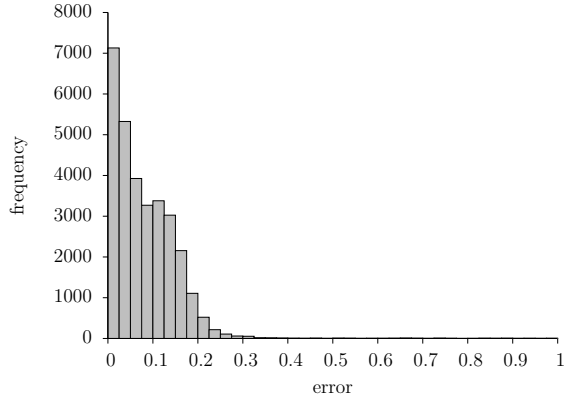


Figure 4.42: Computed crystal size error for test case 1 with $J_o = 21800$.

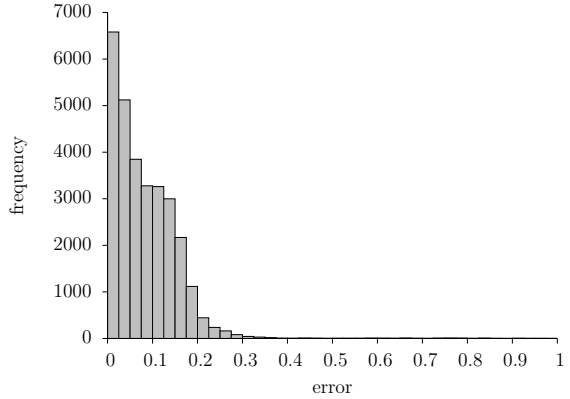


Figure 4.43: Computed crystal size error for test case 1 with $J_o = 10900$.

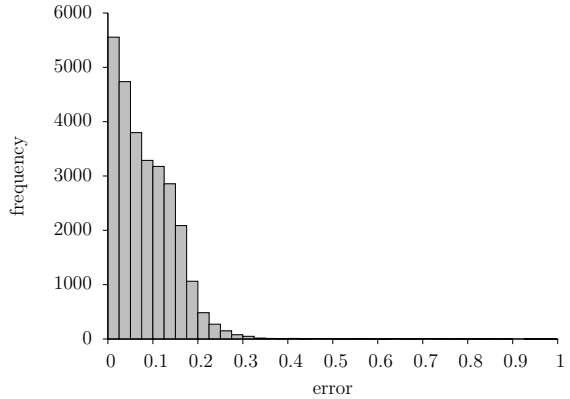


Figure 4.44: Computed crystal size error for test case 1 with $J_o = 4360$.

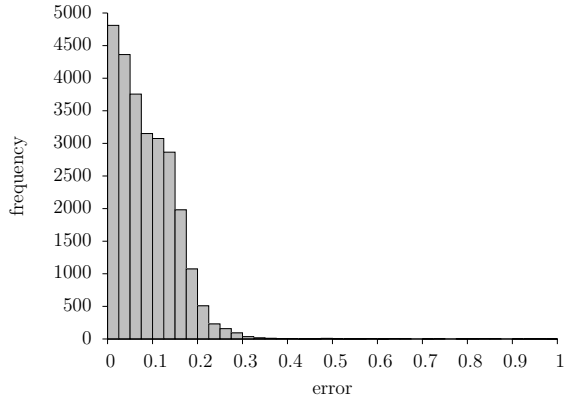


Figure 4.45: Computed crystal size error for test case 1 with $J_o = 2180$.

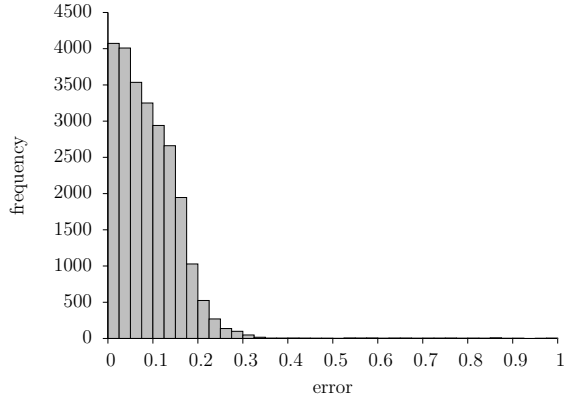


Figure 4.46: Computed crystal size error for test case 1 with $J_o = 1090$.

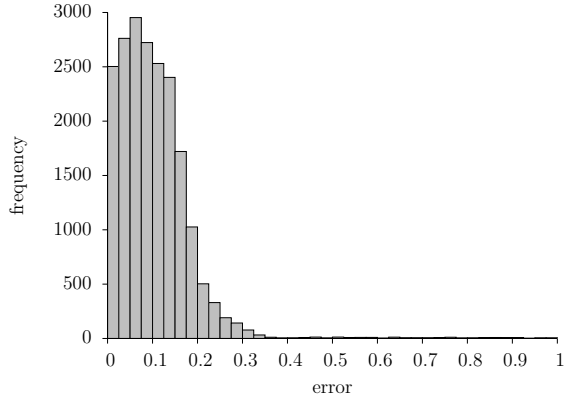


Figure 4.47: Computed crystal size error for test case 1 with $J_o = 218$.

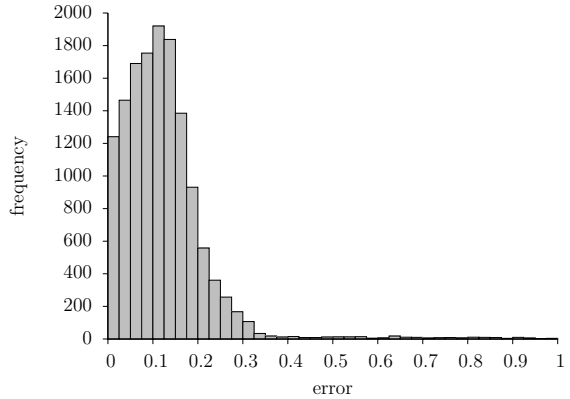


Figure 4.48: Computed crystal size error for test case 1 with $J_o = 43.6$.

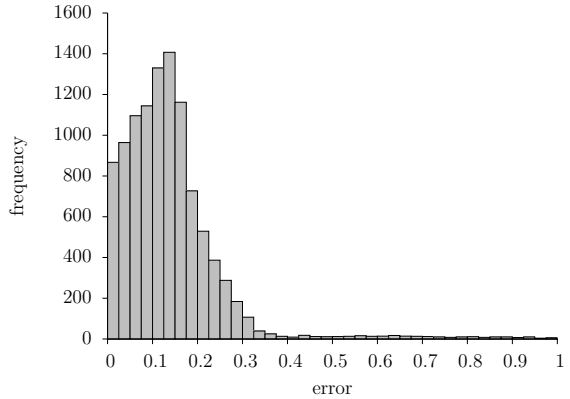


Figure 4.49: Computed crystal size error for test case 1 with $J_o = 21.8$.

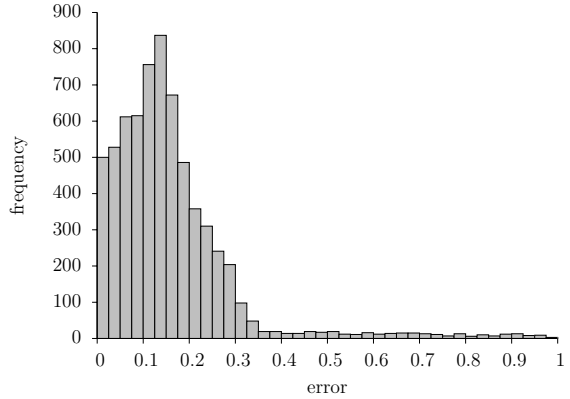


Figure 4.50: Computed crystal size error for test case 1 with $J_o = 10.9$.

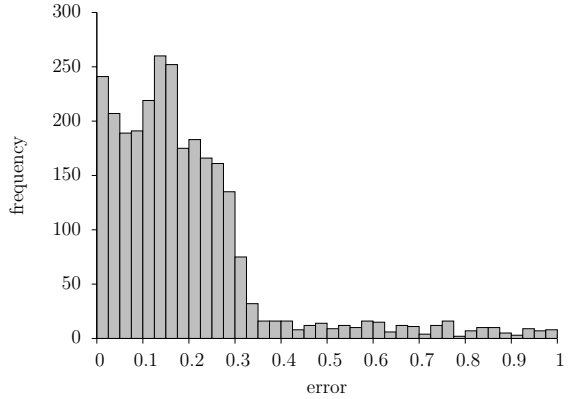


Figure 4.51: Computed crystal size error for test case 1 with $J_o = 4.36$.

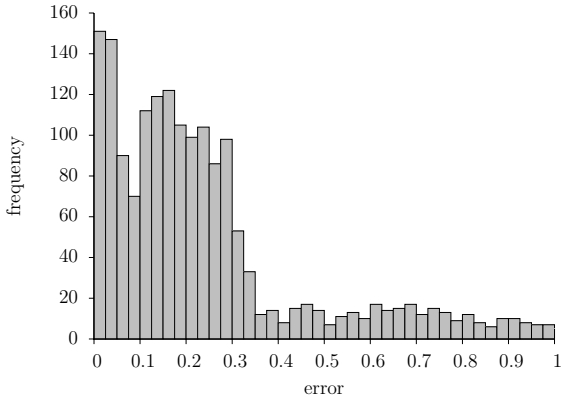


Figure 4.52: Computed crystal size error for test case 1 with $J_o = 2.18$.

4.4.3 Test Case 2

			Error					
J_o	Accepted	Rejected	<.1	.1-.2	.2-.3	.3-.4	.4-.5	>.5
21800	20923	10960	13279	5283	1414	572	203	172
10900	19114	11874	12398	4802	1218	426	147	123
4360	16650	12816	10880	4359	960	261	97	93
2180	14119	13881	9074	3912	795	194	79	65
1090	11726	14319	7544	3315	637	120	51	59
218	6174	14925	3482	2108	460	60	31	33
43.6	2439	12788	1120	970	310	27	4	8
21.8	1537	10983	591	683	239	16	4	4
10.9	874	8997	283	406	171	6	3	5
4.36	357	6121	80	156	113	6	0	2
2.18	152	4125	24	69	57	1	0	1

Table 4.5: Crystal size determination performance for test case 2.

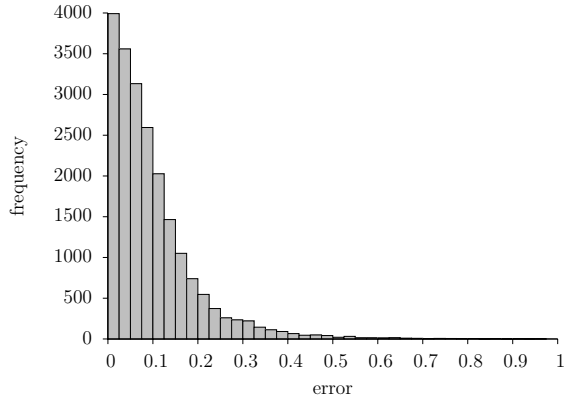


Figure 4.53: Computed crystal size error for test case 2 with $J_o = 21800$.

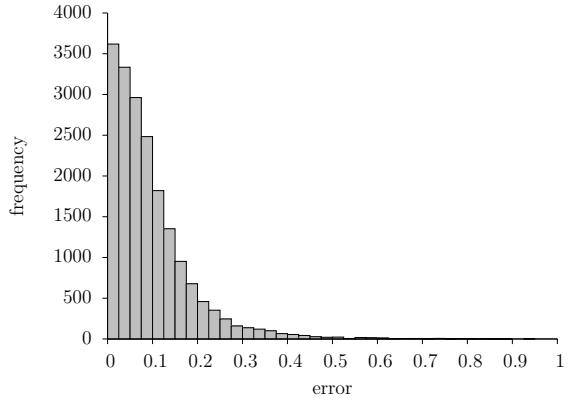


Figure 4.54: Computed crystal size error for test case 2 with $J_o = 10900$.

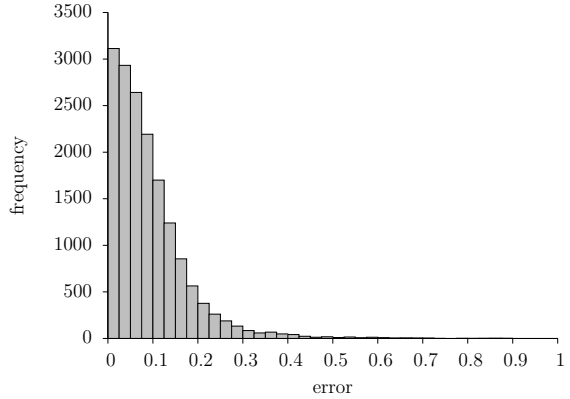


Figure 4.55: Computed crystal size error for test case 2 with $J_o = 4360$.

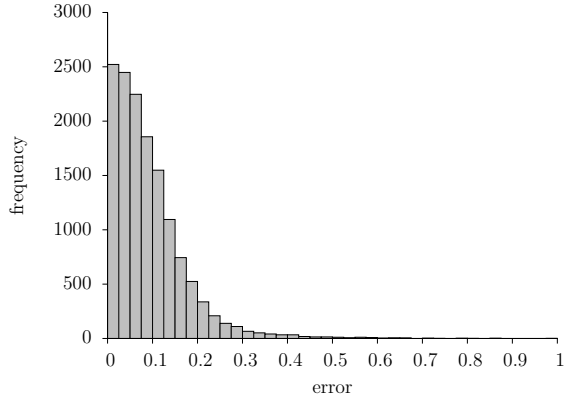


Figure 4.56: Computed crystal size error for test case 2 with $J_o = 2180$.

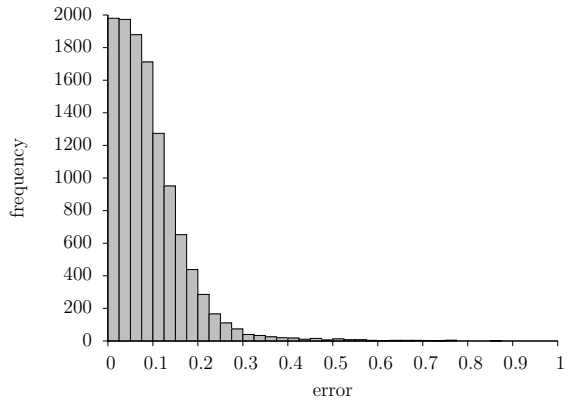


Figure 4.57: Computed crystal size error for test case 2 with $J_o = 1090$.

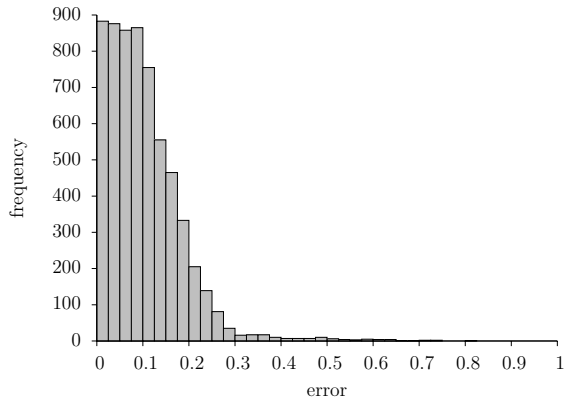


Figure 4.58: Computed crystal size error for test case 2 with $J_o = 218$.

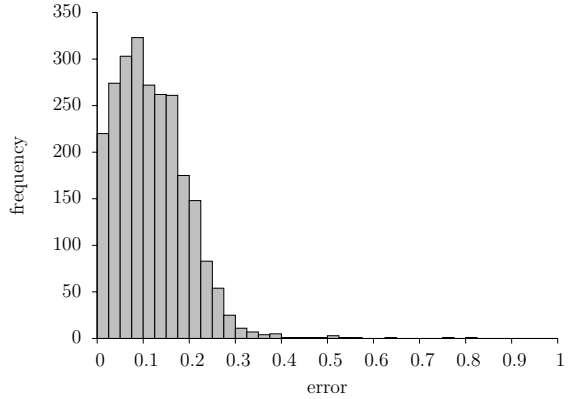


Figure 4.59: Computed crystal size error for test case 2 with $J_o = 43.6$.

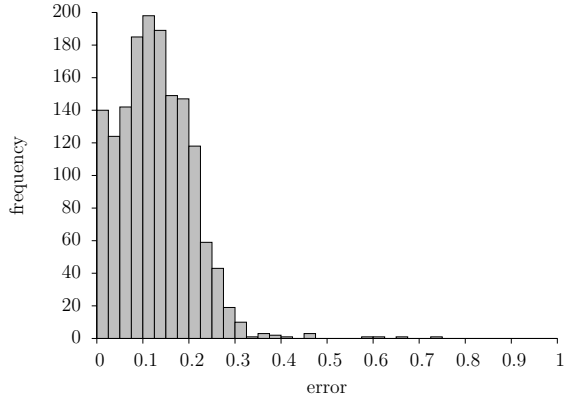


Figure 4.60: Computed crystal size error for test case 2 with $J_o = 21.8$.

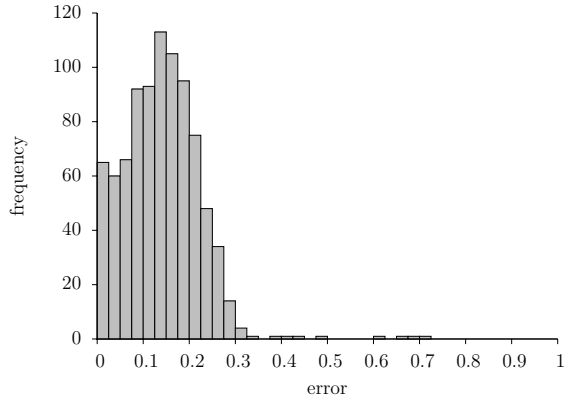


Figure 4.61: Computed crystal size error for test case 2 with $J_o = 10.9$.

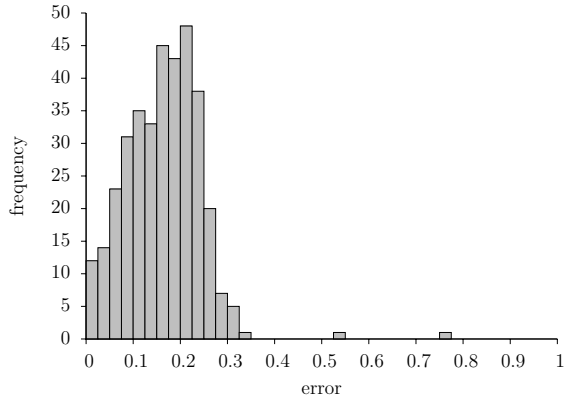


Figure 4.62: Computed crystal size error for test case 2 with $J_o = 4.36$.

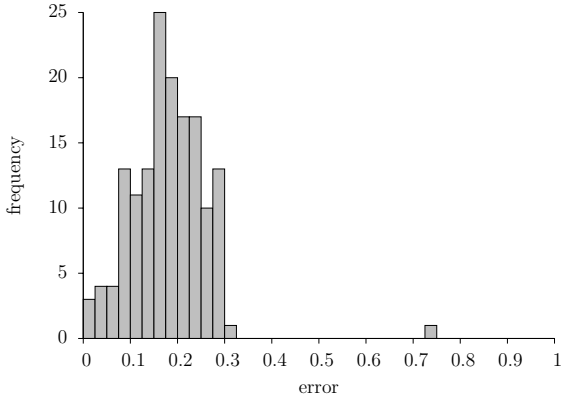


Figure 4.63: Computed crystal size error for test case 2 with $J_o = 2.18$.

4.4.4 Test Case 3

			Error					
J_o	Accepted	Rejected	<.1	.1-.2	.2-.3	.3-.4	.4-.5	>.5
21800	20443	13348	13484	5906	808	162	44	39
10900	19511	14221	12942	5629	724	136	42	38
4360	18209	15332	12182	5169	698	91	35	34
2180	16888	16437	11270	4826	663	75	28	26
1090	15351	15801	9972	4645	620	53	33	28
218	10687	22327	6465	3624	511	44	17	26

Table 4.6: Crystal size determination performance for test case 3.

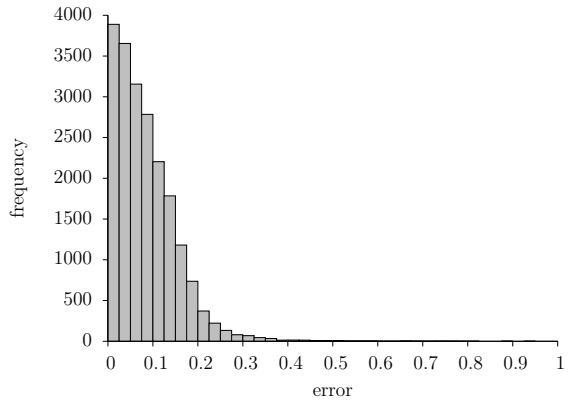


Figure 4.64: Computed crystal size error for test case 3 with $J_o = 21800$.

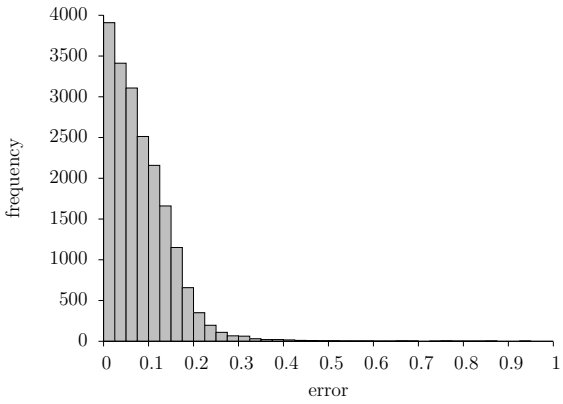


Figure 4.65: Computed crystal size error for test case 3 with $J_o = 10900$.

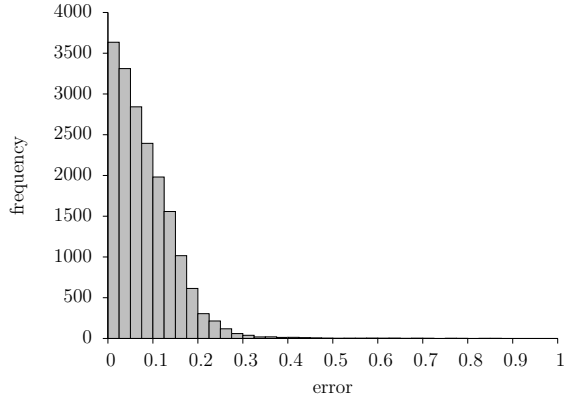


Figure 4.66: Computed crystal size error for test case 3 with $J_o = 4360$.

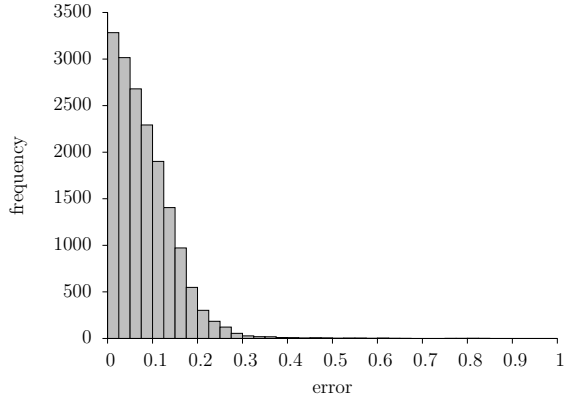


Figure 4.67: Computed crystal size error for test case 3 with $J_o = 2180$.

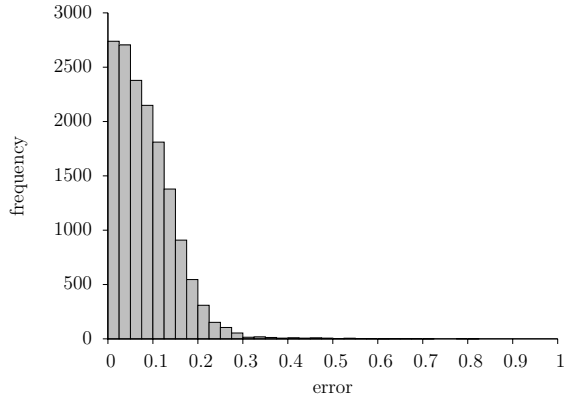


Figure 4.68: Computed crystal size error for test case 3 with $J_o = 1090$.

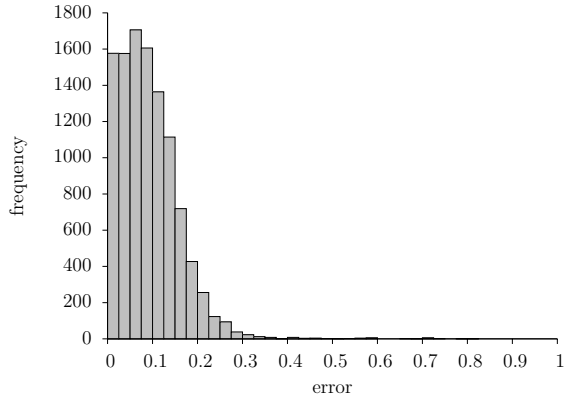


Figure 4.69: Computed crystal size error for test case 3 with $J_o = 218$.

4.5 Structure Factor Magnitude Modeling

4.5.1 Test Description

Here we test our structure factor magnitude modeling technique from Section 3.4. For the expectation maximization step, we perform outlier rejection by removing any of the variance stabilized structure factor magnitudes $w_{i,m}$ in which $\frac{1}{K^2}\sum_{j=1}^{K}\sigma_{i,j}^{(n)}\sum_{k=1}^{K}\mathcal{G}(w_{i,m},\mu_{i,k}^{(n)},\sigma_{i,k}^{(n)}) < \tau_c$, with τ_c set to 10^{-5} for test cases 1 and 2 and 2×10^{-2} for test case 3. We initialized the standard deviations $\sigma_{i,j}$ to be .1 for test cases 1 and 2 and .05 for test case 3. For the scaling correction step, we found that it was best to replace the standard deviations computed in the expectation maximization step with a scaled average over all images $\sigma_{\text{avg}} = \frac{1}{\tau MK}\sum_{i,k}\sigma_{i,k}$,

where M is the number of images, K is the order of the twinning ambiguity, and τ is a scaling constant. We set $\tau = 8$ for test case 1, $\tau = 4$ for test case 2, and $\tau = 2$ for test case 3. We performed 50 iterations of alternating between expectation maximization and scaling, and used 20 iterations for each of the expectation maximization steps. In practice, a good choice of parameters can be found by examining their effects on a select few of the scaled histograms.

In the following subsections, we present histograms of the initial unscaled variance stabilized structure factor magnitude data along with the scaled data and multi-modal model at select reciprocal lattice points.

4.5.2 Test Case 1

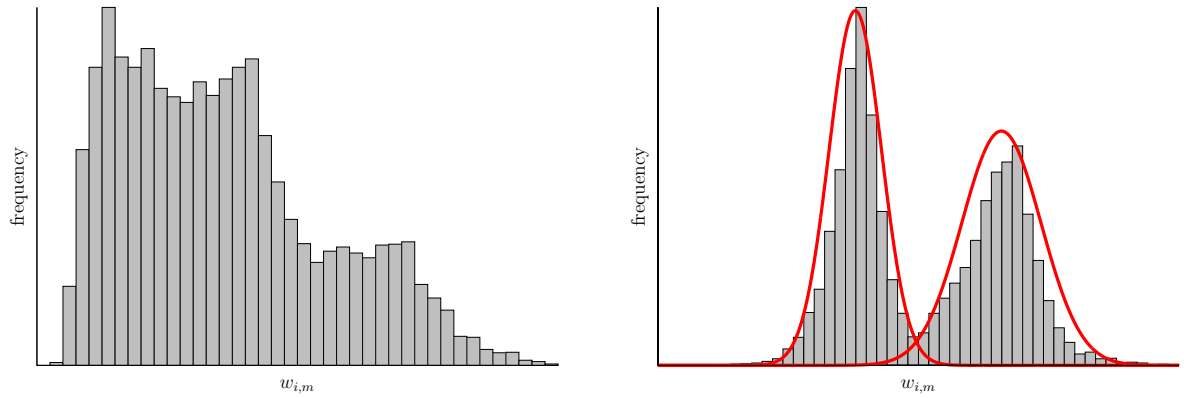


Figure 4.70: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 21800$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

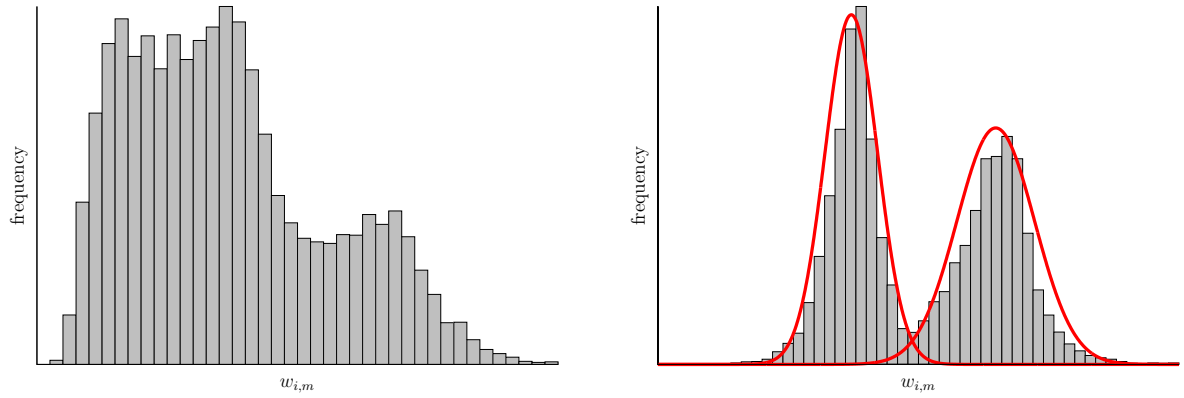


Figure 4.71: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 10900$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

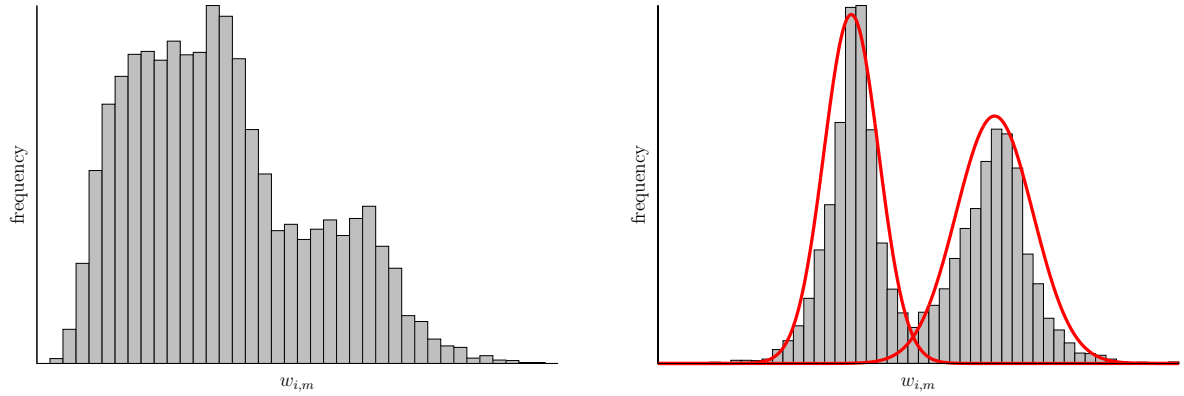


Figure 4.72: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 4360$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

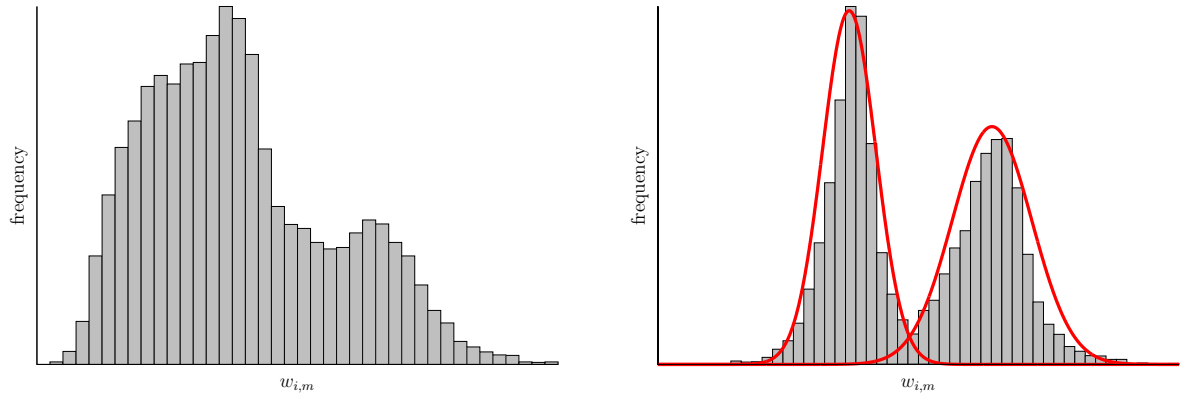


Figure 4.73: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 2180$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

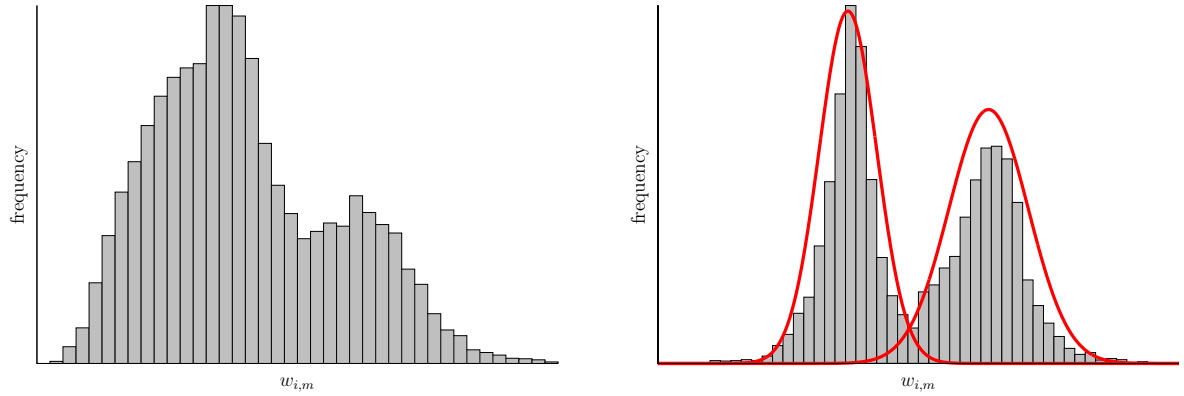


Figure 4.74: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 1090$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

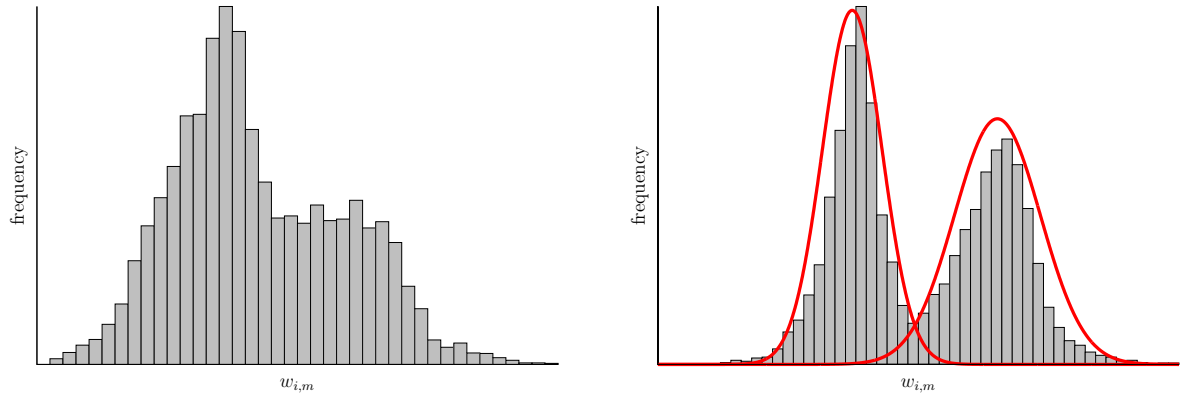


Figure 4.75: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 218$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

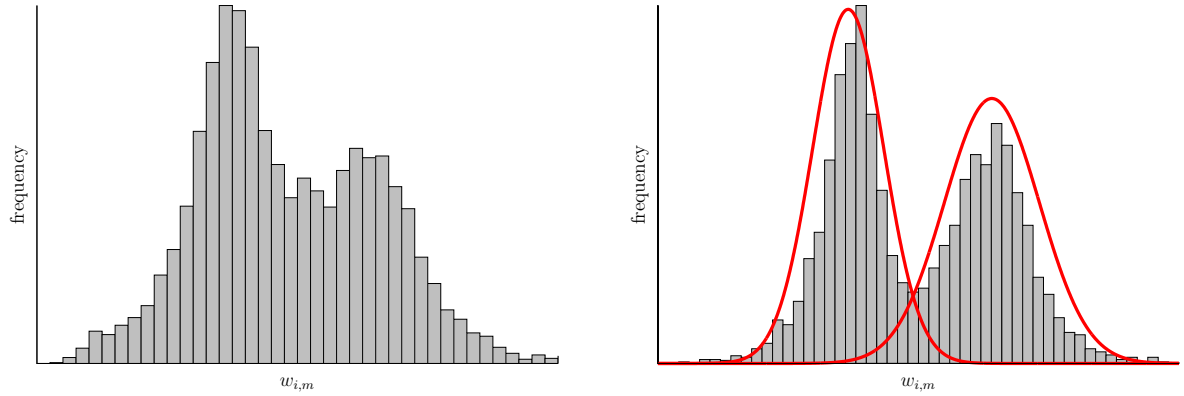


Figure 4.76: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 43.6$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

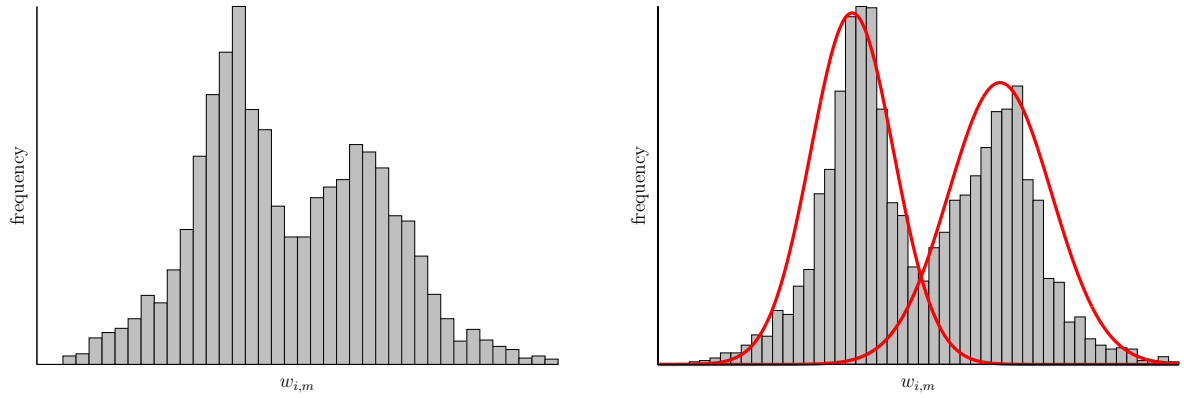


Figure 4.77: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 21.8$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

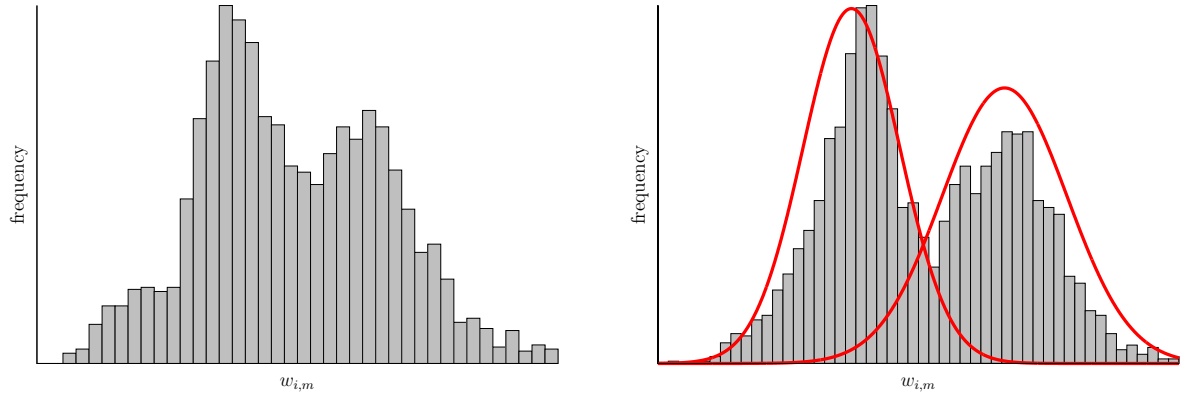


Figure 4.78: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 10.9$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

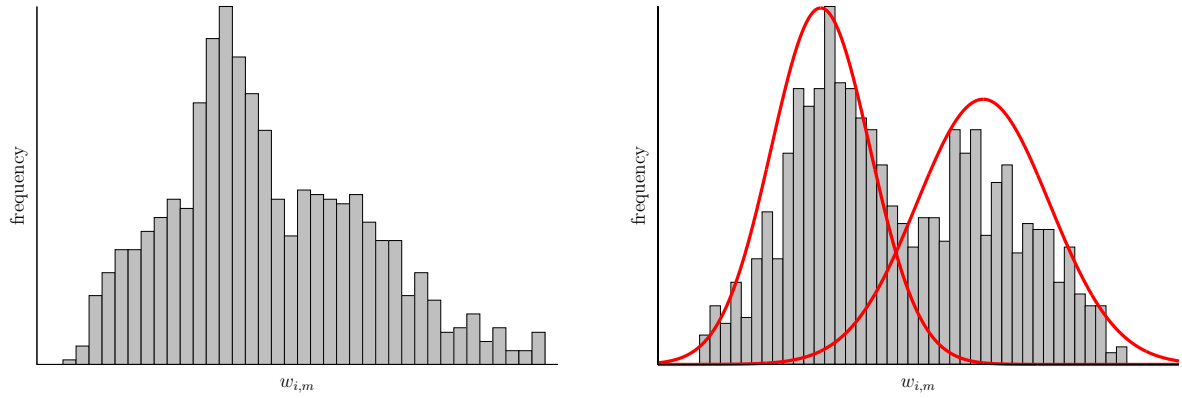


Figure 4.79: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 4.36$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

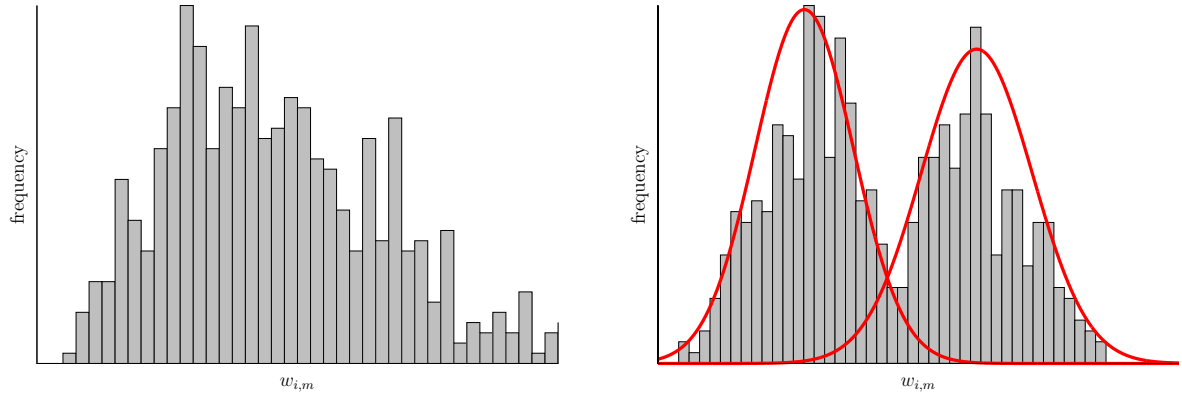


Figure 4.80: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (3,2,1), for test case 1 with $J_o = 2.18$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

4.5.3 Test Case 2

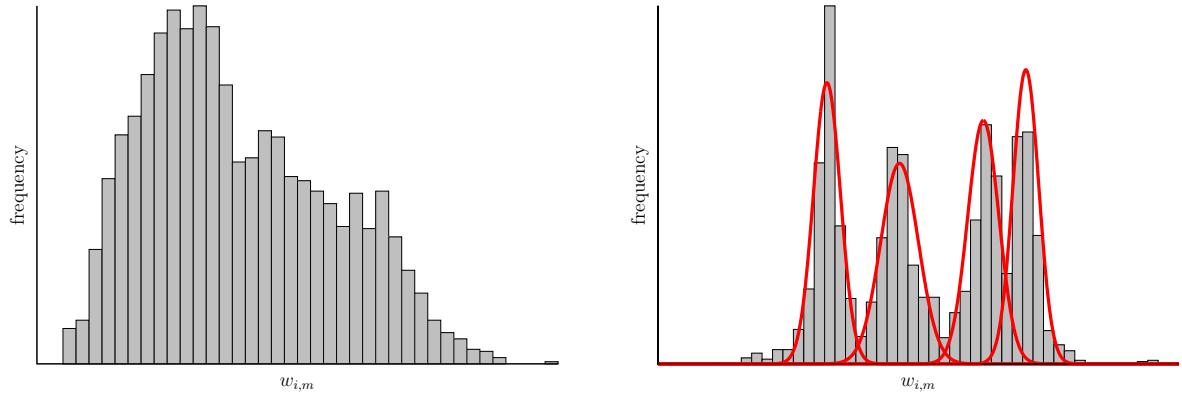


Figure 4.81: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 21800$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

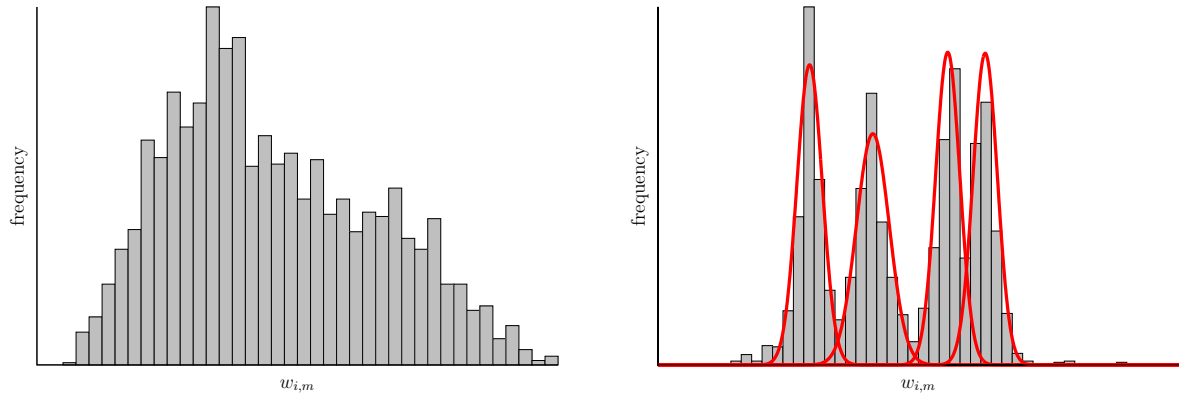


Figure 4.82: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 10900$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

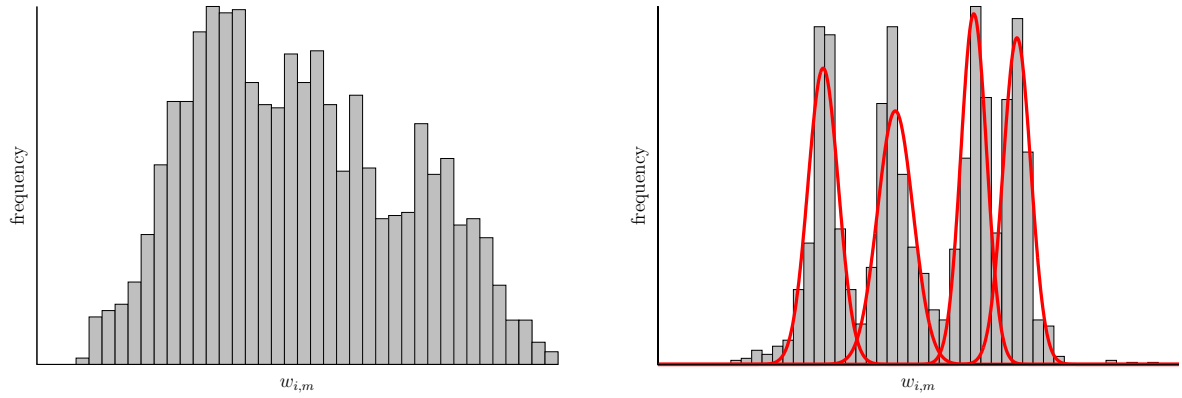


Figure 4.83: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 4360$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

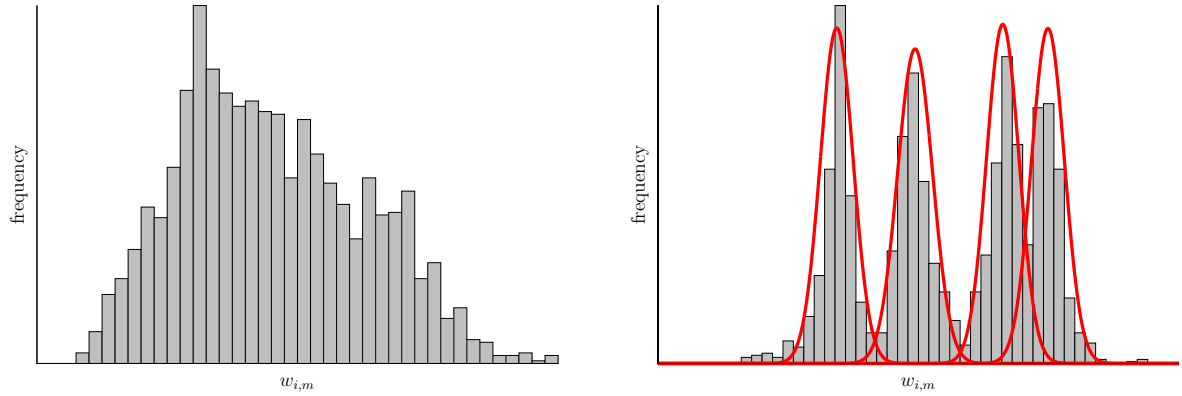


Figure 4.84: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 2180$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

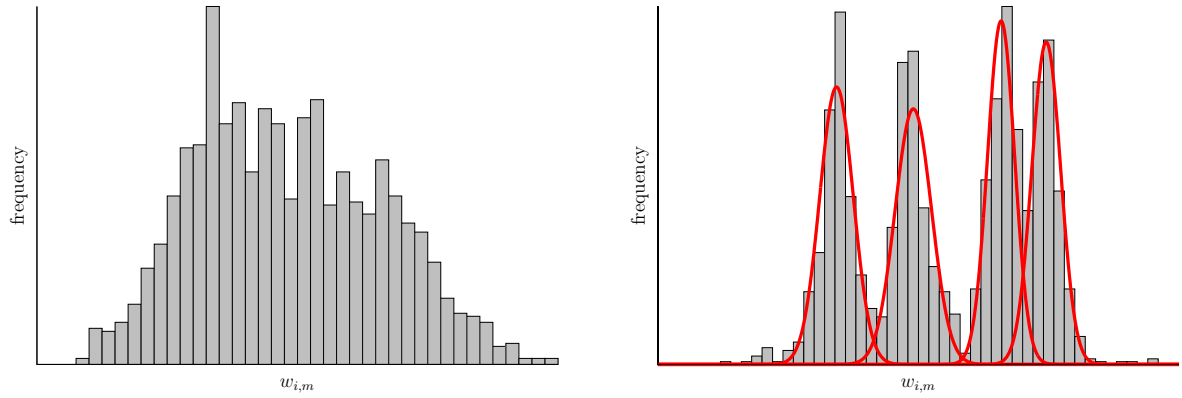


Figure 4.85: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 1090$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

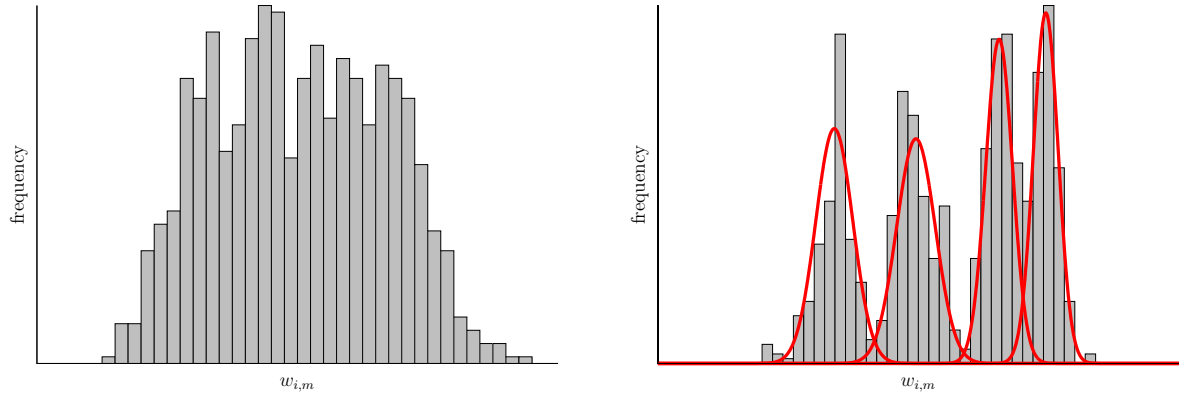


Figure 4.86: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 218$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

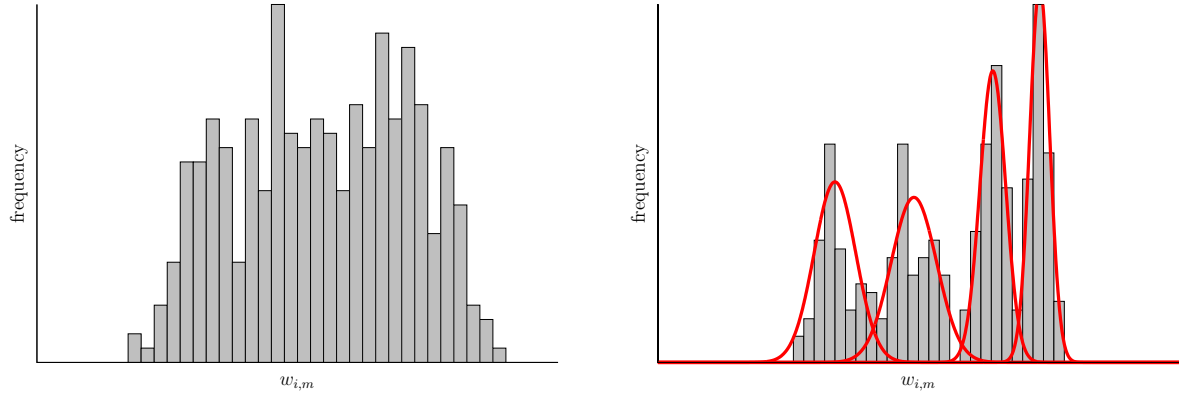


Figure 4.87: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 43.6$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

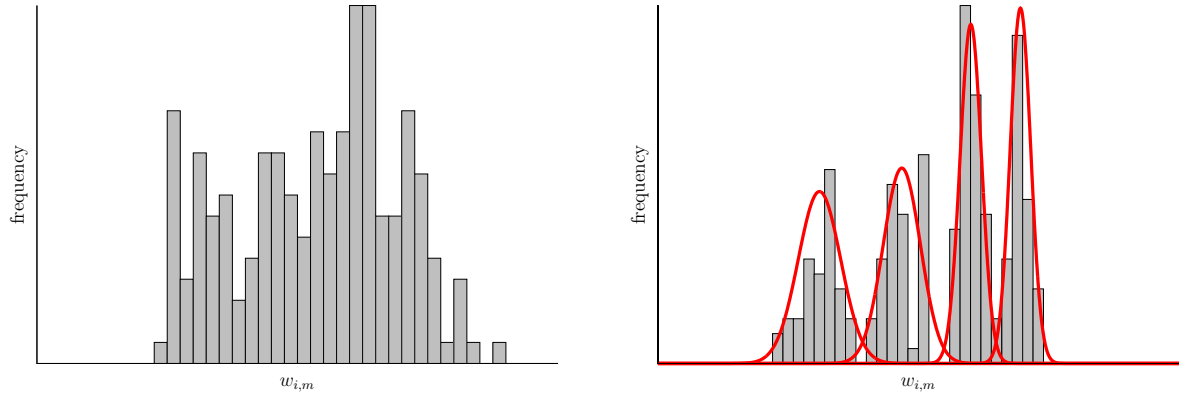


Figure 4.88: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 21.8$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

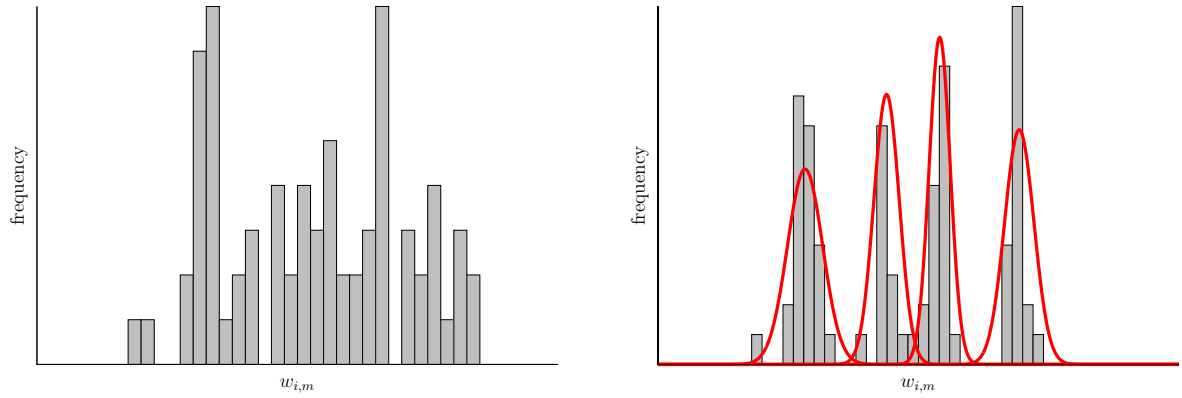


Figure 4.89: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,1), for test case 2 with $J_o = 10.9$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

4.5.4 Test Case 3

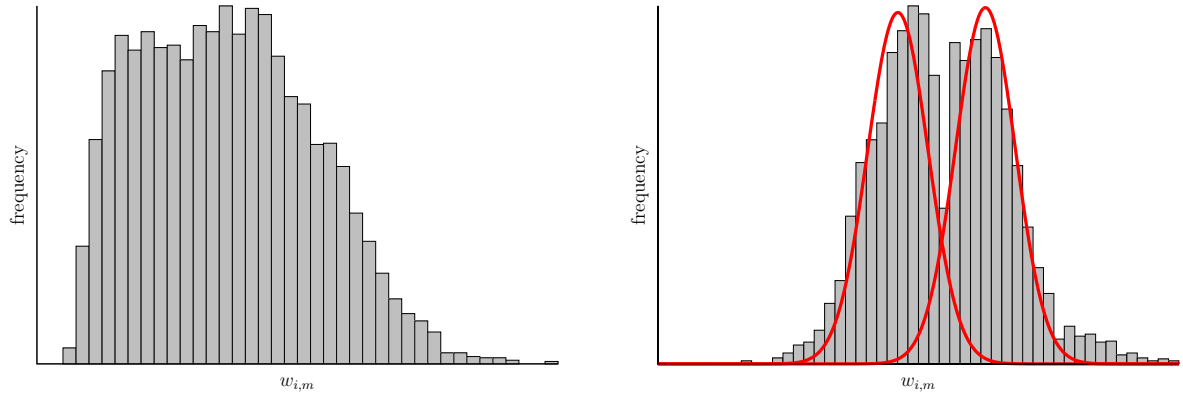


Figure 4.90: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices $(1,1,\frac{1}{2})$, for test case 3 with $J_o = 21800$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

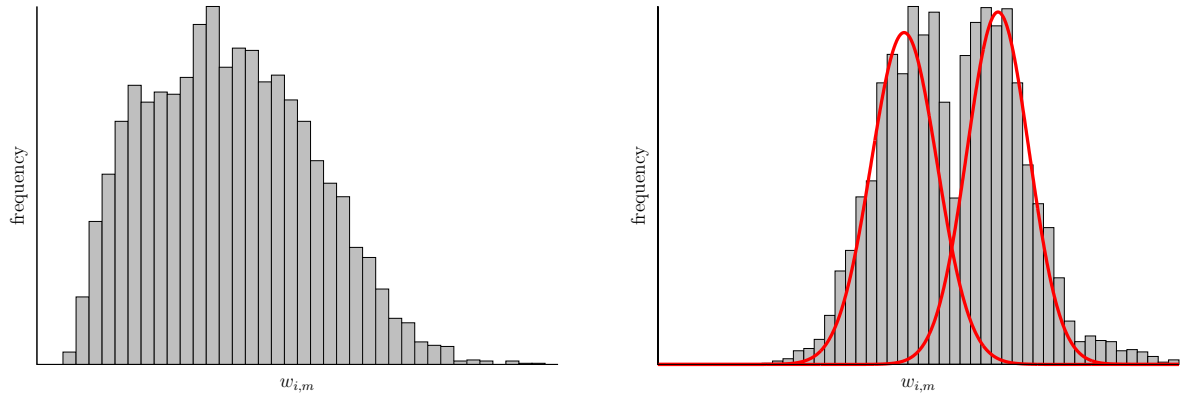


Figure 4.91: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,$\frac{1}{2}$), for test case 3 with $J_o = 10900$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

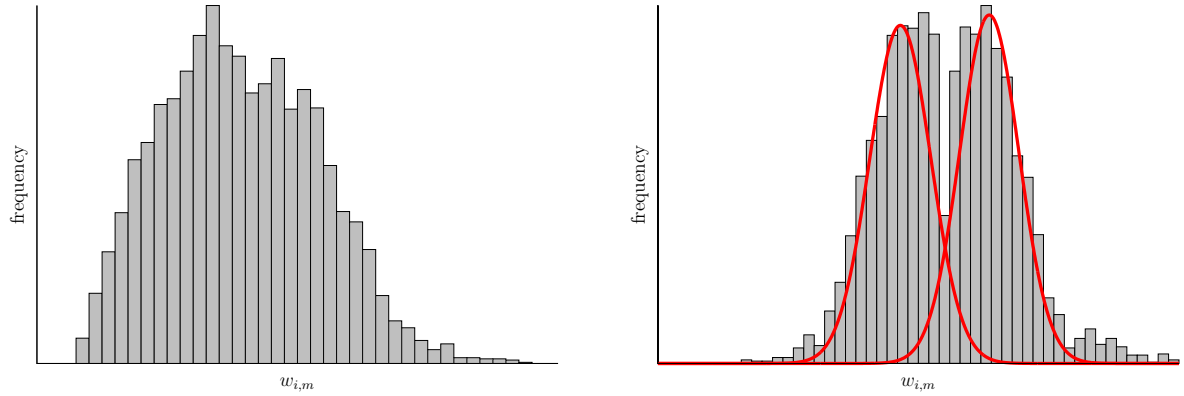


Figure 4.92: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,$\frac{1}{2}$), for test case 3 with $J_o = 4360$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

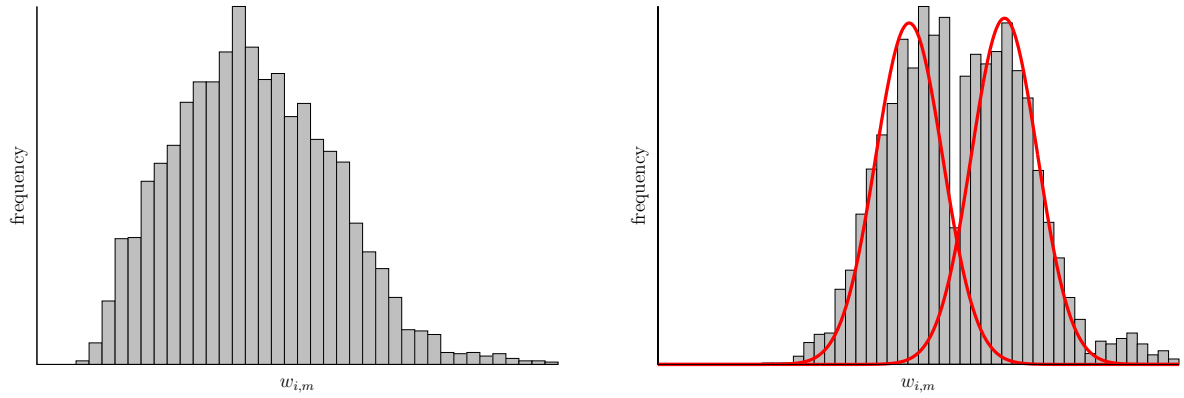


Figure 4.93: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,$\frac{1}{2}$), for test case 3 with $J_o = 2180$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

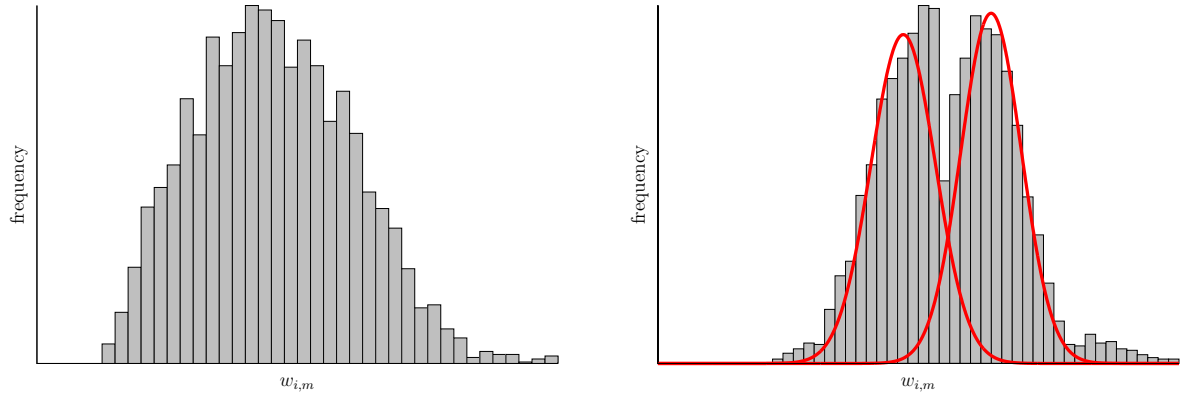


Figure 4.94: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices (1,1,$\frac{1}{2}$), for test case 3 with $J_o = 1090$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

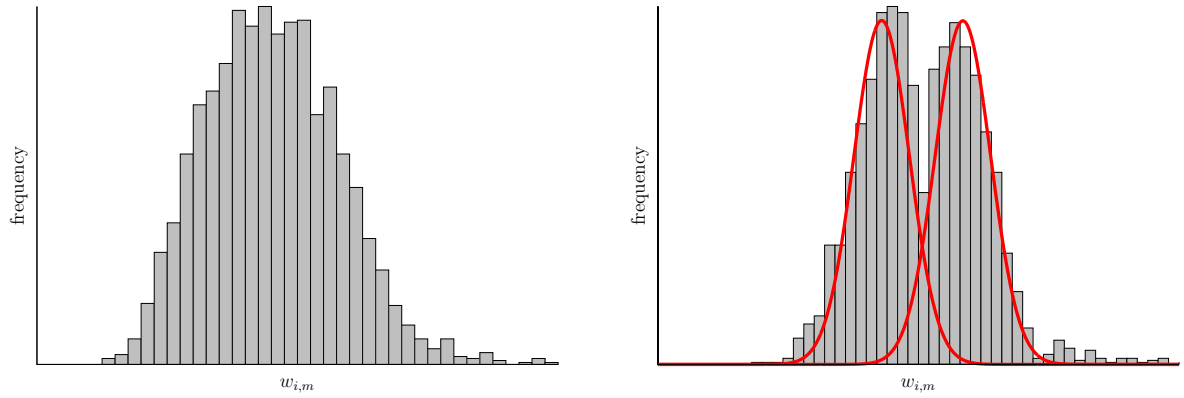


Figure 4.95: Frequency plots of the possible variance stabilized structure factor magnitudes at the reciprocal lattice point with Miller indices $(1,1,\frac{1}{2})$, for test case 3 with $J_o = 218$. Left: Unscaled data. Right: Scaled data with multi-modal Gaussian model (red).

4.6 Solving the Twinning Problem

4.6.1 Test Description

Here we test our approach to solving the twinning problem. We use up to 23 reciprocal lattice points in each Bravais direction, which corresponds to graphical models with node counts on the order of 10^3-10^4 and edge counts on the order of 10^6-10^8. We replace the standard deviations computed from the expectation maximization step with a scaled average, as was done in Section 4.5.1. A detwinned orientation is rejected if there is another possible orientation whose error in (3.43) is less than a factor of 1.2 times the minimum error. This tolerance controls the selectivity of the orientations: every computed orientation will be accepted if it is too close to 1 and everything will be rejected if it is too large. However, apart from these extremes, the accuracy of the orientation calculation is largely insensitive to the value of this tolerance.

Note that, for K-fold twinning, there are up to K valid solutions for the set of detwinned orientations, which are related to each other through multiplication by elements of the lattice rotational symmetry group $\mathcal{S}_R(\mathcal{L})$. Therefore, if $\{R_{c,i}\}_i$ is the set of correct orientations used to generate the images and $\{R_i\}_i$ is a consistent choice for the set of computed orientations, then there is an $R \in \mathcal{S}_R(\mathcal{L})$ such that for every i, $R_i \approx RR_{c,i}$. Hence, we measure the number of correctly detwinned orientations via

$$\max_{R_a \in \mathcal{S}_R(\mathcal{L})} \left| \{R_i : ||R_i - R_a R_{c,i}||_F \leq ||R_i - R_b R_{c,i}||_F \text{ for all } R_b \in \mathcal{S}_R(\mathcal{L})\} \right|. \tag{4.6}$$

Any orientations not part of the maximizer set are considered to be wrong.

In the following subsections we tabulate the number of detwinned orientations that were accepted, rejected, correct, and wrong.

4.6.2 Test Case 1

J_o	Accepted	Rejected	Correct	Wrong
21800	22645	9360	22629	16
10900	21355	9603	21340	15
4360	19420	9607	19400	20
2180	17844	9717	17829	15
1090	16274	9693	16258	16
218	11663	9757	11641	22
43.6	6109	8914	6081	28
21.8	3775	7409	3750	25
10.9	1774	5368	1746	28
4.36	273	2842	143	130
2.18	17	1969	10	7

Table 4.7: Orientation detwinning performance for test case 1.

4.6.3 Test Case 2

J_o	Accepted	Rejected	Correct	Wrong
21800	17607	3316	17554	53
10900	15399	3794	15323	76
4360	12292	4156	12227	65
2180	9976	4230	9915	61
1090	7581	4145	7542	39
218	1178	5106	1133	45
43.6	20	2427	15	5
21.8	0	1537	0	0
10.9	0	874	0	0
4.36	0	357	0	0
2.18	0	152	0	0

Table 4.8: Orientation detwinning performance for test case 2.

4.6.4 Test Case 3

J_o	Accepted	Rejected	Correct	Wrong
21800	7438	13005	7384	54
10900	10178	9404	10082	96
4360	5441	12755	5358	83
2180	881	15983	841	40
1090	435	14765	237	198
218	316	10371	107	209

Table 4.9: Orientation detwinning performance for test case 3.

4.7 Reconstructions

4.7.1 Test Description

Here we test the feasibility of using data processed by our autoindexing, crystal size determination, structure factor magnitude modeling, and orientation detwinning algorithms to perform three-dimensional reconstructions of the molecular structure within a unit cell using the computational phase retrieval strategy discussed in Section 3.6. For every sample point in Ω, described in Algorithm 9, whose value was measured with at least 20 images, we compute structure factor magnitude values via Equation (3.44) with the neighborhood radius r set to 5% of the smallest reciprocal Bravais lattice vector length. We then search for the electron density with the version of the Shrinkwrap algorithm described in Algorithm 9. For test case 3, we also enforced the known set of reflection conditions for $P2_12_12_1$, i.e., that the structure factors are exactly 0 at Miller indices of the form $(2n+1, 0, 0)$, $(0, 2n+1, 0)$, and $(0, 0, 2n+1)$, where $n \in \mathbb{Z}$.

We use the following set of parameters for Algorithm 1. We initialize the phases to 0 and reduce the cutoff τ and Gaussian width σ after every 10,000 HIO/ER iterations, up to three times. In more detail, we use HIO until it has failed to decrease the total error $\varepsilon_S^2(\rho^{(n)}) + \varepsilon_M^2(\rho^{(n)})$ after 200 iterations or until a support refinement step, and in both cases we switch over to ER for 200 iterations. In test case 1, for $J_o \geq 43.6$, we initially set $\tau = .2$ and $\sigma = 1.1$ pixels and decrease σ by steps of .1 and for $J_o < 43.6$, we initially set $\tau = .25$ and $\sigma = 1.8$ pixels and decrease σ by steps of .1. In test case 2 we initially set $\tau = .15$ and $\sigma = 2.5$ pixels and decrease τ by steps of .03 and σ by steps of .4. In test case 3, for $J_o \geq 4360$, we initially set $\tau = .15$ and $\sigma = 2.975$ pixels and decrease τ by steps of .03 and σ by steps of .4 and for $J_o < 21.8$, we initially set $\tau = .18$ and $\sigma = 2.975$ pixels and decrease σ by steps of .1. In particular, one should choose these parameters by studying the supports that they induce, i.e., by choosing those which do not lead to the support completely vanishing

or growing too far outside of the unit cell. In general, convergence was reached in between 60,000-100,000 iterations.

In the following subsections, we present contours of the reconstructed electron densities, for cases where at least 1,000 images were oriented, along with the corresponding exact solution, which is obtained by taking the inverse Fourier transform of the exact structure factors at the same resolution measured by the diffraction images.

4.7.2 Test Case 1

Figure 4.96: Electron density contour of the exact solution for test case 1, displayed at two different orientations.

Figure 4.97: Electron density contour of the computed reconstruction for test case 1 with $J_o = 21800$, displayed at two different orientations.

Figure 4.98: Electron density contour of the computed reconstruction for test case 1 with $J_o = 10900$, displayed at two different orientations.

Figure 4.99: Electron density contour of the computed reconstruction for test case 1 with $J_o = 4360$, displayed at two different orientations.

Figure 4.100: Electron density contour of the computed reconstruction for test case 1 with $J_o = 2180$, displayed at two different orientations.

Figure 4.101: Electron density contour of the computed reconstruction for test case 1 with $J_o = 1090$, displayed at two different orientations.

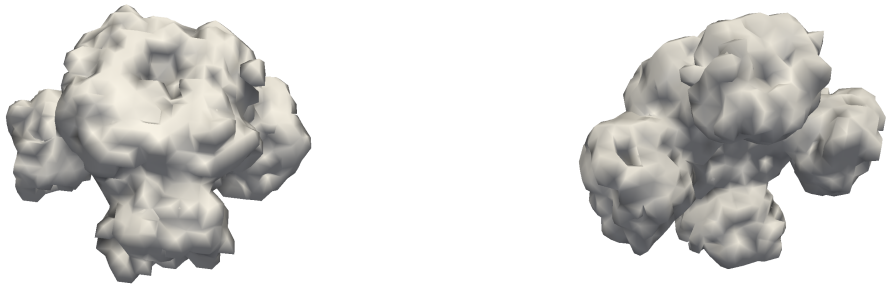


Figure 4.102: Electron density contour of the computed reconstruction for test case 1 with $J_o = 218$, displayed at two different orientations.

Figure 4.103: Electron density contour of the computed reconstruction for test case 1 with $J_o = 43.6$, displayed at two different orientations.

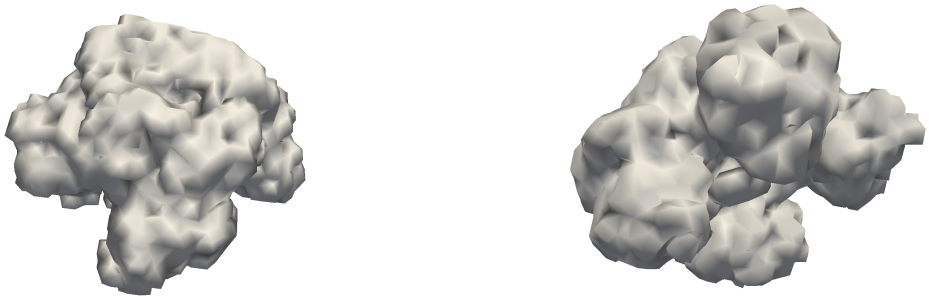


Figure 4.104: Electron density contour of the computed reconstruction for test case 1 with $J_o = 21.8$, displayed at two different orientations.

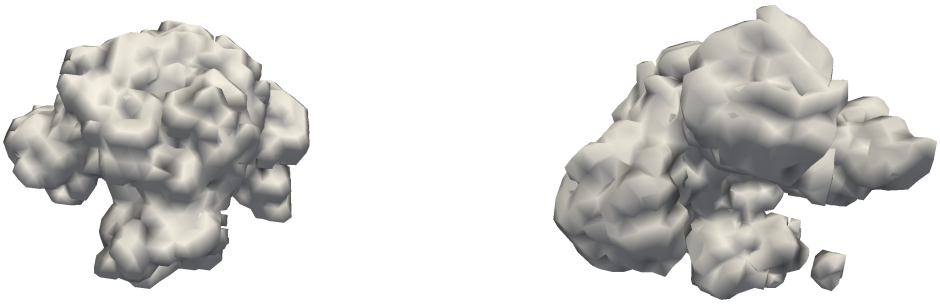


Figure 4.105: Electron density contour of the computed reconstruction for test case 1 with $J_o = 10.9$, displayed at two different orientations.

4.7.3 Test Case 2

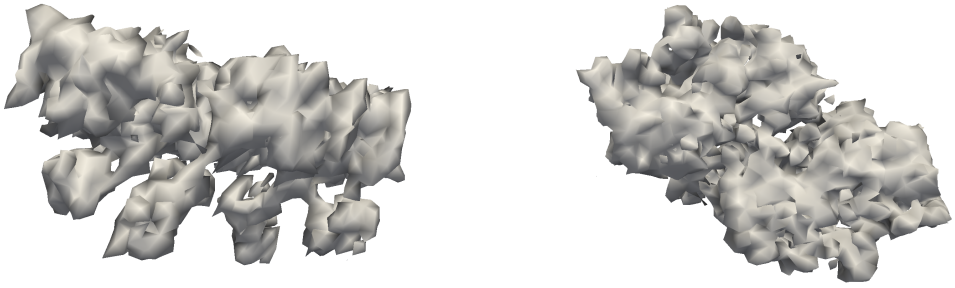


Figure 4.106: Electron density contour of the exact solution for test case 2, displayed at two different orientations.

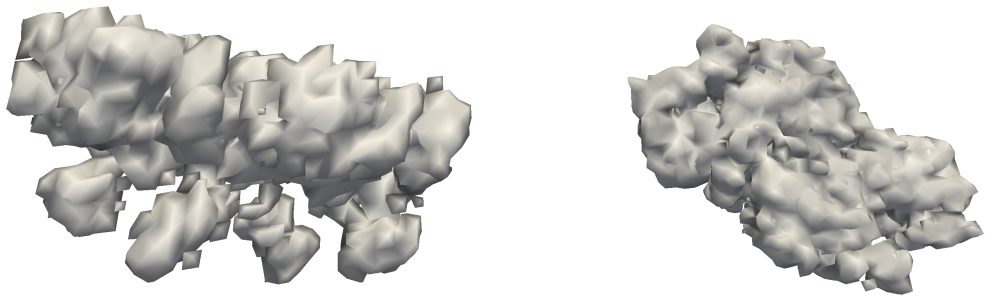


Figure 4.107: Electron density contour of the computed reconstruction for test case 2 with $J_o = 21800$, displayed at two different orientations.

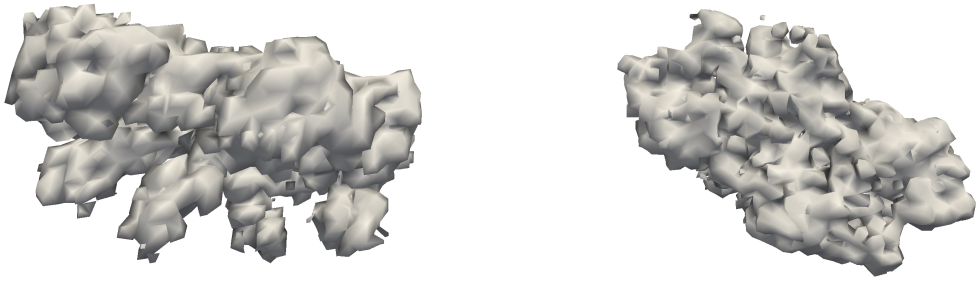


Figure 4.108: Electron density contour of the computed reconstruction for test case 2 with $J_o = 10900$, displayed at two different orientations.

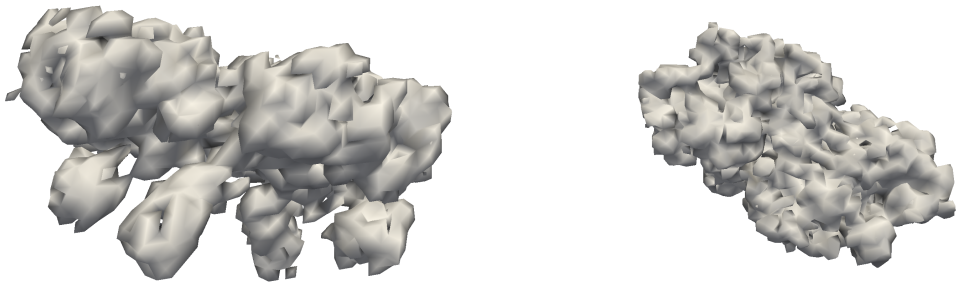


Figure 4.109: Electron density contour of the computed reconstruction for test case 2 with $J_o = 4360$, displayed at two different orientations.

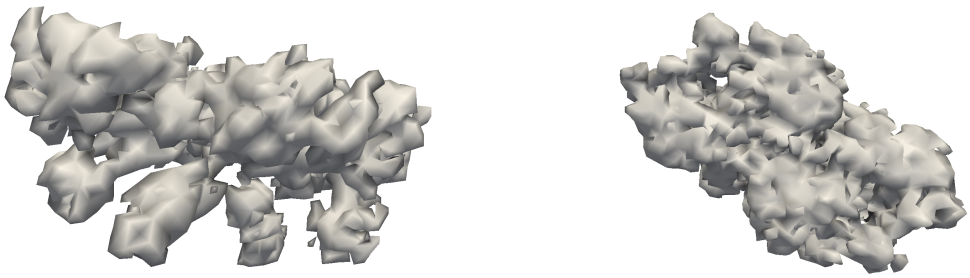


Figure 4.110: Electron density contour of the computed reconstruction for test case 2 with $J_o = 2180$, displayed at two different orientations.

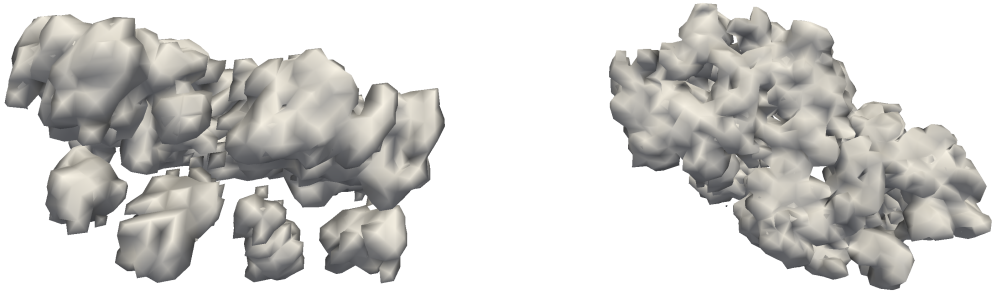


Figure 4.111: Electron density contour of the computed reconstruction for test case 2 with $J_o = 1090$, displayed at two different orientations.

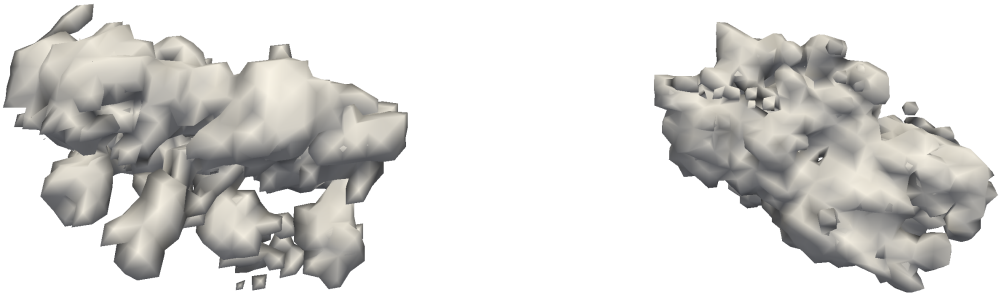


Figure 4.112: Electron density contour of the computed reconstruction for test case 2 with $J_o = 218$, displayed at two different orientations.

4.7.4 Test Case 3

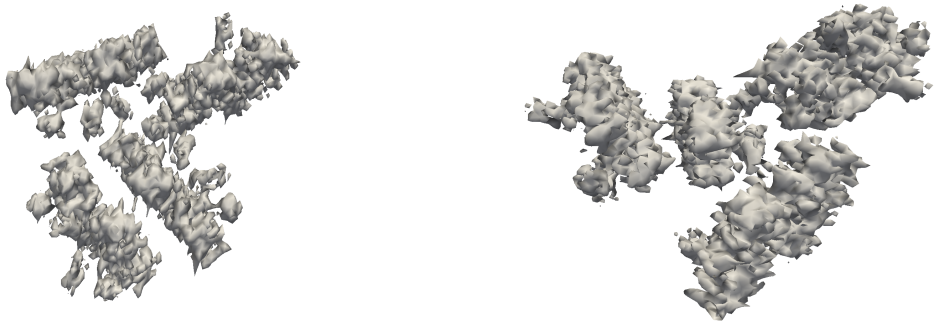


Figure 4.113: Electron density contour of the exact solution for test case 3, displayed at two different orientations.

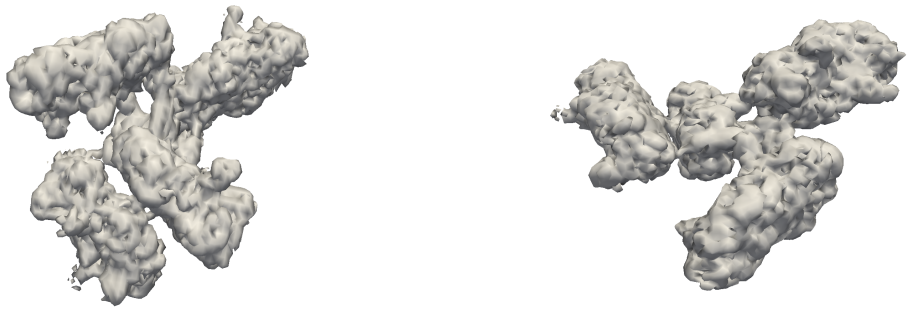


Figure 4.114: Electron density contour of the computed reconstruction for test case 3 with $J_o = 21800$, displayed at two different orientations.

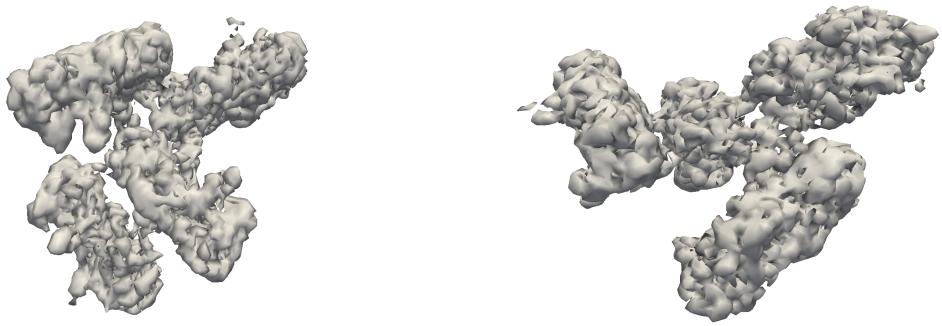


Figure 4.115: Electron density contour of the computed reconstruction for test case 3 with $J_o = 10900$, displayed at two different orientations.

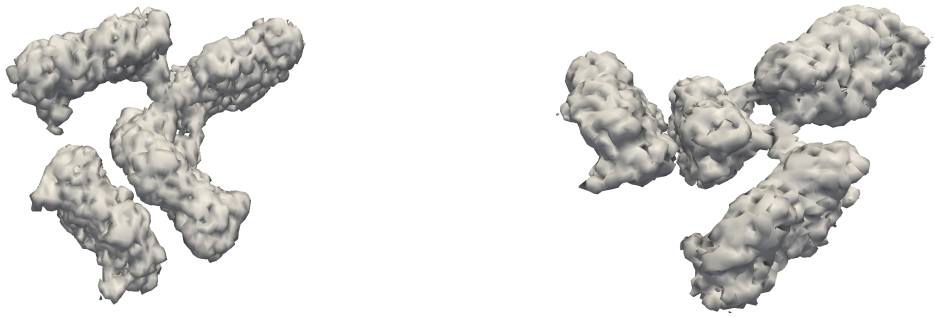


Figure 4.116: Electron density contour of the computed reconstruction for test case 3 with $J_o = 4360$, displayed at two different orientations.

4.8 Summary

The results presented in the previous sections demonstrate that our algorithmic framework has the capability to effectively solve the twinning problem and determine molecular structure with current experimental parameters and noise levels seen in experiments.

Our autoindexing technique was able to accurately determine the orientations of the images, up to lattice symmetry, with a typical accuracy of about 99.9%, for a large range of incident photon flux densities. Its performance only starts to degrade when an insufficient number of peaks are observed in the images.

The majority of the accepted crystal sizes were able to be determined to within about 10-20% relative error. Sizes were typically rejected when the low angle images did not pass

close enough to a reciprocal lattice point. In principle, the rejection rate can be lowered by either using a larger detector or by placing it closer to the interaction point, which would allow detection of more Bragg peaks at high resolution. As was shown, the accuracy of the calculated crystal sizes were sufficient to detwin the data and perform reconstruction.

Note that we were able to successfully scale and find the peaks in the structure factor magnitude modeling step for all of the signal strengths that we tested. However, in order to solve the twinning problem, one requires these models at a sufficient number of reciprocal points. Consequently, failure typically occurs when an insufficient amount of signal is collected from the high-angle Bragg peaks, in which case we may no longer have enough information to form these multi-modal models at higher frequency values. In theory, it may be possible to lower the minimum incident photon flux density required for this approach to work by using a larger number of images. Fortunately, when successful, the detwinning algorithm is shown to be close to 99% accurate, and when it fails, instead of returning incorrect results, it tends to reject most of the orientations, allowing one to determine if the method was successful.

Test case 3 demonstrates that with a sufficient amount signal, well within proposed levels, one may solve the twinning problem beyond Laue symmetry and up to symmetry of the non-Bragg data. In particular, this allows for the application of computational phase retrieval techniques, which typically require detwinned non-Bragg data, to these situations. The slight drop in performance for $J_o = 21800$ Å^{-2} is likely caused by the inclusion of data where the shape function is close to zero, now measurable from the larger signal, which makes the calculation in Equation 3.30 more unstable. This could potentially be fixed by using a more robust cutoff for filtering out weak signals.

We have also demonstrated that the orientation and crystal size calculation is precise enough to allow reconstruction with relatively few images. In particular, our approach converges much more quickly than the Monte Carlo approach, developed in [74], even when utilizing non-Bragg reflections, since we remove most of the variation in our data. Moreover, the reconstructions for test cases 1 and 3 show that the sampling strategy in Section 3.6 coupled with compressive phase retrieval techniques, make it feasible to computationally solve the phase problem for x-ray nanocrystallography using only magnitude information.

Chapter 5

Conclusion

We have presented a new algorithmic framework for reconstructing molecular structure *ab initio* from an ensemble of x-ray nanocrystallography diffraction images in the presence of noise, large variations in the incident photon flux densities, and the twinning problem. In particular, our approach is based on accurately determining orientation and crystal size information and, thus, allows for reconstruction with fewer images than the Monte Carlo approach, developed in [74]. Furthermore, we have shown that this framework is able to successively solve the twinning problem, even beyond Laue symmetry, and determine structure with the parameters and noise levels seen in current experiments.

This framework may be applied to enhance current reconstruction techniques and may allow for the use of additional methods that were previously infeasible in the presence of the twinning problem. For instance, by using this approach, molecular replacement techniques will be able to test models against the full detwinned diffraction information. Additionally, this framework also allows for the application of techniques that do not require the knowledge of an existing similar structure, such as anomalous dispersion and isomorphous replacement, which typically need complete orientation information. Furthermore, we have demonstrated that nanocrystallographic diffraction images provide sufficient oversampling of the electron density power spectrum to solve the phase problem using only Fourier magnitude information, via a compressive phase retrieval algorithm. This computational phasing offers the possibility to greatly simplify the experimental setup for imaging large molecules, as it does not require an already existing model, beam wavelengths near absorption edges, nor the creation of additional crystals loaded with heavy atoms.

This approach could be further enhanced with the development of post-refinement techniques. In particular, once a sufficient number of images are fully oriented, one may be able refine the orientation, crystal size, and scaling information of the rejected images, which may then be included in the final structure factor magnitude calculation. This may even allow one to repeat the steps in the reconstruction framework several times as an iterative

refinement.

Additionally, while we have shown that sampling at and halfway between reciprocal lattice points is sufficient to computationally solve the phase retrieval problem, one may try to enhance this process by incorporating more samples on the line between adjacent lattice points, which also tend to have a noticeable signal in the diffraction images. In particular, this may help in the case when the molecular structure takes up a large percentage of the unit cell, which, otherwise, reduces the effectiveness of seeking a solution with minimal support.

References

[1] *Parametrization of f0 (the non-dispersive part of the atomic scattering factor) vs sin(theta/lambda)*, 2002, URL `http://ftp.esrf.eu/pub/scisoft/xop2.3/DabaxFiles/f0_InterTables.dat`.

[2] J. P. Abrahams and A. G. W. Leslie, *Methods used in the structure determination of bovine mitochondrial f1 atpase*, Acta Cryst. **D52** (1996), 30–42.

[3] B. Alidaee et al., *Solving the maximum edge weight clique problem via unconstrained quadratic programming*, European Journal of Operational Research **181** (2007), 592–597.

[4] A. Aquila et al., *Time-resolved protein nanocrystallography using an x-ray free-electron laser*, Optics Express **20** (2012), no. 3, 2706–2716.

[5] R. Barakat and G. Newsam, *Necessary conditions for a unique solution to two-dimensional phase recovery*, J. Math. Phys. **25** (1984), no. 11, 3190–3193.

[6] A. Barty, C. Caleman, and H. N. Chapman, *Self-terminating diffraction gates femtosecond x-ray nanocrystallography measurements*, Nature Photonics **6** (2011), 35–40.

[7] H. H. Bauschke, P. L. Combettes, and D. R. Luke, *Phase retrieval, error reduction algorithm, and fienup variants: a view from convex optimization*, J. Opt. Soc. Am. A. **19** (2002), no. 7, 1334–1345.

[8] H. H. Bauschke, P. L. Combettes, and D. R. Luke, *Hybrid projection-reflection method for phase retrieval*, J. Opt. Soc. Am. A. **20** (2003), no. 6, 1025–1034.

[9] F. C. Bernstein et al., *The protein data bank: A computer-based archival file for macromolecular structures*, J. of. Mol. Biol. **112** (1977), 5.5.

[10] D. M. Blow and F. Crick, *The treatment of errors in the isomorphous replacement method*, Acta Cryst. **12** (1959), 794–802.

[11] D. M. Blow and M. G. Rossmann, *The single isomorphous replacement method*, Acta Cryst. **14** (1961), 1195–1202.

[12] S. Boutet et al., *High-resolution protein structure determination by serial femtosecond crystallography*, Science **337** (2012), 362–364.

[13] J. W. Campbell, *The practicality of using a three-dimensional Fourier transform in auto-indexing protein single-crystal oscillation images*, J. Appl. Cryst. **31** (1998), 407–413.

[14] E. J. Candes, Y. Eldar, T. Strohmer, and V. Voroninski, *Phase retrieval via matrix completion*, (pre-print), 2011, arXiv:1109.0573 [cs.IT].

[15] E. J. Candes, J. K. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory **52** (2006), no. 2, 489–509.

[16] H. N. Chapman, P. Fromme, A. Barty, T. A. White, and R. A. Kirian, *Femtosecond x-ray protein nanocrystallography*, Nature **470** (2011), 73–77.

[17] F. H. C. Crick and B. S. Magdoff, *The theory of the method of isomorphous replacement for protein crystals. I*, Acta Cryst. **9** (1956), 901–907.

[18] R. A. Crowther, *The molecular replacement method*, ch. The Fast Rotation Function, Gordon & Breach, 1972.

[19] R. A. Crowther and D. W. Blow, *A method of positioning a known molecule in an unknown crystal structure*, Acta Cryst. **23** (1967), 544–548.

[20] A. J. M. Duisenberg, *Indexing in single-crystal diffractometry with an obstinate list of reflections*, J. Appl. Cryst. **25** (1992), 92–96.

[21] V. Elser, *Phase retrieval by iterated projections*, J. Opt. Soc. Am. A. **20** (2003), no. 1, 40–55.

[22] J. R. Fienup, *Reconstruction of an object from the modulus of its Fourier transform*, Optics Letters **3** (1978), no. 1, 27–29.

[23] J. R. Fienup, *Phase retrieval algorithms: a comparison*, Applied Optics **21** (1982), no. 15, 2758–2769.

[24] J. R. Fienup and C.C. Wackerman, *Phase-retrieval stagnation problems and solutions*, J. Opt. Soc. Am. A. **3** (1986), no. 11, 1897–1907.

[25] J. N. Franklin, *Ambiguities in the x-ray analysis of crystal structures*, Acta Cryst. **A 30** (1974), 698–702.

[26] R.W. Gerchberg and W. O. Saxton, *A practical algorithm for the determination of the phase from image and diffraction plane pictures*, Optik **35** (1972), 237–246.

[27] D. W. Green, V. M. Ingram, and M. F. Perutz, *The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method*, Proc. Roy. Soc. A. **225** (1954), no. 1162, 287–307.

[28] Y. Guan, *Variance stabilizing transformations of poisson, binomial and negative binomial distributions*, Stat. Probabil. Lett. **14** (2009), 1621–1629.

[29] A. Guinier, *X-ray diffraction in crystals, imperfect crystals and amorphous bodies*, Dover, 1994.

[30] T. Hahn (ed.), *International tables for crystallography, volume A: Space group symmetry*, 5th ed., vol. A, Springer-Verlag, Berlin, New York, 2002.

[31] H. Hauptman and J. Karle, *Solution of the phase problem I. The centrosymmetric crystal*, no. 3, American Crystallographic Association, 1953.

[32] M. H. Hayes, *The reconstruction of a multidimensional sequence from the phase or mignitude of its Fourier transform*, IEEE Transactions on Acoustics Speech and Signal Processing **30** (1982), no. 2, 140–154.

[33] M. H. Hayes and J. H. McClellan, *Reducible polynomials in more than one variable*, Proceedings of the IEEE, vol. 70, 1982.

[34] W. Hendrickson, *Analysis of protein structure from diffraction measurement at multiple wavelengths*, Trans ACA **21** (1985), 11–21.

[35] W. Hendrickson and C. Ogata, *Phase determination from multiwavelength anomalous diffraction measurements*, Meth Enzymol **276** (1997), 494–523.

[36] W. Hendrickson, J. Smith, and S. Sheriff, *Direct phase determination based on anomalous scattering*, Meth Enzymol **115** (1985), 41–55.

[37] B. L. Henke, E. M. Gullikson, and J. C. Davis, *X-ray interactions: photoabsorption, scattering, transmission, and reflection at e=50-30000 ev, z=1-92*, Atomic Data and Nuclear Data Tables **54** (1993), no. 2, 181–342.

[38] A. M. J. Huiser and P. van Toorn, *Ambiguity of the phase-reconstruction problem*, Optics Letters **5** (1980), no. 11, 499–501.

[39] L. C. Johansson et al., *Lipidic phase membrane protein serial femtosecond crystallography*, Nat. Meth. **9** (2012), no. 3, 263–265.

[40] J. Karle, *Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology*, International Journal of Quantum Chemistry: Quantum Biology Symposium **7** (1980), 357–367.

[41] J. Karle, *Linear algebraic analyses of structures with one predominant type of anomalous scatterer*, Acta Cryst. **A45** (1989), 303–307.

[42] G. Kartha and R. Parthasarathy, *Combination of multiple isomorphous replacement and anomalous dispersion data for protein structure determination I. Determination of heavy-atom positions in protein derivatives*, Acta Cryst. **18** (1965), 745–749.

[43] P. Kiedron, *On the 2-d solution ambiguity of the phase recovery problem*, Optik **59** (1981), 303–309.

[44] R. Koopmann et al., *In vivo protein crystallization opens new routes in structural biology*, Nat. Meth. **9** (2012), no. 3, 259–262.

[45] L. Lovisolo and E. A. B. da Silva, *Uniform distribution of points on a hyper-sphere with applications to vector bit-plane encoding*, Vision, Image and Signal Processing, IEEE Proceedings, vol. 148, 2001.

[46] D. R. Luke, *Relaxed averaged alternating reflections for diffraction imaging*, Inverse Problems **21** (2005), 37–50.

[47] F. Maia, C. Yang, and S. Marchesini, *Compressive auto-indexing in femtosecond nanocrystallography*, Ultramicroscopy **111** (2011), no. 7, 807–811.

[48] S. Marchesini, *Phase retrieval and saddle-point optimization*, J. Opt. Soc. Am. A. **24** (2007), no. 10, 3289–3296.

[49] S. Marchesini, *A unified evaluation of iterative projection algorithms for phase retrieval*, Review of Scientific Instruments **78** (2007), 011301.

[50] S. Marchesini, *Ab initio compressive phase retrieval*, Microscopy and Microanalysis **15** (2009), 742–743.

[51] S. Marchesini, H. He, H. N. Chapman. P. Hau-Riege, A. Noy, M. R. Howells, U. Weierstall, and J. C. Spence, *X-ray image reconstruction from a diffraction pattern alone*, Physical Review B **68** (2003), no. 4, 140101.

[52] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, *Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens*, Nature **400** (1999), 342–344.

[53] R. P. Millane, *Multidimensional phase problems*, J. Opt. Soc. Am. A. **13** (1996), 725–734.

[54] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, *Compressive phase retrieval*, Proc. SPIE 6701, Wavelets XII, vol. 6701, 2007.

[55] Z. Otwinowski and W. Minor, *Processing of x-ray diffraction data collected in oscillation mode*, Macromolecular Crystallography Part A **276** (1997), 307–326.

[56] A. L. Patterson, *A Fourier series method for the determination of the components of interatomic distances in crystals*, Phys. Rev. **46** (1934), 372–376.

[57] A. L. Patterson, *Ambiguities in the x-ray analysis of crystal structures*, Phys. Rev. **65** (1943), no. 5-6, 195–201.

[58] L. Pauling and M. D. Shappell, *The crystal structure of bixbyite and the c-modification of the sesquioxides*, Z. Kristallogr **75** (1930), no. 1-2, 128–142.

[59] M. F. Perutz, *Isomorphous replacement and phase determination in non-centrosymmetric space groups*, Acta Cryst. **9** (1956), 867–873.

[60] G. N. Ramachandran and S. Raman, *A new method for the structure analyisis of noncentrosymmetric crystals*, Curr. Sci. **25** (1956), 348–351.

[61] I. Ramazzina, L. Cendron, C. Folli, R. Berni, D. Monteverdi, G. Zanotti, and R. Percudani, *Logical identification of an allantoinase analog (puuE) recruited from polysaccharide deacetylases*, J. Biol. Chem. **283** (2008), no. 34, 23295–23304.

[62] J. Rosenblatt, *Phase retrieval*, Commun. Math. Phys. **95** (1984), 317–343.

[63] J. Rosenblatt and Paul D. Seymour, *The structure of homometric sets*, SIAM J. Alg. Disc. Meth. **3** (1982), no. 3, 343–350.

[64] M. G. Rossmann, *The molecular replacement method*, Gordon & Breach, New York, 1972.

[65] M. G. Rossmann and D. M. Blow, *The detection of sub-units within the crystallographic asymmetric unit*, Acta Cryst. **15** (1962), 24–31.

[66] J. L. C. Sanz, T. S. Huang, and F. Cukierman, *Stability of unique fourier-transform phase reconstruction*, JOSA **73** (1983), no. 11, 1142–1145.

[67] J. C. H. Spence, R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, T. White, A. Barty, H. N. Chapman, S. Marchesini, and J. Holton, *Phasing of coherent femtosecond x-raydiffraction from size-varying nanocrystals*, Optics Express **19** (2011), no. 4, 2866–2873.

[68] J. C. H. Spence, U. Weierstall, and H. N. Chapman, *X-ray lasers for structural and dynamic biology*, Rep. Prog. Phys. **75** (2012), 102601.

[69] I. Steller, R. Bolotovsky, and M. G. Rossmann, *An algorithm for automatic indexing of oscillation images using Fourier analysis*, J. Appl. Cryst. **30** (1997), 1036–1040.

[70] I. Waldspurger, A. d'Aspremont, and S. Mallat, *Phase recovery, maxcut and complex semidefinite programming*, (pre-print), 2012, arXiv:1206.0102 [math.OC].

[71] B. C. Wang, *Resolution of phase ambiguity in macromolecular crystallography*, Meth Enzymol **115** (1985), 90–111.

[72] B. E. Warren, *X-ray diffraction*, Dover, 1990.

[73] Z. Wen, C. Yang, X. Liu, and S. Marchesini, *Alternating direction methods for classical and ptychographic phase retrieval*, Inverse Problems **28** (2012), no. 11, 115010.

[74] T. A. White, R. A. Kirian, A. V. Martin, A. Aquila, K. Nass, A. Barty, and H. N. Chapman, *Crystfel: a software suite for snapshot serial crystallography*, Journal of Applied Crystallography **45** (2012), no. 2, 335–341.

[75] A. J. C. Wilson (ed.), *Tables for x-ray crystallography*, vol. C, Kluwer Academic Publishing, 1995.

[76] X. F. Zheng, C. D. Zheng, Y. X. Gu, Y. D. Mo, M. F. Fan, and Q. Hao, *Use of single isomorphous replacement data of proteins-resolving the phase ambiguity and a new procedure for phase extension*, Acta. Cryst. **D53** (1997), 49–55.

[77] A. Zouni, H. T. Witt, J. Kern, P. Fromme, N. Krauss, W. Saenger, and P. Orth, *Crystal structure of photosystem II from synechococcus elongatus at 3.8 A resolution*, Nature **409** (2001), no. 739, 11217865.