Aspects of Phase Retrieval in X-ray Crystallography

Romain Arnal

A thesis presented for the degree of
Doctor of Philosophy

in
Electrical and Computer Engineering
at the University of Canterbury,
Christchurch, New Zealand.

June 2019

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 2 describes work by the candidate that is reported in "Millane and Arnal, Uniqueness of the macromolecular crystallographic phase problem, Acta Cryst., A71, 592-598, 2015."

Please detail the nature and extent (%) of contribution by the candidate:

This publication describes joint work by the candidate and Millane. The theory was derived jointly by the candidate and Millane. The simulations, the required software development, and presentation of the results were conducted by the candidate. Writing was a joint effort.

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: Rick Millane Signature:  Date: *31 May 2019*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 3 is a reproduction of the paper "Arnal and Millane, The phase problem for two-dimensional crystals. I. Theory, Acta Cryst., A73, 438-448, 2017." Permission to reproduce this paper in the thesis has been obtained from the publisher IUCr.

Please detail the nature and extent (%) of contribution by the candidate:

The candidate was the lead author of this publication. The candidate developed key aspects of the theory with assistance from Millane. Design of the simulations, algorithm and software development, experimental design, and development and presentation of the results were conducted by the candidate. The candidate wrote most of the paper with assistance from Millane.

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Rick Millane* R. P. Millane Signature: Date: *31 May 2019*

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 4 is a reproduction of the paper "Arnal, Zhao, Mitra, Spence and Millane, The phase problem for two-dimensional crystals. II. Simulations, Acta Cryst., A74, 537-544, 2018." Permission to reproduce this paper in the thesis has been obtained from the publisher IUCr.

Please detail the nature and extent (%) of contribution by the candidate:

The candidate was the lead author of this publication. The candidate developed the algorithms and software described in the paper, and used these to generate, interpret and present the results described. The candidate wrote most of the paper with assistance from Millane. Mitra provided data for the simulations. Spence provided some general guidance for this study. Zhao was Spence's student and provided little input to this work.

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Rick Millane* Signature:  Date: *31 May 2019*

ACKNOWLEDGEMENTS

First and foremost, I must express my sincere gratitude to my supervisor Prof. Rick Millane. For all his support and guidance over the last few years I am especially indebted to him.

During the course of my research I had the extraordinary chance to take part in many experiments at the Linac Coherent Light Source (LCLS) at Stanford. I am very grateful to principal investigators Henry Chapman, Carolin Seuring, Richard Bean and Lourdu Paulraj, for giving me the opportunity to participate in and learn from these experiments. I express my thanks to all the people involved in experiments cxils2616, cxilr7416, mfxn1116, amok8916 and amok9916, especially Steve Aplin and David Wojtas. Thanks again to Henry Chapman, for the invitations to visit and work at the Center for Free-electron Laser Science (CFEL) in Hamburg, and to the various people I met during my trips there.

I also owe special thanks to my friend Joe Chen, for introducing me to this field and for the many discussions, valuable insights and for being an outstanding teacher. I am also thankful to Rick Kirian, for his invitations to visit and work at Arizona State University, and the help provided by many people there.

Finally I would like to thank Andy Tan for always being there for me, Dr. Steve Weddell for inviting me to New Zealand back in 2013, and my family in France for allowing me travel across the world without seeing them for years.

Contents

Preface

X-ray crystallography is the primary technique for imaging the structures, or the positions of the atoms, of molecules. Knowledge of the geometrical atomic structures of molecules is key information in physics, chemistry, biology, geology and many other areas of science and technology. Structures are determinants of the properties of molecular systems. In the case of biology, knowledge of the structures of biological molecules provides essential information that allows us to understand the biological functionality of biomolecules and biomolecular systems. This knowledge is used to understand the fundamental molecular basis of biological function and processes, disease processes, and is also important in rational, or structure-based, drug design.

X-ray crystallography involves irradiating a crystal specimen of the molecule under study with a beam of X-rays, and measuring the resulting pattern of diffracted X-rays. The data consisting of measured diffraction patterns is then inverted computationally to produce an image of the molecule. This is often referred to as computational imaging or computational microscopy. If both the phase and amplitude of the diffracted X-ray could be measured, then inversion of the data to produce the image would be straightforward. However, in practice, one can measure only the amplitude, but not the phase, of the diffracted X-rays. This results in the famous so-called "phase problem" in crystallography. A method of determining the phases must be devised before the structure can be calculated.

The phase problem in crystallography has been studied for over one hundred years, and a number of clever methods have been devised for determining the phases in order for structures to be calculated. However, the phase problem is still an active area of research as current phasing techniques have significant limitations, and also because of the emergence of new kinds of instrumentation, specimens, and diffraction experiments.

This thesis is concerned with the phase problem and phase retrieval algorithms for biological (macromolecular) crystallography that have arisen, in part, through the recent introduction of a new kind of X-ray source called an X-ray free-electron laser, and through new kinds of specimens that can be used with these sources.

The thesis is divided into six chapters. The first chapter provides background information on diffraction imaging, X-ray crystallography, the phase problem, phase retrieval algorithms, and X-ray free-electron lasers and serial femtosecond crystallogra-

phy. Original material is contained in Chapters 2 through 5. Concluding remarks are made in Chapter 6.

Chapter 2 is concerned with properties of the phase problem for 3D crystals. New relationships are derived that more carefully formalise uniqueness for this problem, the problem for the case of an unknown molecular support is studied in detail and the theoretical results are supported by simulations, and the effects of crystallographic and noncrystallographic symmetry are elucidated.

Chapters 3 and 4 form the first main part of the thesis and consider the phase problem for 2D crystals, a new kind of specimen that has been investigated with X-ray free-electron lasers. The two chapters are presented as two published journal papers for which the candidate is the primary author. In Chapter 3, the fundamental uniqueness properties of the phase problem for 2D crystals are derived, the nature of the solution set is elucidated, and the effects of various kinds of *a priori* information are evaluated by simulation. Chapter 4 follows up the results in Chapter 3, using simulations to investigate practical aspects of *ab initio* phase retrieval for 2D crystals using minimal molecular envelope information, and considering the characteristics of data available from X-ray free-electron laser sources.

Chapter 5 forms the second main part of the thesis and develops a new kind of *ab initio* phasing technique called *ab initio* molecular replacement phasing. This method uses diffraction data from the same molecule crystallised in two or more crystal forms. Uniqueness of the solution for such a dataset is evaluated, and a suitable phase retrieval algorithm is developed and tested by simulation using a small protein of known structure.

Chapter 6 contains a brief summary of the outcomes of the thesis and suggestions for future research.

The following publications have been published as part of this research:

Journal papers:

Millane, R. P. and Arnal, R. D. (2015). "Uniqueness of the macromolecular crystallographic phase problem," *Acta Crystallographica A*, vol. 71, pp. 592-598.

Arnal, R. D and Millane, R. P. (2017). "The phase problem for two-dimensional crystals. I. Theory," *Acta Crystallographica A*, vol. 73, pp. 438-448.

Arnal, R. D., Zhao, Y. , Mitra, A. K., Spence, J. C. and Millane, R. P. (2018). "The phase problem for two-dimensional crystals. II. Simulations," *Acta Crystallographica A*, vol. 74, pp. 537-544.

Published conference papers:

Chen, J., Kirian, R., Beyerlein, K., Bean, R., Morgan, A., Yefanovc, O., Arnal, R. D., Wojtas, D., Bones, P., Chapman, H., Spence, J. and Millane, R."Image reconstruction in serial femtosecond nanocrystallography using X-ray free-electron lasers," in Proceedings of SPIE: Image Reconstruction from Incomplete Data VIII, San Diego, CA, USA, Sep. 2015.

Millane, R. P., Arnal, R. D., Chen, J. "Phase retrieval for multiple objects," in Proceedings of SPIE: Image Reconstruction from Incomplete Data VIII, San Diego, CA, USA, Sep. 2015.

Arnal, R. D and Millane, R. P. "Effects of non-uniform sampling on phase retrieval," in *Proceedings of the Electronics New Zealand Conference (ENZCON)*, Wellington, New Zealand, Nov 2016.

Arnal, R. D and Millane, R. P. "The Phase Problem with Structured Sampling," in *Proceedings of the Image and Vision Computing Electronics New Zealand conference (IVCNZ)*, Palmerston North, New Zealand, Nov 2016.

Arnal, R. D. and Millane, R. "Phase retrieval for crystalline specimens," in Proceedings of SPIE: Unconventional and Indirect Imaging, Image Reconstruction, and Wavefront Sensing, San Diego, CA, USA, Sep. 2017.

Arnal, R. D and Millane, R. P. "Phase retrieval for 1D and 2D crystals," *Proceedings of the Electronics New Zealand Conference (ENZCON)*, Christchurch, New Zealand, Dec. 2017.

Wojtas, D., Arnal, R. D., Millane, R. "Molecular imaging with X-ray free-electron lasers," in Proceedings of SPIE: Unconventional and Indirect Imaging, Image Reconstruction, and Wavefront Sensing, San Diego, CA, USA, Sep. 2018.

Conference abstracts:

Millane, R. P., Chen, J., Arnal, R. D., Morgan, A., Bean, R., Beyerlein, K., Yefanovc, O., Chapman, H. and Kirian, R. "Phasing XFEL Diffraction Data from Very Small Crystals," Coherence Conference, Saint-Malo, France, Jun. 2016.

Metz, M., Arnal, R. D., Chapman, H. and Millane, R. P. "Phasing in Protein X-ray

Crystallography using Data from Multiple Unit Cells," Coherence Conference, Saint-Malo, France, Jun. 2016.

Millane, R. P. and Arnal, R. D. (2015). "Prospects for *ab initio* phasing and XFEL imaging of 1D and 2D crystal," BioXFEL Conference, New Orleans, USA, Feb. 2018.

Arnal, R. D., Metz, M., Morgan, A., Chapman, H. and Millane, R. "Molecular-replacement ab initio phasing in protein crystallography," BioXFEL Conference, San Diego, USA, Feb. 2019.

1 INTRODUCTION

1.1 DIFFRACTION IMAGING

In imaging, radiation from an object of interest is measured and processed to obtain an image of the object. In some cases, for instance astronomy, the object is the source of the radiation, but usually, a separate and controlled source of radiation is used to illuminate the object. In the course of passing through the object a portion of the radiation interacts with the matter and, while doing so, encodes information on the object composition and structure. Typically, the encoding of the outgoing radiation is decoded with lenses specific to that radiation type, to directly form an image of the object. However, if lenses are not available, lensless techniques are used where the outgoing radiation is recorded on a detector to be later decoded using numerical computations.

In diffraction imaging, measurements of the diffraction from an object are used to obtain a high resolution image of the object. In the past century, complementary diffraction imaging techniques have been developed around different types of radiation such as electromagnetic waves (e.g. light, microwaves, soft and hard X-rays) and matter waves (e.g. electrons, neutrons). Short wavelength radiation allows imaging of matter down to the atomic scale i.e. about one Ångström, with $1\,\text{Å} = 10^{-10}\,\text{m}$, while the penetrating power of the radiation allows imaging of the full three-dimensional structure.

Three-dimensional high resolution imaging enabled by diffraction imaging techniques is a powerful tool that is used in many fields such as chemistry, structural biology, and material science, to name but a few. These techniques enable life to be imaged down to its smallest actors, help develop new advanced materials, design better drugs with better specificity, and study electronic transfers in chemical reactions for the development of more efficient catalysts or even bio-mimicked solar energy generation systems [Hasnain, 2015, Blundell and Patel, 2004, Verschueren et al., 1993].

Until recently, diffraction from small non-crystalline objects was too weak to be measured. Natural gratings, such as molecular crystals and crystalline solids, offer an attractive solution to increase the strength and signal-to-noise ratio (SNR) of the diffraction patterns, as shown in Section 1.2.2. This crystal requirement is one of the largest shortcomings of conventional diffraction imaging techniques, as not all matter

readily crystallises. On the bright side, technological advances in radiation sources and facilities coupled with new methods has rendered crystallisation the most difficult step in diffraction imaging, such that if a crystal of the object can be produced, it is likely that the structure can be determined [Chayen, 2004].

1.2 X-RAY CRYSTALLOGRAPHY

X-ray crystallography (XRC) is currently the most successful diffraction imaging technique for the determination of the structure (the position of atoms in space) of crystalline matter at atomic resolution. This technique spurred the development of the field of structural biology, beginning with the determination of the first protein, the sperm whale myoglobin, by Kendrew in 1958 [Kendrew et al., 1958, Jaskolski et al., 2014]. Currently, more than 90% of all protein structures deposited in the protein databank (PDB) were obtained by XRC. The structure of entire viruses and complex assemblies such as the ribosome (about a quarter of a million atoms) are known to high resolution owing to XRC [Khatter et al., 2015, Prasad et al., 1999].

From the early experiments determining the structure of a copper sulphate crystal using an X-ray tube and photographic plates lead by Friedrich and Knipping in 1912 [Friedrich et al., 1912], to the serial femtosecond X-ray crystallography (SFX) experiments using X-ray free-electron lasers (XFELs) and megahertz CCD detectors [Allahgholi et al., 2015, Henrich et al., 2011], the advancement of this field is both evolutionary and revolutionary. This latter fact is better envisioned when considering the natural advances of XRC where smaller crystals, briefer X-ray pulses and more intense and highly coherent sources have led to the structure determination of key protein structures (of top clinical, technological or environmental significance) and higher resolutions [Johansson et al., 2017, Fromme, 2015, Ishchenko et al., 2018]. At the forefront of XRC, single-particle imaging (SPI), where crystals are no longer needed [Aquila et al., 2015, Oberthür, 2018, Spence and Doak, 2004], has been the main driver and the object of all hopes in the field [Chapman, 2009].

1.2.1 X-ray diffraction crystallography methods

In single crystal X-ray diffraction (SXRD), the oldest and most precise method of XRC, a crystal of the molecule under study is placed in a collimated and monochromatic X-ray beam. The intensities and angles of the diffracted photons are recorded on a detector forming an image called a *diffraction pattern*. A number of diffraction patterns are collected by rotating the crystal about an axis, ideally perpendicular to the X-ray beam, to obtain a three-dimensional dataset. Each diffraction pattern presents a number of spots with varying intensities called *reflections* that encodes for the molecule structure. A reconstituted diffraction pattern is shown in Fig. 1.1. Each reflection corresponds to the constructive interference of scattered X-rays within the crystal that is equivalent to

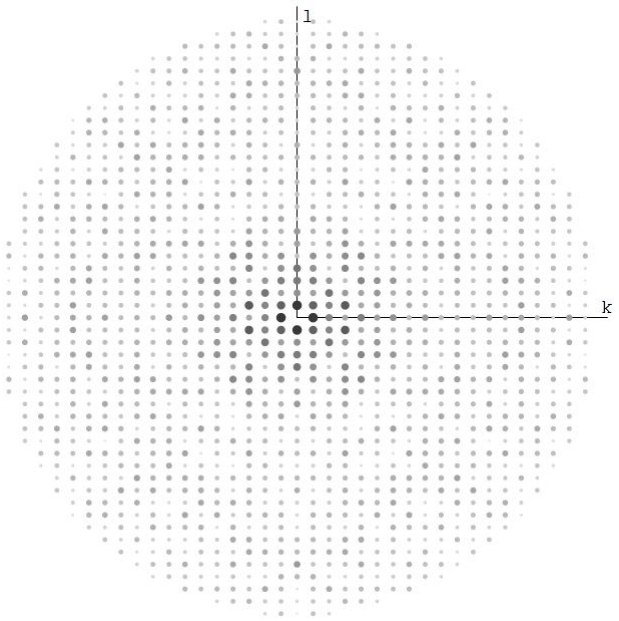


Figure 1.1 Reconstituted diffraction pattern showing spots (reflections) of different intensities [Rypniewski et al., 1993].

the reflection of X-rays on a set of equally spaced crystal planes. Each spot is indexed by three integers (the miller indices) that describe a particular set of planes.

One of the main drawbacks affecting SXRD relates to X-ray induced radiation damage. Amongst the high energy X-ray photons that interact with the crystal, the vast majority are photoabsorbed causing damage to the molecules and to the crystal structure. The radiation dose that a cryo-cooled crystal can accept before half of the diffraction intensity vanishes, compared to that of the undamaged structure, is called the Henderson limit and is about 30 MGy at 100K (where one Gray is one joule per kilogram) [Henderson, 1990]. As the same crystal is exposed multiple times in SXRD, this limits the minimum size of the crystal to about a few microns [Holton and Frankel, 2010, Robin et al., 2016].

The advent of XFEL sources opened a new era in XRC and a major shift of the X-ray imaging paradigm. XFELs produce pulses with a peak brilliance* a billion times higher than the 3$^{\text{rd}}$ generation X-ray synchrotron radiation sources used in SXRD. With such brillances, any crystal lased by a single XFEL pulse undergoes a Coulomb explosion, or vaporization, in a process which starts by the emission of photoelectrons [Spence, 2017, Lomb et al., 2011]. Such destruction of the crystal is not immediate however and noticeable damage will only be manifest after about 30 fs. As an XFEL pulse duration

*Brilliance describes the maximum number of photons of a given energy that are emitted per unit time, unit cross-sectional area, and unit solid angle. Brilliance is usually given in units of photons /s/mm^2/mrad^2, but for comparison purposes between X-ray sources with different spectra, the brilliance is often normalised to 0.1% of the source bandwidth i.e. in units of photons /s/mm^2/mrad^2/0.1%BW.

is in the order of femtoseconds, meaningful diffraction from an undamaged structure can be recorded. These ideas were at the origin of a new paradigm in XRC using XFEL, the so-called *diffract-before-destruct* paradigm [Chapman et al., 2011, Chapman et al., 2014, Neutze et al., 2000].

A new XRC method, based on this paradigm, was developed: serial femtosecond crystallography (SFX). In SFX, in stark contrast to SXRD, for each XFEL pulse a new crystal is presented and destroyed. Up to a few million crystals can be shot in a single SFX experiment, with plans to record diffraction patterns at the full MHz XFEL pulse repetition rate. A number of complimentary specimen delivery approaches were developed for this purpose, including, fixed target systems, liquid jets and gas dynamic virtual nozzles [Weierstall et al., 2014, Hunter et al., 2014, DePonte et al., 2008, Roedig et al., 2015]. Contrary to SXRD, the orientations of the patterns are not known as the crystal orientation at the interaction point is random. These orientations must first be recovered, which can be difficult if the diffraction is weak.

The intensity and position of the reflections in the diffraction pattern encodes the strength and location of the X-ray scattering density in the crystal, namely the *electron density*. Interestingly, the encoding is none other than the squared magnitude of the Fourier transform of the electron density, as outlined in the following section. The Fourier phases which are necessary to inverse Fourier transform the diffraction data to obtain a map of the electron density, are lost in XRC experiments as the detector cannot measure them. The lack of phase measurement is a common problem in lensless imaging and other fields [Millane, 1996, Millane, 1990], and is known colloquially as the *phase problem*. A number of computational techniques to recover the phases in XRC are given in Section 1.3.2. After successful recovery of the phases and, therefore, of the electron density, the three-dimensional structure of the molecule can be determined by fitting a molecular model to the calculated electron density map.

1.2.2 Diffraction by a crystal

The oscillating electric field of X-ray radiation impinging on electrons in the crystal forces them to oscillate. In turn, the oscillating motion of the electron produces dipole radiation which is the basis of scattering. Here, only electron scattering where no energy is imparted to the electron and the wavelength of the scattered X-ray photon is the same as the incident photon is considered. The measured diffraction corresponds to the superposition of all the scattered waves from the electron density in the crystal. In order to simplify the description of the interaction of X-ray photons with the electron density in the crystal, a number of approximations are used in this section:

1. Fraunhofer or far-field approximation: The incident X-rays and the detector plane are essentially at infinity with respect to the crystal (optically speaking). This approximation allows the simplification with plane waves used in Fig. 1.2.

2. Born approximation: The scattering of the X-rays is weak and so occurs at most once within the crystal (multiple scattering does not occur).

A description of diffraction based on these principles is known as the kinematical theory of diffraction and is the main description used in X-ray protein crystallography. If the crystal is perfect, the dynamical theory of diffraction is used where multiple scattering and other phenomenon are taken into account.

Consider diffraction from the electron density cloud of an atom with density denoted $\rho_{\text{atom}}(\mathbf{r})$, with the origin at the center of the atom. The incident X-ray beam is defined by the wave vector $\mathbf{s}_i$, and we consider coherent scattering in the outgoing direction given by the wave vector $\mathbf{s}_o$, both of length $1/\lambda$ (Thompson scattering). The scattered waves from all points $\mathbf{r}$ in the electron density cloud are superimposed in the contribution to the outgoing scattered wave. The amplitude and phase of this scattered wave is thus given by

$$f(\mathbf{s}_i) = \int_{\mathbf{r}} \rho_{\text{atom}}(\mathbf{r}) \exp(2\pi i \mathbf{r} \cdot (\mathbf{s}_o - \mathbf{s}_i)) \mathrm{d}\mathbf{r}. \tag{1.1}$$

Assuming that the electron density cloud is spherically symmetric, the scattering from an atom can be reduced to an atomic scattering factor, denoted $f(|\mathbf{u}|)$ that depends only on the length of the vector $\mathbf{u} = \mathbf{s}_o - \mathbf{s}_i$ and $|\mathbf{u}| = 2\sin(\theta)/\lambda$, with θ the angle shown in Fig. 1.2.

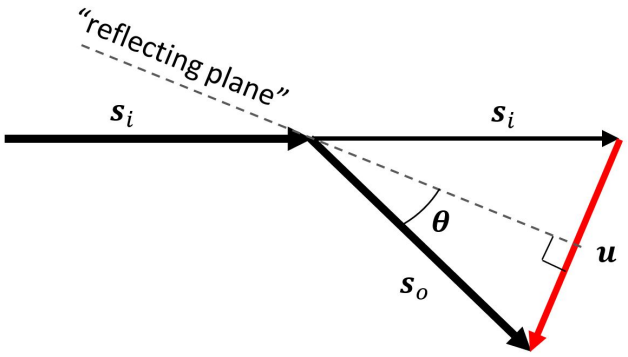


Figure 1.2 The incident wave, denoted $\mathbf{s}_i$ diffracted in the direction of $\mathbf{s}_o$ can be viewed as being reflected against a plane, adapted from [Drenth, 2007].

The scattering from a molecule can be similarly found from the superposition of the scattering from each of its constituent atoms. The scattering contribution from an atom j at position $\mathbf{r}_j$ is given as a phase change of $2\pi i \mathbf{r}_j \cdot \mathbf{u}$ to the atomic scattering factor $f_j(|\mathbf{u}|)$ and summing over the n atoms gives the total scattering from the molecule as

$$F(\mathbf{u}) = \sum_{j=1}^{n} f_j(|\mathbf{u}|) \exp(2\pi i \mathbf{r}_j \cdot \mathbf{u}). \tag{1.2}$$

Equation (1.2) can be rewritten by integration over the continuous electron density of the molecule, denoted $\rho(\mathbf{r})$, occupying the volume V, giving

$$F(\mathbf{u}) = \int_V \rho(\mathbf{r}) \exp(2\pi i \mathbf{r} \cdot \mathbf{u}) \mathrm{d}\mathbf{r}, \tag{1.3}$$

which corresponds to the Fourier transform.

In order to increase the diffraction to measurable levels, the molecule is crystallised. Unfortunately, crystallising introduces translational periodicity which samples the diffraction and limits the information available in the diffraction pattern. To understand why, let us first model the crystal electron density.

A crystalline object repeats a motif, called the *unit cell*, translationally along three dimensions. The set of all translation vectors of the motif is known as the crystal lattice L, and is given as the set of vectors

$$\mathbf{t}_{mnp} = m\mathbf{a} + n\mathbf{b} + p\mathbf{c}, \tag{1.4}$$

where m, n and p are integers and the volume encompassed by the lattice vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ is the unit cell. For an infinite crystal, the crystal electron density can be described as the convolution of the lattice of L and the unit cell contents $\rho(\mathbf{r})$, so that

$$g(\mathbf{r}) = \rho(\mathbf{r}) \otimes \sum_m \sum_n \sum_p \delta(\mathbf{r} - \mathbf{t}_{mnp}), \tag{1.5}$$

where $\otimes$ denotes convolution. The diffraction by the crystal is the Fourier transform of $g(\mathbf{r})$. This is given by

$$G(\mathbf{u}) = P(\mathbf{u}) \sum_h \sum_k \sum_l \delta(\mathbf{R} - \mathbf{t'}_{hkl}), \tag{1.6}$$

where $P(\mathbf{u})$ is the Fourier transform of $\rho(\mathbf{r})$ and $\mathbf{t'}_{hkl}$ is the set of *reciprocal lattice vectors* given by

$$\mathbf{t'}_{hkl} = h\mathbf{a'} + k\mathbf{b'} + l\mathbf{c'}, \tag{1.7}$$

where $\mathbf{a'}$, $\mathbf{b'}$, $\mathbf{c'}$ are the reciprocal unit cell vectors. The unit cell and reciprocal lattice unit cell vectors are related by

$$\mathbf{a'} = \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}} \qquad \mathbf{b'} = \frac{\mathbf{c} \times \mathbf{a}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}} \qquad \mathbf{c'} = \frac{\mathbf{a} \times \mathbf{b}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}}, \tag{1.8}$$

where $\times$ denotes the vector cross product. Inspection of equation (1.6) therefore shows that the diffraction by the crystal is equivalent to diffraction by the motif (unit cell contents), but sampled on the reciprocal lattice.

1.3 THE PHASE PROBLEM

Despite tremendous advances in XRC, the phase problem remains one of the limiting factors hampering routine protein structure determination. Many phase problems faced in XRC experiments may be classified as ill-posed NP-complete problems [Zwick et al., 1996]. Solutions to NP-complete problems can be verified in polynomial time but no algorithm has yet been developed to obtain the solution in under polynomial time. Additionally, ill-posed problems have more parameters than measurements (data) and are thus not uniquely solvable. A solution to the phase retrieval thus requires additional independent data or *a priori* knowledge to render the solution to the phase problem unique, and a way to use the additional data to either change the problem to a simpler one, or one with reduced dimensionality with a solution landscape that can be algorithmically searched. Additional data in X-ray crystallography can originate either from *a priori* knowledge on the solution, or from experiments.

1.3.1 Nature of the phase problem

To each reflection in the diffraction pattern corresponds a wave amplitude, measurable, but also a wave phase that is unfortunately not measurable and therefore lost during the experiment. The diffraction at position $\mathbf{h}$ on the reciprocal lattice is thus a complex quantity $F_{\mathbf{h}}$, in X-ray crystallography called the *structure factor*, given in terms of a magnitude, $|F_{\mathbf{h}}|$ and a phase, $\phi_{\mathbf{h}}$, where

$$F_{\mathbf{h}} = |F_{\mathbf{h}}| \exp(i\phi_{\mathbf{h}}). \tag{1.9}$$

Once the phases are known, the electron density map of the unit cell can be obtained by inverse Fourier transforming equation (1.9),

$$\rho(\mathbf{x}) = \frac{1}{V} \sum_{\mathbf{h}} F_{\mathbf{h}} \exp(i2\pi\mathbf{h} \cdot \mathbf{x}), \tag{1.10}$$

where V is the volume of the unit cell.

The inverse Fourier transform of the square of the structure amplitudes is referred to as the Patterson function [Drenth, 2007], denoted here by $P(\mathbf{x})$, i.e,

$$P(\mathbf{x}) = \frac{1}{V} \sum_{\mathbf{h}} |F_{\mathbf{h}}|^2 \exp(i2\pi\mathbf{h} \cdot \mathbf{x}). \tag{1.11}$$

The Patterson function can be calculated without knowledge of the phases, and has a number of applications in X-ray crystallography [Drenth, 2007].

1.3.2 Phase retrieval

Solving a phase problem and recovering the phases is known as *phase retrieval*. A number of complementary methods are used and are briefly described in this section.

1.3.2.1 Direct methods

Direct methods refers to a collection of techniques that use the relationships between the structure factor amplitudes and phases to directly recover the latter from the former [Harker and Kasper, 1948, Giacovazzo, 1999]. These relationships exist because the electron density is not random and constraints can be defined upon it.

The constraint used to derive phase relationships in direct methods is the atomicity of the electron density, i.e. the electron density consists of separated atomic peaks. Other constraints such as positivity of the electron density and a random distribution of atoms are also used to formulate probabilitic phase relationships [Woolfson, 1987].

Unfortunately, the constraints used in the direct methods do not scale well with the structure size. For large molecules the phase probability distributions become flat and no additional information is obtained. Furthermore, high resolution data on which the atomicity constraint depends becomes harder to collect for large molecules due to disorder and the constraint becomes ineffective. For these reasons, direct methods are today very successfully used in small molecule crystallography, for molecules containing up to about a thousand non-hydrogen atoms, and are not effective for large biological molecules [Usón and Sheldrick, 1999].

1.3.2.2 Isomorphous replacement

Isomorphous replacement (IR) attempts to recover the phases experimentally and was the first method used to solve the phase problem in protein crystallography [Brito and Archer, 2013]. This is an effective approach that is used if no *a priori* information is known about the structure of the molecule. The method starts with the measurement of diffraction data from one (single isomorphous replacement - SIR) or more (multiple isomorphous replacement - MIR) heavy atom crystal derivatives along with the native structure crystal. Here, heavy atoms refers to high Z atoms such as Hg, U, Pb, Pt, which present a much higher atomic scattering factor to that of the CHNOPS (carbon, hydrogen, nitrogen, oxygen, phosphorus, sulfur) atoms that make up most of biological molecules.

Crystal derivatives are obtained by soaking the native structure crystal in a reagent solution containing the heavy atoms. The reagent permeates the crystal and delivers the heavy atoms to reactive sites on the structure. As the term “isomorphous” suggests, this technique requires that the addition of heavy atoms to the native structure does not significantly alter the structure or the packing arrangement of molecules in the crystal (same unit cell parameters). This can be a difficult prospect in practice.

The locations of the heavy atoms in the structure is usually available from Patterson methods or direct methods. Equations are derived that can be solved for the phases, using data consisting of the sets of diffraction data and the heavy atom positions [Drenth, 2007].

1.3.2.3 Anomalous scattering

Each atom type has specific electronic transitions at which X-ray photons of the corresponding energy are absorbed. The atomic scattering factor used in equation (1.2) assumes elastic scattering and does not include the change in the diffracted X-ray phase observed near the absorption edges of an atom. Generally, anomalous scattering is included in the atomic scattering factor as

$$\mathbf{f} = f + \Delta f' + i\Delta f'', \tag{1.12}$$

where f is the normal atomic scattering factor far from an absorption edge, and $\Delta f'$ and $\Delta f''$ are dispersion corrections that depend on the atomic number Z and the X-ray wavelength or photon energy, and are listed in the International Tables for Crystallography, Volume C [Prince, 2006]. This results in changes in the diffraction patterns, the most visible being the breakdown of Friedels law, i.e. $|F(\mathbf{h})| \neq |F(-\mathbf{h})|$.

For the photon energies usually used in X-ray crystallography, light atoms do not contribute to anomalous scattering. In practice then, a crystal derivative must be used. The most common derivative uses selenium atoms. Selenium atoms can easily be incorporated in the protein by replacing the amino acid methionine with selenomethionine (SeMet) [Hendrickson et al., 1990]. Selenomethionine can be incorporated in proteins with no effects on the protein structure, an advantage compared to the heavy atom derivatives used in IR.

Diffraction data from SeMet derivatives and native structures can be used to determine the phases in a method similar to that for MIR. This is referred to as multiple anomalous dispersion (MAD) and has become an important and widely used method to solve the phase problem in protein crystallography. Crystal derivatives used in MIR can also be used, allowing MAD and MIR to be used together to solve a structure.

1.3.2.4 Molecular replacement

The Euclidean distance preserving property of the Fourier transform shows that if two objects are similar in real space then they have similar structure factors in reciprocal space. Accordingly, if a known structure, called the *model*, is known to be analogous to that of an unknown structure, called the *target*, then the Fourier phases of the model can be used along with the Fourier amplitudes of the target to determine an approximation to the structure of the target. This constitutes, in a nutshell, the simple idea behind molecular replacement (MR) [Rossmann, 1972].

The success of MR phasing is dependent on the quality of the structural homology between the model and the target. In practice, structural homology between two proteins is assessed by comparing the proteins' amino acid sequences. Sufficient structural homology is expected for MR phasing, with at least 35% sequence identity, which generally corresponds to a C^{α} root-mean-square deviation (RMSD) of about 1.5 Å [Abergel, 2013]. The method proceeds by recreating the unit cell of the target crystal from the model protein. This involves the determination of the rotation and translation of the model within the unit cell. The orientation of the target protein can be determined by comparing the Patterson map to that of a set of Patterson maps derived from the model in different orientations [Drenth, 2007]. Similarly, the positions of oriented models can be made to match that of the target using a probabilistic maximum likelihood translation function. After optimisation of the position and orientation of the model, the diffraction from the model crystal is simulated to obtain approximate starting phases. This step is often followed by subsequent refinement steps to obtain more accurate phases.

The more protein structures that are known, the greater the chance of finding an homologous structure leading to a successful determination of the target structure. This virtuous circle and the leverage of past structural knowledge has promulgated the MR technique as the most successful phase retrieval technique. About 70% of all the structures deposited in protein databanks used the MR phasing technique [Berman et al., 2000]. In fact, MR phasing of diffraction data can be near-entirely automated and run on most laptop computers in matter of hours. That being said, molecular replacement can be affected by model bias (where the solution resembles the model rather than the target) and, in the adverse case for which no homologous models can be found, is ill-suited for finding new structural folds. The reuse of incorrect model structures can also lead to a compounding of structural errors and erroneous folds.

1.3.2.5 Ab initio phasing

Ab initio phasing is really an extension of the direct methods to large proteins. In its purest form, *ab initio* phasing utilises only the information in the diffraction pattern of the native protein and general information that can be found with minimal effort such as the molecular weight, the protein occupied portion of the unit cell, the presence of non-crystallographic symmetry, or the fact that the protein is made up of a chain of amino acids residues of known sequence, that is easily obtainable. Assumptions about the electron density can also be used such as positivity, prediction of the secondary structure of the protein, the likely globular shape of the protein, and likely position in the unit cell to form non-overlapping inter-cell contacts. The *ab initio* phasing approach is insensitive to model bias, does not require additional beam time or crystal derivatives, and is not limited to small proteins, and so is the holy grail of all phasing techniques. *Ab initio* phasing has, however, so far been unsuccessful in general in

protein crystallography.

Iterative projection algorithms are one of the algorithms being used in attempts to implement *ab initio* phasing, and are described in the next section. Uniqueness of the solution to the phase problem is an important aspect when considering *ab initio* phasing, as the amount of additional data may not be sufficient to uniquely recover the phases. *Ab initio* phasing and uniqueness are the major focus of this thesis.

1.4 ITERATIVE PROJECTION ALGORITHMS

In *ab initio* phasing, the *a priori* information and the data available can generally be expressed as constraints. With this description, the phase problem can be formulated as a constraint satisfaction problem, where solutions are found at the intersection of all the constraints.

Iterative projection algorithms (IPAs) are search algorithms for constraint satisfaction problems. They recursively apply a combination of projections to an iterate, denoted $\mathbf{w}$, according to an update rule. In most cases only two constraints are considered, and this is the case considered here.

1.4.1 Constraints and projections

Projections and constraints are conveniently described as operations in a vector space, rather than the physical space itself. In this description, a vector $\mathbf{w} = [w_1, ..., w_n]^T$ in the n-dimensional space represents the function $f(\mathbf{x})$, where each vector component w_j corresponds to the value of one sample of $f(\mathbf{x})$. Each point in the n-dimensional vector space corresponds to a different function (or electron density) $f(\mathbf{x})$.

A constraint set $\mathcal{C}$ is a region encompassing all vectors whose corresponding function satisfies the constraint. These regions can be characterised by their convexity. Graphically, a constraint set is convex if all points in the line segment between any two points in the set are also in the set, as shown in Fig. 1.3. Mathematically, if and only if $\mathbf{w}$ and $\mathbf{y}$ are any two vectors in a convex constraint set $\mathcal{C}$, then

$$\forall \eta \in [0, 1], \mathbf{w} + \eta(\mathbf{y} - \mathbf{w}) \in \mathcal{C}. \tag{1.13}$$

As shown in Fig. 1.3, non-convex sets do not satisfy this condition.

As a Euclidean metric space, the distance between two points $\mathbf{y}$ and $\mathbf{w}$ is given by the Euclidean norm

$$||\mathbf{y} - \mathbf{w}|| = \sqrt{\sum_{j=1}^{n} (y_j - w_j)^2}. \tag{1.14}$$

The projection of a point $\mathbf{w}$ onto the constraint $\mathcal{C}$ is defined as the closest point $\mathbf{y} \in \mathcal{C}$ to $\mathbf{w}$. Mathematically, denoting by $P_{\mathcal{C}}$ the projection onto the constraint $\mathcal{C}$, the projection

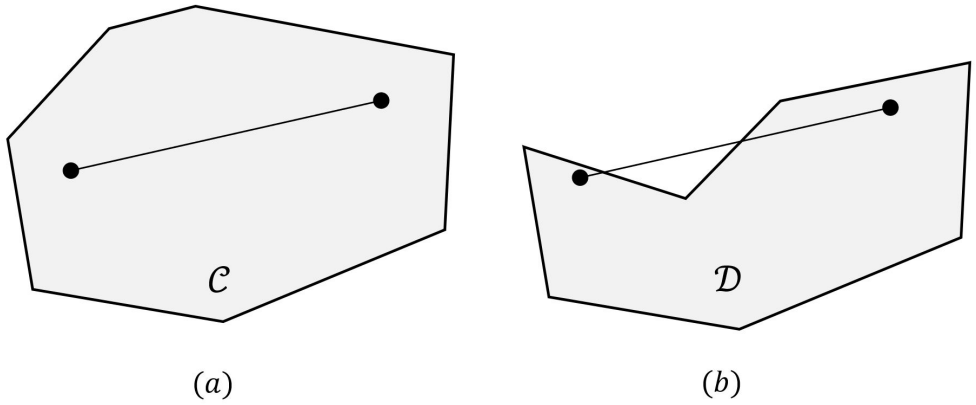


Figure 1.3 (a) A convex set $\mathcal{C}$ and (b) a non-convex set $\mathcal{D}$.

of $\mathbf{w}$ onto $\mathcal{C}$ is defined by

$$P_{\mathcal{C}}\mathbf{w} = \underset{\mathbf{y}\in\mathcal{C}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{w}||, \tag{1.15}$$

where $\operatorname{argmin}_{\mathbf{w}\in\mathcal{C}} f(\mathbf{w}) = \{\mathbf{w} \mid \mathbf{w} \in \mathcal{C} \wedge \forall \mathbf{y} \in \mathcal{C} : f(\mathbf{w}) \leq f(\mathbf{y})\}$. The projection is idempotent, i.e.

$$P_{\mathcal{C}}P_{\mathcal{C}}\mathbf{w} = P_{\mathcal{C}}\mathbf{w}. \tag{1.16}$$

Often, relaxing a projection can help with the global search properties of IPAs. A relaxed projection for the constraint $\mathcal{C}$, denoted $T_{\mathcal{C}}\mathbf{w}$ is defined as

$$T_{\mathcal{C}}\mathbf{w} = P_{\mathcal{C}}\mathbf{w} + \gamma_{\mathcal{C}}(P_{\mathcal{C}}\mathbf{w} - \mathbf{w}), \tag{1.17}$$

where $\gamma_{\mathcal{C}}$ is a relaxation parameter. In the special case $\gamma_{\mathcal{C}} = 1$ the projection is called a reflection, denoted $R_{\mathcal{C}}\mathbf{w} = 2P_{\mathcal{C}}\mathbf{w} - \mathbf{w}$.

1.4.2 Error reduction algorithm

The error reduction (ER) algorithm is the simplest form of IPA and consists in alternatively projecting the iterate back-and-forth between two constraint sets [Fienup, 1982]. The ER update rule is given by

$$\mathbf{w}_{i+1} = P_B P_A \mathbf{w}_i, \tag{1.18}$$

where $\mathbf{w}_i$ denotes the "iterate" at iteration i.

Three situations are depicted in Fig. 1.4. In the first case both constraints are convex, the ER reduces the error (distance of the iterate to the solution) after each iteration and is assured to converge, albeit possibly slowly, to the solution. The second case illustrates the situation where at least one of the constraint sets is non-convex.

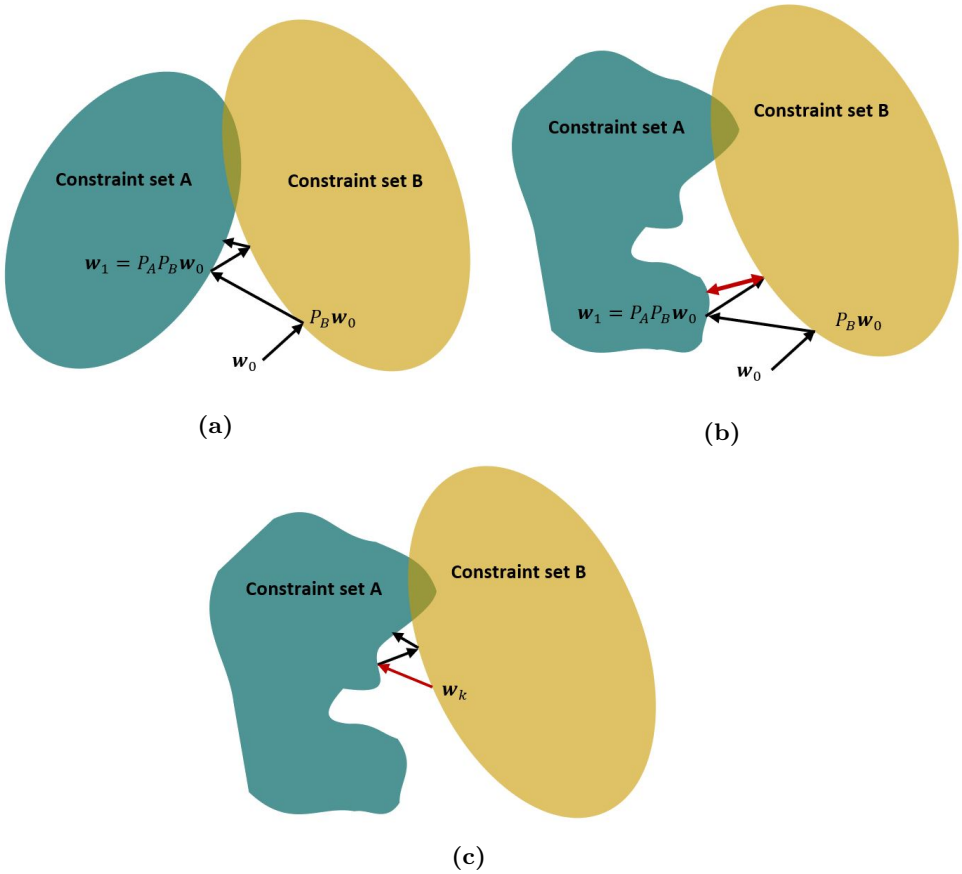


Figure 1.4 Behaviour of the ER algorithm for (a) two convex constraint sets, (b) at least one constraint set is non-convex and, (c) used as a refinement algorithm when constraint sets are "locally" convex.

In such cases the algorithm will usually stagnate at a local minimum which is not a solution, i.e. does not satisfy all the constraints. In the last case, similar to the second case except that the iterate is already close to a solution and the ER algorithm is used as a refinement algorithm. In general then, the ER algorithm is not suitable for non-convex problems, such as phase retrieval where an initial estimate of the solution is not available.

A related algorithm, the relaxed projection (RP) algorithm replaces the ER projections with their relaxed versions giving

$$\mathbf{w}_{i+1} = T_B T_A \mathbf{w}_i. \tag{1.19}$$

With the relaxation parameters $\gamma_{A,B}$ usually chosen such as $0 < \gamma_{A,B} < 1$. This algorithm can sometimes avoid stagnation and converge more rapidly, but it is not generally effective with convex constraints.

1.4.3 Hybrid input-output algorithm

The hybrid input-output (HIO) algorithm, developed by Fienup [Fienup, 1982] for astronomy, is an extension to the ER algorithm that is used in those case where only the Fourier amplitude, support and positivity constraints need to be enforced. Its update rule is given as [Millane and Lo, 2013]

$$\mathbf{w}_{i+1} = (P_A P_B + P_{A'}(I - \beta P_B))\mathbf{w}_i, \tag{1.20}$$

where A' is the complement set of A and β a parameter usually chosen as 0.7.

Contrary to the ER algorithm, the HIO algorithm is able to avoid stagnation even in the case of the non-convex Fourier amplitude constraint. The HIO algorithm was the first algorithm that effectively avoids stagnation and is still a popular IPA, although the constrains that it accomodates are restricted [Millane and Lo, 2013]. A generalisation of the HIO algorithm to accept for a wider range of constraints has been described by Millane and Stroud [Millane and Stroud, 1997].

1.4.4 Difference map algorithm

The difference map algorithm (DM) was derived by Elser [Elser, 2003a, Elser, 2003c] and is designed to overcome stagnation for non-convex problems. The update rule for the DM algorithm is given by

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \beta(P_A T_B \mathbf{w}_i - P_B T_A \mathbf{w}_i), \tag{1.21}$$

where T_A and T_B are the relaxed projections of P_A and P_B with relaxation parameters γ_A and γ_B, respectively. The nonzero DM parameter β is often chosen with $0.7 \leq |\beta| \leq 1$ and the relaxation parameters are often fixed to $\gamma_A = -1/\beta$ and $\gamma_B = 1/\beta$ [Elser, 2003b, Elser, 2003a]. A block diagram representation of the DM algorithm is given in Fig. 1.5.

At convergence, $\mathbf{w}_{i+1} = \mathbf{w}_i$, referred to as a fixed point of the algorithm, gives from equation (1.21),

$$P_A T_B \mathbf{w}_i = P_B T_A \mathbf{w}_i = \tilde{\mathbf{w}}. \tag{1.22}$$

Because P_A and P_B are the final operations in equation (1.22), $\tilde{\mathbf{w}}$ satisfies both constraints and is the solution sought. This is in contrast to ER where the solution cannot usually be obtained from the iterate at convergence. The DM algorithm therefore avoids stagnation near local minima and has good search properties. Progression and convergence of the DM can be monitored by calculating the difference between the

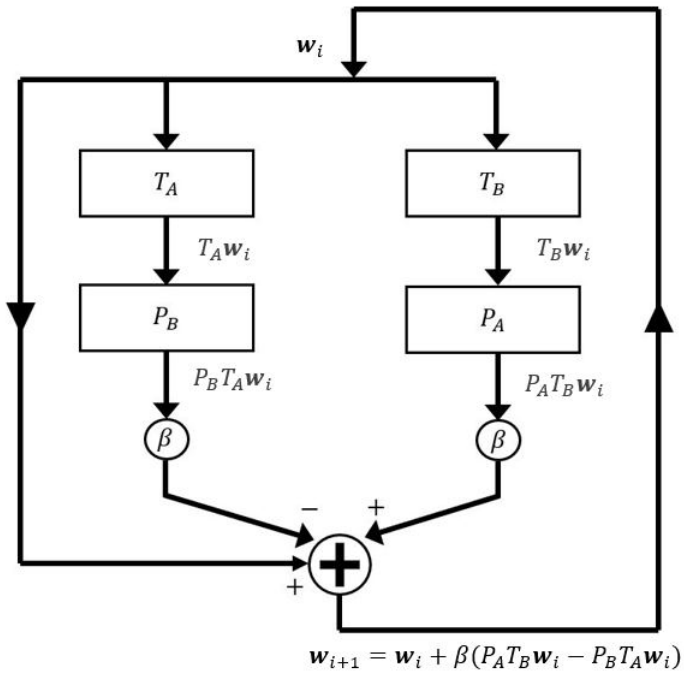


Figure 1.5 Block diagram of the difference map algorithm.

projections

$$\Delta = |P_A T_B \mathbf{w}_i - P_B T_A \mathbf{w}_i|, \tag{1.23}$$

and Δ decreases to zero at a fixed point.

1.4.5 Specific constraints, projections and error metrics

In this section common constraints and their projections are described. Usually, for each constraint an error metric can be defined by calculating the distance of the current iterate to the corresponding constraint set. The progression and convergence of the algorithm are monitored by calculating error metrics. Examples of error metrics are also given.

1.4.5.1 Support constraint

The support constraint corresponds to the fact that outside a support region, denoted $\mathcal{S}$, the electron density has no crystalline order (liquid phase) and only contributes to the background scattering in diffraction patterns. The corresponding projection

operator, $P_{\mathcal{S}}$, is defined as,

$$\forall j = 1, \ldots, n \qquad P_{\mathcal{S}} w_j = \begin{cases} w_j & j \in \mathcal{S} \\ 0 & \text{otherwise,} \end{cases} \tag{1.24}$$

where the electron density sample values outside the support $\mathcal{S}$ are set to zero while all values inside the support are left untouched.

In simulations, the true solution denoted $\mathbf{w}_{\text{true}}$ is known. Monitoring the distance between the true solution to the estimated solution $\mathbf{w}_{\text{est}}$, i.e. $||\mathbf{w}_{\text{true}} - \mathbf{w}_{\text{est}}||$ can be used to detect convergence of the iterative projection algorithm. An often used error metric is the root-mean-square (RMS) image error, defined as

$$e = \sqrt{\frac{\sum |\mathbf{w}_{\text{est}} - \mathbf{w}_{\text{true}}|^2}{\sum |\mathbf{w}_{\text{true}}|^2}}, \tag{1.25}$$

where the sum is over all the vector components.

1.4.5.2 Fourier amplitude constraint

The Fourier amplitude constraint is the constraint imposed by the experimental Fourier amplitude data. The Fourier amplitude projection, $P_{\mathcal{M}}$, is given as

$$P_{\mathcal{M}} \mathbf{w}_i = \mathcal{F}^{-1} \left(\sqrt{\frac{\mathbf{I}_{\text{true}}}{|\mathcal{F}\{\mathbf{w}_i\}|^2}} \mathcal{F}\{\mathbf{w}_i\} \right) \tag{1.26}$$

where $\mathbf{I}_{\text{true}}$ is the vector of measured intensities. Note that the Fourier amplitude constraint is a non-convex constraint as it corresponds to the intersection of $(2n-2)$-dimensional hyper-cylinders in $\mathbb{R}^{2n}$.

The RMS Fourier error is given as

$$E = \sqrt{\frac{\sum (\sqrt{\mathbf{I}_{\text{true}}} - \sqrt{\mathbf{I}_{\text{est}}})^2}{\sum \mathbf{I}_{\text{true}}}}, \tag{1.27}$$

where the summation is over all measured intensities. In most IPA implementations, this error metric generally equals zero after application of the Fourier space projection and thus must be calculated after the real space projection.

Uniqueness of the solution to the phase problem is important, because if it is non-unique, false solutions will satisfy the constraints and some of the error metrics will approach zero. In simulations, because the true solution is known, this problem is avoided by computing e in equation (1.25), false solutions can therefore be detected.

1.5 UNIQUENESS OF THE PHASE PROBLEM

Additional independent data, or *a priori* knowledge, are required to solve the crystallographic phase problem. In most phasing techniques, with the notable exception of *ab initio* phasing, the additional data constitute extremely powerful constraints that are sufficient to render the solution to the problem unique. However this cannot be said for the weak constraints generally used in *ab initio* phasing.

Before attempting to solve the phase problem using *ab initio* phasing techniques, the uniqueness of the solution to the phase problem must first be considered. For this purpose, Elser and Millane [Elser and Millane, 2008] introduced the constraint ratio, denoted Ω, that is defined as the ratio of the number of independent data to the number of unknown parameters of the phase problem. An ill-posed problem will have more unknown parameters than independent data and thus $\Omega < 1$. On the contrary, a phase problem with $\Omega > 1$ will have more independent data than unknown parameters and a unique solution to the problem is thus likely. In practice, even with $\Omega > 1$ but close or equal to 1, a unique solution is not assured due to the non-linearity of the problem.

For the case of a single non-crystalline object, such as in SPI experiments or other imaging applications that do not involve crystals, where the Fourier intensity can be measured continuously, the number of parameters is proportional to the volume of the object, i.e. the number of samples in the object support $\mathcal{S}$, denoted by $|\mathcal{S}|$, with $|\cdot|$ denoting the volume or number of samples. From the Fourier amplitude data, one can calculate the autocorrelation of the object, denoted $\mathcal{A}(\mathbf{x})$, by inverse Fourier transforming the square of the Fourier magnitudes. Since the autocorrelation is conjugate centrosymmetric, i.e. $\mathcal{A}(\mathbf{x}) = \mathcal{A}^*(\mathbf{x})$, only half of this data are independent, and the number of independent data is proportional to half the volume of the autocorrelation of the object support, or $0.5|\mathcal{A}|$. The constraint ratio for a single non-crystalline object is thus given as [Elser and Millane, 2008]

$$\Omega = \frac{|\mathcal{A}|}{2|\mathcal{S}|}. \tag{1.28}$$

The constraint ratio is bounded below as

$$\Omega \geq 2^{D-1}. \tag{1.29}$$

where D is the dimensionality of the problem. For objects with a convex and centrosymmetric support, $\Omega = 2^{D-1}$.

In the 3D crystal case, where the intensity data is only available at the Bragg positions, the constraint ratio, denoted Ω_c, was determined by Millane and Lo [Millane

and Lo, 2013] and given as

$$\Omega_c = \frac{1}{2p}, \tag{1.30}$$

where p is the protein fraction of the unit cell volume. A solvent content of more than 50% is thus sufficient to obtain $\Omega_c > 1$ needed for solving the phase problem uniquely.

1.6 X-RAY SOURCES

The history of X-ray crystallography parallels the development of X-ray sources. In fact, the various X-ray sources are often referred to as generations to emphasise the various evolutions in the X-ray source technologies.

1.6.1 X-ray tubes and synchrotron light sources

The first generation includes the past and modern laboratory X-ray tube sources. As a simplified description, in X-ray tubes, electrons emitted from the cathode are accelerated towards the anode by an applied *tube voltage*, the electrons collide with the anode and two processes, characteristic X-ray emission and the Bremsstrahlung effect, are involved in the production of X-rays.

The Bremsstrahlung effect, from the German words *Bremsen* (to brake) and *Strahlung* (radiation), is the radiation that occurs when the electrons are decelerated, in this case in the anode material, converting kinetic energy into radiation. The emission spectrum from the Bremsstrahlung effect in an X-ray tube is continuous and with an intensity dwarfed by the characteristic X-ray emission lines of the anode material, usually copper or chromium that are used in X-ray crystallography. Characteristic X-rays are produced when a high energy electron ejects an inner shell bound electron of the anode material, the vacant energy level is then filled by outer shell electrons with the production of photons with a quantized energy corresponding to that of the difference between the energy levels specific to that of the anode material. Two energy levels are often used in X-ray crystallography, corresponding to transitions from the L shell to the K shell and known as K-alpha emissions. Filters, monochromators or X-ray mirrors are often used to select specific X-ray emission lines [Drenth, 2007].

Synchrotron light sources are much brighter than X-ray tubes. Synchrotrons circulate electrically charged particles at relativistic speeds in a storage ring. When the beam changes direction due to the presence of magnets (bending magnets, wigglers or undulators) in their paths, X-rays are produced, this is called *magnetobremsstrahlung radiation* or synchrotron radiation.

1.6.2 X-ray free electron laser sources

The latest addition, the X-ray free-electron laser (XFEL) source, at a cost of about a billion US dollars, is often considered as the fourth generation, but because of their extreme brilliance, ultra-short and spatially coherent pulses they are in a class of their own, and have ushered in new kinds of X-ray diffraction experiments.

X-ray free-electron laser sources are constructed from three main sections: an electron source (electron injector), a linear accelerator and an undulator arranged linearly in that order. A short description of these follows.

The injector produces short bunches of electrons by extracting them from a solid cathode with a conventional laser and then shaping and compressing the electron bunch. The quality of this step is crucial as any variations would be amplified in the next steps. The ultra short pulses of XFELs are in part due to the very compact electron bunch created at this step.

The electron bunches are then accelerated within a linear accelerator (linac). The latest generation of XFELs uses a superconducting linac that can reach hundreds and even thousands of meters in length and accelerates electrons to energies of up to 20 GeV. The energy of these electrons is tunable and so is the final wavelength of the X-rays.

Relativistic electron bunches are collimated before entering the undulator. Undulators are special arrangements of permanent magnets with alternating fields that change the course of the electron bunches in a sinusoidal path. X-ray photons are emitted by magnetobremsstrahlung effect when the electrons have an acceleration perpendicular to their velocity.

1.7 SERIAL FEMTOSECOND CRYSTALLOGRAPHY

1.7.1 Sample delivery

Ideally, in SFX, for each X-ray pulse a new crystal is introduced in the focus of the XFEL beam. Different approaches have been developed for this, including, but not limited to: fixed targets, gas dynamic virtual nozzle (GDVN), lipidic cubic phase (LCP) and electrospinning.

Liquid microjet injectors carry fully hydrated crystals at the XFEL beam interaction point in a continuous liquid jet. The diameter of the jet must ideally be similar to that of the X-ray beam diameter ($1 - 10\,\mu m$) for increased hit rates and minimal crystal waste [Boutet et al., 2018]. Due to the high pulse repetition rates of new XFEL sources, flow speeds of $10 - 100\,m\,s^{-1}$ must be reached in order to clear the debris from previous hits and to present a new crystal in the beam focus before the next pulse occurs. Such an injector must be reliable and not be clogged by the crystals thus, requiring a large nozzle orifice diameter. Small jet diameters and large nozzle orifice sizes are

made possible through the use of a gas dynamic virtual nozzle (GDVN). In a GDVN, a large inner sample solution capillary (typically $40 - 75\,\mu m$) is surrounded by an outer capillary carrying a high pressure gas, typically helium [DePonte et al., 2008]. The co-flowing gas accelerates the sample solution and shear and pressure forces reduce the diameter of the inner jet by a factor of 10 to 50 [Boutet et al., 2018]. GDVN requires extremely tight manufacturing tolerances, and state-of-the-art 3D printed nozzles with 500nm resolution are generally used to assure the proper centering of the capillaries.

Double flow focussing nozzles (DFFN) are an extension to the GDVN where a second liquid can be used to reduce the sample flow rate [Oberthür et al., 2017]. Three concentric capillaries are used: the outer capillary carries the focussing gas, the innermost capillary carries the sample in its crystallisation buffer and an accelerated focusing liquid miscible with the buffer is pumped in the middle capillary. With control of the flow rates of the sample and focussing liquid, the sample consumption can be minimised without affecting the jet stability, with the added advantage that the sample can be changed without disturbing the jet.

High viscosity injectors such as lipidic cubic phase (LCP) injectors are used when the viscosity of the sample is too high for using GDVNs. Membrane proteins are notoriously difficult to crystallise and produce in great quantities. The crystallisation of membrane proteins is complicated due to their flexibility, hydrophobic surfaces and lack of stability [Caffrey, 2015]. A popular crystallisation approach for membrane proteins uses lipids as the crystallisation medium, membrane proteins can freely diffuse in the lipid, as it mimics a membrane-like environment, and can therefore be concentrated and form crystals [Landau and Rosenbusch, 1996]. The cubic phase has the consistency of thick toothpaste and has shown to be a convenient approach to deliver membrane protein crystals to the XFEL beam. The LCP injector consists of a hydraulic plunger pushing LCP out a capillary with pressure of up to 10,000psi. A co-flowing gas is used to keep the LCP stream aligned. Contrarily to GDVNs, the flow rate of LCP injector can be optimised for exposing a fresh section of LCP stream, dramatically reducing crystal waste [Weierstall et al., 2014].

Fixed target delivery approaches use a thin solid support membrane to carry the crystals to the XFEL beam [Hunter et al., 2014]. Support membranes are ideally as thin as possible to reduce background scattering and are often etched from a silicon wafer coated with silicon nitride, creating regularly spaced silicon nitride membrane windows where the crystals are deposited. The fixed target must be moved after each shot to present a new window, and thus a new crystal, to the XFEL beam. This can be done by a motorised setup moving the fixed target synchronised to the XFEL pulse rate. Keeping crystals well hydrated on fixed targets has also shown to be a great challenge.

1.7.2 Hit finding

In a typical SFX experiment, hundreds of thousands, up to a few million frames, are recorded. Many of those frames (typically above 60%) are blanks as no crystal was present in the X-ray beam [Boutet et al., 2018]. Furthermore, frames containing too few Bragg peaks and frames abnormally noisy are usually discarded. The first analysis step in an SFX experiment consists of selecting the good hits from the mountain of data generated by the X-ray detector in a step known as hit-finding. The hit finding task seems trivial at first but is complicated by the fluctuations inherent to each XFEL pulse and variability of the background noise. Hit finding is usually performed by software, such as Cheetah or psocake, by first finding intensity peaks in the frame [Barty et al., 2014]. These potential Bragg spots are then tested against several metrics such as SNR, area of the peak, sum of the peak's pixels values. If a frame has a certain number of peaks (typically 15), then it is considered to be a hit [Boutet et al., 2018].

1.7.3 Indexing and integrating

The next analysis step in SFX consists of assigning the Miller indices to the Bragg peaks known as *indexing* the diffraction patterns. For this, the lattice parameters and the diffraction pattern orientations must be determined. The lattice parameters and orientations can be geometrically determined from the peaks positions and experiment parameters such as detector dimensions, detector panels positions and camera length. Often, a known sample will be shot beforehand to refine the experiment and detector geometries. Many patterns are discarded during this step as it is unlikely that a lattice corresponds to spurious intensity peaks. Once indexed, the intensities of peaks of the same indices are averaged, this approach is referred to as the *Monte Carlo integration method* [Kirian et al., 2010].

1.8 OPEN PROBLEMS

There are, of course, many open problems in protein X-ray crystallography and the application of XFELs.

Although phasing of the diffraction data is an advanced area, with many powerful methods available, there is still an active interest in developing new phasing methods. In particular, methods that requires less *a priori* information or use data from new experimental techniques. Some new phasing methods are developed in this thesis, that make use of diffraction data from two-dimensional crystals, and from multiple crystal forms of the same molecule.

2 ASPECTS OF THE PHASE PROBLEM FOR 3D CRYSTALS

2.1 INTRODUCTION

The basic properties, particularly uniqueness, of the phase problem for crystalline objects were described in Section 1.5. In this chapter, aspects of some of these properties are studied in more depth. These include the relationship to the phase problem for a single object via the Patterson function, the problem in the presence of a volume constraint (i.e. an envelope of unknown shape but known volume), and the effect of crystallographic and non-crystallographic symmetry. An expression for the constraint ratio for crystals involving the Patterson function allows the effects of different real space constraints to be evaluated. The results assist in understanding the nature of the macromolecular crystallographic phase problem and the potential for *ab initio* phasing.

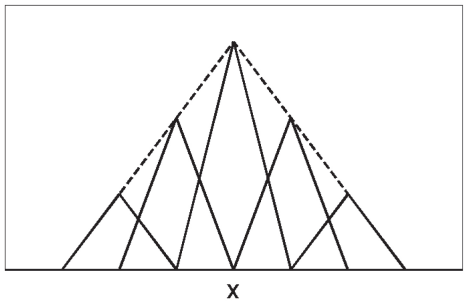
2.2 UNIQUENESS FOR A CRYSTAL

Uniqueness properties of the phase problem for crystalline specimens were described briefly in Section 1.5. The problem is considered here in more detail in terms of the Patterson function.

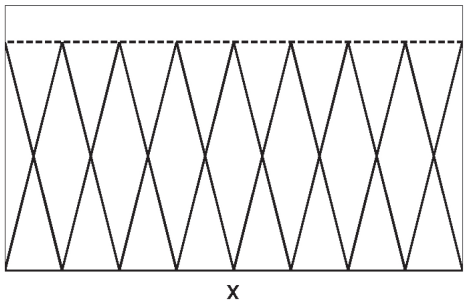
Consider first a finite crystal (object) of $N \times N \times N$ unit cells. The volume of the object is $|\mathcal{S}_N| = N^3 V$, where $\mathcal{S}_N$ is the support of the crystal and V is the volume of the unit cell. Since all the unit cells are the same, the number of independent object parameters is proportional to $|\mathcal{S}_N|/N^3 = V$. The normalised autocorrelation of the crystal, $A_N(\mathbf{x})$, can be written as

$$A_N(\mathbf{x}) = \frac{1}{N^3} \sum_{\mathbf{m}=-(\mathbf{N}-\mathbf{1})}^{\mathbf{N}-\mathbf{1}} (N - |m_1|)(N - |m_2|)(N - |m_3|)\, A(\mathbf{x} - \mathbf{m}\boldsymbol{\Lambda}) \,, \tag{2.1}$$

where $\mathbf{m} = (m_1, m_2, m_3)$, the matrix $\boldsymbol{\Lambda} = (\mathbf{a}|\mathbf{b}|\mathbf{c})^T$, where $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ are the unit cell vectors, i.e. the rows of $\boldsymbol{\Lambda}$ are the unit cell vectors, $\mathbf{N} = (N, N, N)$, and $A(\mathbf{x})$ is the autocorrelation of a single unit cell. This is illustrated for 1D for $N = 3$ in Fig. 2.1. The volume of the support of the autocorrelation of the crystal is $|\mathcal{A}_N| = 8N^3 V$.



(a)



(b)

Figure 2.1 The weighted autocorrelations of a single unit cell (solid lines) that make up the autocorrelation of a 1D crystal with $N = 3$ unit cells (dashed line) as in equation (2.1). (b) The Patterson function (dashed line) for an infinite crystal that is made up of an infinite number of equally-weighted autocorrelations of a single unit cell (solid lines).

However, as a result of equation (2.1), not all sample values of the autocorrelation are independent. Inspection of Fig. 2.1 shows that, in the 1D case, the whole of $A(\mathbf{x})$ can be determined from information on the boundary of $A_N(\mathbf{x})$, where there is no overlap. Similarly, in the 3D case, $A(\mathbf{x})$ can be determined from the boundary region of $A_N(\mathbf{x})$ where there is no overlap. The volume of this region, that contains independent data, is $8V$. Therefore, substituting into equation (1.28), the constraint ratio for the finite crystal is $\Omega_N = 8V/2V = 4$. The result is therefore the same as for a single object, as expected, since a finite crystal is a single object. In principle then, the whole finite crystal could be reconstructed from a measurement of its continuous diffracted intensity. In practice however, for all but very small crystals (small N), it would be difficult to measure the continuous diffracted amplitude between the Bragg reflections, due to its small values in these regions.

For a realistic crystal, N is large and we have to consider the limit $N \to \infty$. The autocorrelation $A_N(\mathbf{x})$ then extends to infinity and reduces to the Patterson function

$P(\mathbf{x})$, i.e.

$$\lim_{N\to\infty} A_N(\mathbf{x}) = \sum_{\mathbf{m}=-\infty}^{\infty} A(\mathbf{x}-\mathbf{m}\boldsymbol{\Lambda}) = P(\mathbf{x}) \ , \tag{2.2}$$

which is illustrated for 1D in Fig. 2.1(b). The boundary region of $A_N(\mathbf{x})$ is now not accessible, and all that is available is $P(\mathbf{x})$, which is periodic with a period that has volume V. Therefore, for a crystal, the number of data is proportional to V, and the constraint ratio, denoted Ω_c, is

$$\Omega_c = \frac{V}{2V} = \frac{1}{2} \ . \tag{2.3}$$

This is consistent with the result described in Section 1.5.

If additional real space information is available, the degrees of freedom in, or the unique region of, the unit cell and the Patterson function will be modified, and the constraint ratio is then given by

$$\Omega_c = \frac{|\mathcal{P}_u|}{|\mathcal{U}_u|} \ , \tag{2.4}$$

where $\mathcal{U}_u$ and $\mathcal{P}_u$ denote the unique region of the unit cell and of the Patterson function, respectively. Note that the 2 in the denominator of equation (1.28) is now absorbed into $|\mathcal{P}_u|$ since $\mathcal{P}_u$ is always centrosymmetric. Equation (2.4) gives the constraint ratio for a crystal, and is a function of only the shape and symmetry of the molecule and the unit cell (since $\mathcal{U}_u$ can be calculated from this information, and $\mathcal{P}_u$ can be calculated from $\mathcal{U}_u$). The constraint ratio in equation (2.4) can be used to characterise uniqueness of the crystallographic phase problem and the effects of different kinds of real space information.

2.3 REAL SPACE CONSTRAINTS

In this section, the constraint ratio is evaluated for four kinds of real space constraints: (1) a known molecular envelope, (2) an unknown molecular envelope of a known volume, (3) crystallographic symmetry, and (4) noncrystallographic symmetry.

2.3.1 Known molecular envelope

Consider the case where the molecule does not occupy all of the unit cell, which is essentially always the case in protein crystallography. Consider first the case where the molecular envelope is known *a priori*. The shape of the envelope can sometimes be obtained from experimental techniques such as solution scattering, electron microscopy, or solvent contrast variation [Hao, 2006, Carter et al., 1990, Lo et al., 2009]. If the shape of the molecular envelope is known, and assuming it can be positioned in the unit cell, then the number of unknowns is proportional to its volume, i.e. $|\mathcal{U}_u| = pV$, where p is the fraction of the unit cell occupied by the molecule.

Since a restricted molecular support (envelope) gives rise to a restricted autocor-

relation support, we need to consider the possibility that the Patterson function does not occupy the whole of the unit cell, reducing the size of its unique region to less than $V/2$. Let $|\mathcal{P}_u| = qV/2$, where q denotes the proportion of the unit cell that is occupied by the Patterson function, and substitution into equation (2.4) gives

$$\Omega_c = \frac{q}{2p} \, . \tag{2.5}$$

Since macromolecules must pack in a crystal in such a way that they make contacts with molecules in adjacent unit cells, they must occupy the unit cell in a fairly homogeneous manner. The result is that it is unlikely that the autocorrelation (of a single molecule) will not occupy all of the unit cell. It is even more unlikely that the Patterson function will not occupy all of the unit cell. This is illustrated in Fig. 2.2. In almost all cases then, $q = 1$, and equation (2.5) reduces to

$$\Omega_c = \frac{1}{2p} = \frac{1}{2(1-s)} \, , \tag{2.6}$$

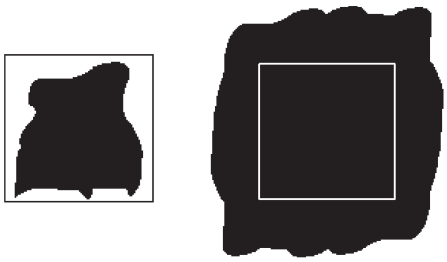
where s is the solvent content of the crystal. The constraint ratio then increases with increasing solvent content, as expected, and uniqueness ($\Omega_c < 1$) requires that $p < 0.5$, i.e. a protein content of less than 50%, or a solvent content of greater than 50%.

It is interesting to note that since generally $q = 1$, the constraint ratio Ω_c depends, as shown by equation (2.6), on only the volume of the envelope, relative to that of the unit cell, and not on its shape. This is in contrast to the single object case where Ω depends on the shape of the object, and not its volume.

2.3.2 Unknown molecular envelope

An important caveat of the previous section is that it assumes that the molecular envelope is known. Using $\mathcal{U}_u$ as the number of object variables in the constraint ratio definition implicitly assumes that it is known, at least for reconstruction purposes, what those variables are. This is not the case if the envelope is unknown, since it is not known which samples are inside the envelope. However, in many cases in protein crystallography, the protein envelope (or solvent) *volume*, rather than the envelope *shape*, can be estimated [Weichenberger and Rupp, 2014]. Consider now the case where the volume of the protein envelope, rather than the envelope itself, is known.

Consider a unit cell of $M \times M \times M$ samples, with a total of $Q = M^3$ samples, and known protein content p, so that the protein is known to occupy $P = pQ$ samples. The location of these P samples is unknown, however. For a particular known envelope, the solution (electron density) lies on a P-dimensional hyperplane in R^Q (i.e. with the other $Q - P$ sample values fixed at zero). Furthermore, there are only a finite number of possible envelopes, i.e. there is a finite number of ways of selecting P samples from the Q samples. Under these conditions then, the object belongs to a P-dimensional



(a)

(b)

Figure 2.2 (a) A molecular envelope (left) and its corresponding autocorrelation (right) that fills the unit cell. (b) A molecular envelope (left) and one period of its corresponding Patterson function (right) that does not fill the unit cell. The case (b) requires a low occupancy molecule, that is unlikely to occur in practice.

subspace, or manifold, in R^Q, that is the union of ${}^QC_P$ P-dimensional hyperplanes (where ${}^QC_P$ denotes the number of combinations). If there are P diffraction data, then the dimensionality of the solution space will be reduced from P to zero, i.e. to a point set. An additional datum will likely select out the correct solution from this point set. The number of independent data is $Q/2$, so uniqueness requires that $Q/2 > P = pQ$, or $p < 0.5$, i.e. a protein content less that 50%, or a solvent content greater than 50%. The result is therefore the same as for a known envelope, and the constraint ratio is still given by equation (2.6). Since the real space constraint manifold is larger than a single P-dimensional hyperplane, the size of the point set may be be larger than for the known envelope case, but there is still a data excess of $(\frac{1}{2} - p)Q - 1$ when $p < 1/2$. We therefore conclude that the solution to the crystallographic phase problem with only knowledge that the crystal protein content, or volume, is less that 50% of the unit cell, is also unique. The increased size and complexity of the real space constraint manifold will likely make the solution more difficult to find, however, compared to the known envelope case.

2.3.2.1 Number of hyperplanes

It is interesting to consider the number of hyperplanes in the solution set. For each envelope shape there are Q possible positions of the envelope (including those that wrap

around the unit cell edges), and these should be treated as redundant since they all give the same Fourier amplitude. Therefore, the number of envelope-position-independent hyperplanes, denoted $\mathcal{N}_h(p, Q)$, is

$$\mathcal{N}_h(p, Q) = Q^{-1}\, {}^{Q}C_{pQ} \; . \tag{2.7}$$

Equation (2.7) can be approximated as follows. Since Q is large, applying Stirling's approximation to ${}^{Q}C_{pQ}$ gives

$$\mathcal{N}_h(p, Q) \approx \sqrt{\frac{2}{\pi}}\, \frac{1}{2\sqrt{p(1-p)}}\, Q^{-3/2}\, \left[p^{-p}(1-p)^{p-1}\right]^{Q} \; . \tag{2.8}$$

For fixed Q, $\mathcal{N}_h(p, Q)$ is symmetric about $p = 0.5$, where it is a maximum. At $p = 0.5$, equation (2.8) reduces to

$$\mathcal{N}_h(0.5, Q) \approx \sqrt{\frac{2}{\pi}}\, Q^{-3/2}\, 2^{Q}, \tag{2.9}$$

which is shown in Fig. 2.3 for $Q = 10^3$. In terms of the overall scale of $\mathcal{N}_h(p, Q)$, equation (2.9) gives an estimate of $\mathcal{N}_h(p, Q)$ for $0.3 < p < 0.7$ that is sufficient for our purposes.

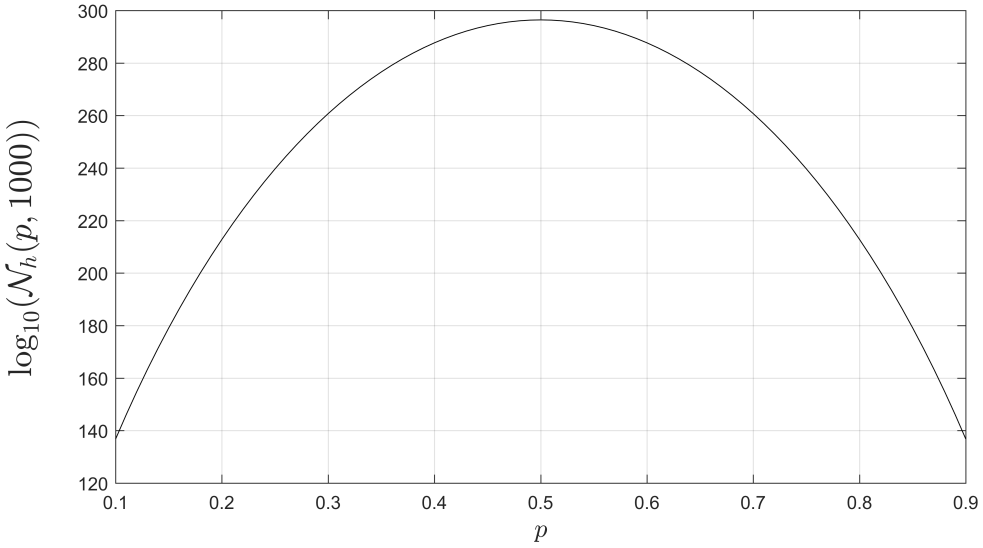


Figure 2.3 Number of hyperplanes in the solution set for $Q = 10^3$ as a function of the protein content p.

Protein crystal solvent contents between 70 and 30% (i.e. $0.3 < p < 0.7$), represent 95% of the entries in the PDB, so the approximation equation (2.9) is appropriate. The number of hyperplanes is large. For example, for $Q = 10^3$, equation (2.9) gives $\mathcal{N}_h(p, Q) \approx 10^{296}$! This number of hyperplanes emphasizes the extreme non-convexity

of the real space constraint set.

2.3.2.2 Simulations

Uniqueness for the case of an unknown envelope was investigated numerically by simulation. The idea is that, since there are multiple solutions to the problem there will be many such solutions, an effective reconstruction algorithm will find one of those solutions rather easily, demonstrating non-uniqueness. By setting up an IPA such as the DM algorithm with the appropriate constraints, the nature of the solution space can be examined by running the algorithm multiple times with different initial conditions. If multiple runs of the algorithm either converge to only the correct solution, or do not converge, then uniqueness is strongly supported. If the problem is not unique, then the algorithm will frequently converge rather quickly to an incorrect solution.

Since the unknown envelope constraint is rather weak and highly non-convex, the difficulty of the reconstruction problem is increased, increasing the number of iterations required for convergence, potentially to an impractically large value. This necessitates simulations with small objects. On the other hand, since the interest here is in uniqueness rather than reconstruction, nonconvergence is almost as informative as convergence.

A 2D unit cell was used for convenience (the same behaviour is expected in 3D since in the crystallographic case, Ω_c is independent of the dimensionality). In real space, the only constraints applied are the size of the envelope (i.e. the number of non-zero sample values) and positivity of the electron density. In reciprocal space, the constraint is to match the structure amplitudes of the true molecule. In addition to the usual positivity and Fourier amplitude projections [Millane and Lo, 2013], the projection for the envelope size is easily shown to consist of setting the $Q - P$ smallest density values to zero and leaving the other P values unchanged, at each iteration [Elser, 2003a]. The difference map algorithm was used with the DM parameter $\beta = 0.8$ as defined in Section 1.4.4 as a trade-off between navigating the search space and remaining in the attractor region.

A 29×29 sample unit cell was used and a single square "molecule" of various sizes was placed in the unit cell in P1 to vary the protein (or solvent) content, and thus vary Ω_c given by equation (2.6). The reconstruction algorithm was run for 10^6 iterations, starting with 10 different random molecules, for each molecule size. For each run, the solution was taken as that which gives the minimum mean-square error between the resulting structure amplitudes and the data. With an unknown support in P1, the structure amplitudes are insensitive to the absolute position of the support, and convergence of the algorithm can be slowed by "drifting" of the support. Therefore, the reconstruction was constrained to have its center of mass coincident with the center of mass of the true molecule.

The weak and highly non-convex real space volume constraint renders the search

for a solution difficult. The number of iterations needed is a function of the unit cell computation grid size and the number of hyperplanes such that the computation time becomes the main drawback in using the volume constraint. In fact the Fourier projection, despite being implemented using the FFT algorithm remains the slowest step. The DM algorithm, implemented in a GPU, was thus used for these simulations, allowing for faster computation and larger problem sizes. The GPU implementation use CUDA libraries made available by NVIDIA® [Nickolls et al., 2008], in particular, cuFFT was used for the fast Fourier transform and the Thrust library for parallelisable functions.

The results of the simulations are summarized in Table 2.1. The table shows the number of runs that converged and the number of correct reconstructions for the converged runs. For the converged runs, the mean-square error in reciprocal space approached very small values. The average number of iterations required in the converged cases is also shown in the table. Convergence was obtained for $\Omega_c > 1.4$ and $\Omega_c < 0.8$ in less than 10^6 iterations. However, for $0.8 < \Omega_c < 1.4$ the algorithm did not converge within 10^6 iterations. This is due to the weak and highly non-convex real space constraint, particularly for values of Ω_c close to unity, as mentioned above. Inspection of the table shows that in all cases for which $\Omega_c > 1$ ($p < 0.5$), the algorithm either converged to the correct solution (which therefore automatically had the correct envelope), or it did not converge. In no cases did it converge to an incorrect solution. This shows strong support for uniqueness in the case $\Omega_c > 1$. For $\Omega_c = 0.73$, multiple incorrect solutions were easily found by the algorithm. This indicates that, indeed, non-unique solutions are likely to be found if they exist. The results show that the structure amplitude data are able to select out the correct hyperplane corresponding to the solution, in spite of this large number.

Table 2.1 Summary of simulation results.

Object size	p	Ω_c	Runs converged	Correct solutions	Average iterations for convergence
15×15	0.27	1.87	10/10	10/10	4×10^4
16×16	0.30	1.64	5/10	5/5	1×10^5
17×17	0.34	1.46	1/10	1/1	8×10^5
24×24	0.68	0.73	10/10	0/10	1×10^4

The algorithm described above is useful for investigating uniqueness, but it is not a practical approach in protein crystallography where the number of sample values is much larger and many more iterations would be required. However, in practice, more is known about protein envelopes. In particular, protein envelopes are generally quite compact. This property substantially reduces the number of possible envelopes and the number of hyperplanes, significantly easing the reconstruction problem. Supplementing

the reconstruction algorithm with additional compactness constraints through the use of, for example, smoothing and shrinking of the support [Wang, 1985, Marchesini et al., 2003] or other schemes [Lo et al., 2009], should allow *ab initio* phasing without initial envelope information for practical problems. Indeed, the recent results of [He and Su, 2015] support this conclusion.

In summary then, even in the case where the molecular envelope is not known *a priori*, the macromolecular crystallographic phase problem has a unique solution if the protein content of the crystal is less than 50%.

2.3.3 Crystallographic symmetry

Consider now the effect of crystallographic (space group) symmetry on the constraint ratio. For non-centric crystallographic symmetry of order R, the Patterson function has symmetry of order $2R$ (as illustrated in Fig. 2.4(a)). Then, $|\mathcal{U}_u| = V/R$ and $|\mathcal{P}_u| = V/2R$, and substitution into equation (2.4) gives

$$\Omega_c = 1/2 \; , \tag{2.10}$$

i.e. the same as for the case without symmetry.

For centric crystallographic symmetry of order R, the Patterson function has symmetry of order R (as illustrated in Fig. 2.4(b)). In this case, $|\mathcal{P}_u| = V/R$, and substitution into equation (2.4) gives

$$\Omega_c = 1 \; . \tag{2.11}$$

This is then the marginal case that corresponds to a countable number of phase solutions (i.e. two choices for each reflection) and only a small amount of additional *a priori* information is required to render the solution unique.

These results are consistent with the well-known fact the reduction in the number of parameters due to the crystallographic symmetry is exactly matched by the same number of relationships between the structure amplitudes, and the overall data/parameter ratio remains unchanged. Crystallographic symmetry does not therefore constrain the phase problem, except in the centric case which does not occur with biomolecules.

2.3.4 Noncrystallographic symmetry

Consider now NCS of order R. NCS does not lead to increased symmetry in the Patterson function (see the illustration in Fig. 2.4(c)), so that $|\mathcal{U}_u| = V/R$ and $|\mathcal{P}_u| = V/2$, and substitution into equation (2.4) gives

$$\Omega_c = R/2 \; . \tag{2.12}$$

The redundancy of the phase problem is therefore improved by a factor R, and a unique solution is expected in principle if $R > 2$. Therefore, as a result of equation (2.12),

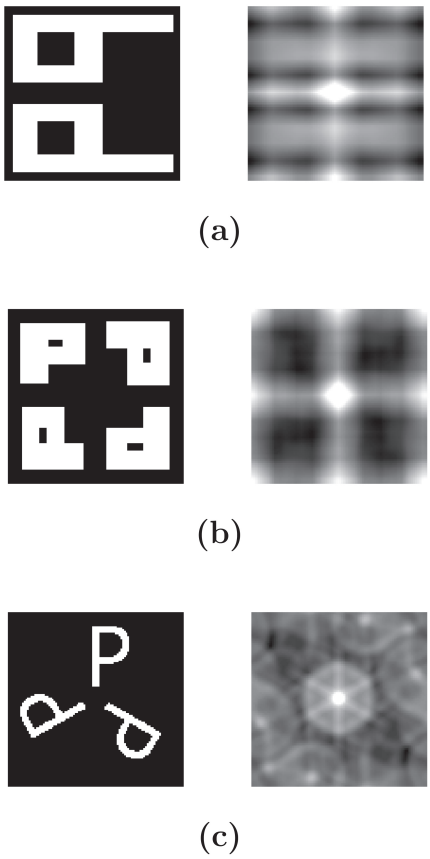


Figure 2.4 Examples of 2D unit cells (left) and one period of their corresponding Patterson functions (right) for (a) noncentric pm crystallographic symmetry, (b) centric p4 crystallographic symmetry, and (c) noncrystallographic 3-fold symmetry in plane group P1, as described in the text. The corresponding Patterson function symmetries are (a) p2mm, (b) p4, and (c) p2.

NCS is a significant factor for *ab initio* phasing. This result coincides with early considerations of the effect of NCS on constraining the phases [Crowther, 1969, Bricogne, 1974], and is related to the fact that NCS, unlike crystallographic symmetry, does not lead to relationships between the structure factor amplitudes, and so the number of independent data is not reduced. An alternative interpretation is that R-fold NCS leads to a denser sampling, by a factor R, relative to the Bragg sampling, of the continuous Fourier amplitude of the contents of the unit, increasing the number of data by a factor R [Millane, 1990, Millane, 1993].

As with the case of a known molecular envelope, the above analysis assumes that the NCS operators are known (so that the number of electron density parameters can be reduced by a factor R). This problem is not so difficult however, as the order of the NCS can be determined from a self-rotation function [Tong and Rossmann, 1997], although positioning of the NCS origin in the unit cell can present difficulties.

NCS is always accompanied by a restricted molecular envelope, and combining the above results gives

$$\Omega_c = \frac{R}{2p} \tag{2.13}$$

in the presence of both constraints. Therefore, with both constraints, solution to the phase problem is expected to be considerably eased. For example, with 2-fold NCS and 50% solvent content, or with 3-fold NCS and 25% solvent content, $\Omega_c = 2$ and the problem is expected to be well-determined in practice.

2.4 SUMMARY

As described previously, the constraining power of real space information in protein crystallography is conveniently characterised by a constraint ratio, which is useful in that it gives guidance on the likely success of *ab initio* phasing. Equation (2.4), utilising the Patterson function, is a new, rigorous, expression for the constraint ratio for a crystal. Properties of the Patterson function allow the constraint ratio to be reduced to the form equation (2.6) under most circumstances. The results also show that a volume constraint is as effective of an envelope constraint, in principle, although finding the solution is more difficult in the former case. Use of equation (2.4) allows a transparent derivation of known results for crystallographic and non-crystallographic symmetry

Recent results indicate that, as a result of errors and missing data, a value of Ω greater than about 1.5 might be needed for *ab initio* phasing in practice [Liu et al., 2012, Millane and Lo, 2013]. Equation (2.4) allows the constraint value to be calculated for specific kinds of real space information in order to make this assessment.

For the case of known protein content and NCS, the constraint ratio is given by equation (2.13). Evaluation of this equation suggests that, with the use of suitable reconstruction algorithms, *ab initio* phasing should be feasible with quite modest values of these parameters. Recent results using iterative projection algorithms indicate that

this is the case [Liu et al., 2012, He and Su, 2015, Lo et al., 2015]. NCS is a particularly powerful constraint if incorporated into iterative projection algorithms [Millane and Lo, 2013, Lo et al., 2015].

Although an estimate of the molecular envelope is desirable if available, uniqueness does not depend on *a priori* knowledge of the envelope, and envelope volume and compactness are a powerful constraint.

3 THE PHASE PROBLEM FOR TWO-DIMENSIONAL CRYSTALS. I. THEORY

Arnal, R. D and Millane, R. P. (2017). "The phase problem for two-dimensional crystals. I. Theory," *Acta Crystallographica A*, vol. 73, pp. 438-448.

Acta Cryst A FOUNDATIONS ADVANCES

ISSN 2053-2733

The phase problem for two-dimensional crystals. I. Theory

Romain D. Arnal and Rick P. Millane*

Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand. *Correspondence e-mail: rick.millane@canterbury.ac.nz

Received 8 May 2017
Accepted 24 September 2017

Edited by H. Schenk, University of Amsterdam, The Netherlands

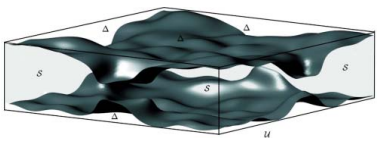
Keywords: phase problem; two-dimensional crystals; XFELs; *ab initio* phasing.

Properties of the phase problem for two-dimensional crystals are examined. This problem is relevant to protein structure determination using diffraction from two-dimensional crystals that has been proposed using new X-ray free-electron laser sources. The problem is shown to be better determined than for conventional three-dimensional crystallography, but there are still a large number of solutions in the absence of additional *a priori* information. Molecular envelope information reduces the size of the solution set, and for an envelope that deviates sufficiently from the unit cell a unique solution is possible. The effects of various molecular surface features and incomplete data on uniqueness and prospects for *ab initio* phasing are assessed. Simulations of phase retrieval for two-dimensional crystal data are described in the second paper in this series.

1. Introduction

The phase problem is of key importance in macromolecular crystallography. For *ab initio* phasing, *i.e.* in the absence of additional experimental information, such as from using isomorphous replacement or anomalous dispersion, the question of what real-space information is required to obtain a unique solution is of practical importance. Protein crystallography generally uses three-dimensional crystals and uniqueness properties of the phase problem in this case are well understood (*e.g.* Millane & Arnal, 2015). The problem is also well understood for isolated (*i.e.* non-crystalline) particles (Bates, 1982; Millane, 1990; Miao *et al.*, 1998) and for one-dimensional crystals (Millane, 2017). However, some macromolecular systems, notably membrane proteins, prefer to form two-dimensional crystals. These are not suitable for conventional crystallography due to their small size and weak scattering, but which have traditionally been used in cryo-electron crystallography (Kühlbrandt & Wang, 1991; Grigorieff *et al.*, 1996; Frank, 2006), and have recently been proposed for use with X-ray free-electron lasers (XFELs) (Frank *et al.*, 2014). In this paper we consider properties, particularly uniqueness properties, of the phase problem for a two-dimensional crystal. The results have particular significance for *ab initio* phasing in two-dimensional crystallography using XFELs. In a second paper (Arnal *et al.*, 2018) we illustrate the implications of the results obtained here using simulations of phase retrieval from two-dimensional crystal data.

The high intensity, small beam focus and short X-ray pulse duration of XFELs can potentially overcome the difficulties of two-dimensional crystallography with synchrotron sources (Frank *et al.*, 2014). The high pulse intensity allows measurable diffraction data to be obtained from small and thin two-dimensional crystals. The small beam focus allows exposure of single two-dimensional crystal grains, and the short pulse



duration allows the diffraction data to be collected before radiation damage occurs. Preliminary application of these ideas to two-dimensional protein crystals has demonstrated that good data in one projection can be obtained to 7 Å resolution (Frank *et al.*, 2014; Pedrini *et al.*, 2014). Avoiding the necessity for cryo-freezing presents significant advantages of this approach over cryo-electron microscopy. It allows studies at room temperature, under physiological conditions, and in two-dimensional crystals grown in a lipid bilayer that mimics their native environment. The nature of XFEL sources means that they are also eminently suitable for time-resolved studies. It is likely that improvements in sample preparation and instrumentation, including X-ray pulse brightness, beam size and pulse duration, will extend the resolution beyond the 7 Å reported in initial experiments.

In the case of two-dimensional cryo-electron crystallography, diffraction amplitudes are obtained using a tilt series and the phases may be obtained by molecular replacement if the structure of a related molecule is available. Alternatively, experimental phases may be obtained by Fourier transforming images obtained in a tilt series in the electron microscope. These phases are combined with the measured diffraction amplitudes and then refined to produce a high-resolution electron-density map. Hence, either a related molecule or micrographs of sufficient quality are required to solve the phase problem. An additional problem in electron microscopy is that, due to a limited range of accessible specimen tilts, there is a missing cone of diffraction data, as well as image data, about the axis normal to the crystal plane.

A potential advantage of crystallography with two-dimensional crystals is that the phase problem should be alleviated to some degree, relative to that of crystallography with three-dimensional crystals. This is because, since the specimen is only one unit-cell thick, the Fourier amplitude can in principle be measured effectively continuously in the corresponding direction in reciprocal space, as opposed to only at the reciprocal-lattice points (Bragg reflections). Since the two-dimensional crystal is periodic in the two transverse directions, the Fourier amplitude is sampled in the two corresponding directions. The result is that the amplitude can be measured along a set of one-dimensional lattice lines in reciprocal space. This increased sampling of the Fourier amplitude is expected to further constrain the phases compared with the three-dimensional crystal case. Similar data might also be obtainable from stacks of two-dimensional crystals that exhibit lateral translational disorder between the individual two-dimensional crystals in the stack.

A similar situation occurs with one-dimensional crystals, in which case there is continuous sampling of the Fourier amplitudes along lattice planes in Fourier space. This places considerable constraints on the phases, as recently shown by Millane (2017), and a unique solution to the *ab initio* problem is expected with fairly minimal additional *a priori* information.

The significance of continuous measurements along lattice lines and the effect of a finite thickness of the specimen in constraining the electron density were recognized quite early in electron crystallography. Stroud & Agard (1979) showed that the one-dimensional projected density consistent with the continuous diffraction amplitudes is restricted to a small set of solutions. Agard & Stroud (1982) showed that, for two-dimensional crystals, the amplitude and phase data could be extended into the missing cone by using a constraint on the specimen thickness. All of these approaches used a simple density-modification type of algorithm. A more sophisticated optimization algorithm has been applied to noisy, missing cone amplitude and phase data (Gipson *et al.*, 2011). However, these methods all require reasonably good initial phase estimates obtained by one of the methods described above.

It is well known that measurement of the continuous Fourier amplitude from an isolated object renders the solution to the phase problem unique in the absence of additional experimental data, and effective algorithms are available for reconstructing the object (Bates, 1984; Fienup, 1982; Elser, 2003; Marchesini, 2007). For three-dimensional crystals, however, the solution to the phase problem is highly non-unique, since the Fourier amplitudes are available only at the reciprocal-lattice points (Millane, 1990; Millane & Arnal, 2015). For one-dimensional crystals, the problem is highly constrained, although some weak additional information is required to obtain a unique solution (Millane, 2017). The phase problem for two-dimensional crystals therefore lies between a highly constrained case (one-dimensional crystal) and a highly under-constrained case (three-dimensional crystal). Uniqueness properties in the case of two-dimensional crystals, and the potential for *ab initio* phasing, are therefore not clear, and are the subject of this paper.

Ab initio phasing for two-dimensional crystals in the context of cryo-electron crystallography has been investigated by Spence *et al.* (2003). They attempted reconstructions of lysozyme at 3 Å resolution from two-dimensional crystal simulated diffraction data using the hybrid input–output (HIO) algorithm (Fienup, 1982). They found that *ab initio* phasing was not successful using the diffraction data alone, but it was successful if the diffraction data were supplemented by phases from sufficient images to fill in the reciprocal lattice with tilts between 0 and 15°. This represents an improvement on conventional phasing in electron crystallography, in terms of the experimental effort required, which generally requires an image tilt series up to about 60° to obtain sufficient phase information. For two-dimensional X-ray crystallography using XFEL sources, however, images are not available to provide initial phase estimates so that, in the absence of molecular replacement phases, the feasibility of *ab initio* phasing takes on more importance.

2. Background

Uniqueness properties of the phase problem for a single object, a three-dimensional crystal and a one-dimensional crystal have been well characterized, and the case of a two-dimensional crystal is the last remaining case to be examined for the class of crystalline objects. To put the latter case into context, it is therefore useful to briefly summarize uniqueness properties of the phase problem for the first three cases.

2.1. A single object

Consider an object with scattering density $f(\mathbf{x})$, where $\mathbf{x} = (x, y, z)$ is position in real space, that occupies a region $\mathbf{x} \in \mathcal{S}$ [*i.e.* $f(\mathbf{x}) = 0$ for $\mathbf{x} \notin \mathcal{S}$], which we refer to as the envelope or the support of $f(\mathbf{x})$. The measured diffracted amplitude from the object is equal to the amplitude $|F(\mathbf{u})|$ of the Fourier transform $F(\mathbf{u})$ of $f(\mathbf{x})$, where $\mathbf{u} = (u, v, w)$ is position in reciprocal space, and reconstruction of $f(\mathbf{x})$ from $|F(\mathbf{u})|$ constitutes the phase problem. A useful reconstruction is obtained only if $f(\mathbf{x})$ is uniquely related to $|F(\mathbf{u})|$ and thus uniqueness is of fundamental practical importance. Uniqueness of the phase problem for a single object has been well studied and it is known that, for an object of finite size, the problem has a unique solution, aside from some trivial ambiguities, in two or more dimensions, as long as the amplitude is measured continuously, or is sufficiently sampled, in reciprocal space (Bruck & Sodin, 1979; Bates, 1982, 1984; Barakat & Newsam, 1984; Millane, 1990; Miao *et al.*, 1998).

Uniqueness of the phase problem can be characterized by considering the ratio of the number of independent amplitude data that are available divided by the number of independent parameters describing the object, which we refer to as the constraint ratio, denoted Ω (Elser & Millane, 2008). A constraint ratio $\Omega > 1$ is a necessary condition for a unique solution. It is not a sufficient condition, but the number of multiple solutions is then severely restricted. If $\Omega = 1$ then the problem is marginal in the sense that multiple solutions exist but uniqueness is restored if a small amount of additional *a priori* information is available. If $\Omega < 1$ then the problem is highly non-unique and a multitude of objects are consistent with the Fourier amplitude data.

The constraint ratio for a single object can be expressed in terms of only the shape of the support region of the object, $\mathcal{S}$, as (Elser & Millane, 2008)

$$\Omega = \frac{|\mathcal{A}|}{2|\mathcal{S}|}, \tag{1}$$

where $\mathcal{A}$ is the support region of the autocorrelation of the object (or of $\mathcal{S}$) and $|\cdot|$ denotes the size (area or volume). Note that, for a real-valued object, the amplitude is centrosymmetric and so the number of independent amplitude data is proportional to $|\mathcal{A}|/2$ and the number of object parameters is proportional to $|\mathcal{S}|$. For a complex object, the number of data is proportional to $|\mathcal{A}|$ and the number of object parameters (real and imaginary parts) is proportional to $2|\mathcal{S}|$. The constraint ratio is given by equation (1) in both cases, and there is no distinction between real and complex objects.

The constraint ratio is bounded by $\Omega \geq 2^{N-1}$ for an object in N dimensions, and so the phase problem is better determined in higher dimensions for a single object (Elser & Millane, 2008). The problem is therefore well determined in two or three dimensions, where $\Omega \geq 2$ and $\Omega \geq 4$, respectively. For a one-dimensional object with connected support, $\Omega = 1$ and there is not a unique solution. The one-dimensional phase problem is relevant to the two-dimensional crystal case and is discussed in more detail in §3.

2.2. A three-dimensional crystal

For a three-dimensional crystal, uniqueness can again be evaluated using the constraint ratio. For a crystal, both the object and the autocorrelation are infinite in extent and the autocorrelation in equation (1) is replaced by the Patterson function. Taking into account any symmetry, the constraint ratio, denoted Ω_c, is then given by (Millane & Arnal, 2015)

$$\Omega_c = \frac{|\mathcal{P}_u|}{|\mathcal{U}_u|}, \tag{2}$$

where $\mathcal{U}_u$ and $\mathcal{P}_u$ denote the unique region of the unit cell and of the Patterson function, respectively. Note that the 2 in the denominator of equation (1) has been absorbed into $|\mathcal{P}_u|$ since the Patterson function is centrosymmetric. In general, in the absence of additional real-space information, this gives $\Omega = 1/2$ (or $\Omega = 1$ for centrosymmetric space groups) and the solution is highly non-unique. If the molecular envelope occupies a proportion p of the unit cell, then equation (2) reduces to

$$\Omega_c = \frac{1}{2p} \tag{3}$$

for non-centrosymmetric space groups (Millane & Arnal, 2015). A protein content p less than 50%, or a solvent content greater than 50%, then gives a unique solution. Interestingly, while the constraint ratio Ω for a single object depends on the shape of the object and not its volume, the constraint ratio Ω_c for a three-dimensional crystal depends on the volume (relative to the volume of the unit cell) of the molecule and not its shape.

2.3. A one-dimensional crystal

Properties of the phase problem for a one-dimensional crystal have recently been examined by Millane (2017), based in part on his earlier work on the phase problem in three dimensions (Millane, 1996). This gave a number of interesting results. While the constraint ratio for a one-dimensional crystal, denoted Ω_{1dc}, satisfies $\Omega_{1dc} \geq 2$, in the absence of other information, the solution is not unique, although it belongs to a low-dimensional set. Fairly minimal additional information, such as positivity or molecular envelope information, is expected to reduce this set to a single solution. For a restricted molecular envelope, the key requirement is that the envelope cross section varies with position along the crystal axis. A useful parameter that describes the constraining power of the molecular envelope, denoted there by Λ and denoted here by Λ_{1dc}, is given by

$$\Lambda_{1dc} = \frac{|\mathcal{C}|}{|\mathcal{S}|}, \tag{4}$$

where C denotes the support of the smallest circumscribing cylinder (of any cross-sectional shape, that need not be simply connected) that encloses the molecular envelope $\mathcal{S}$. If $\Lambda_{1dc} > 1$ then a unique solution is highly likely and the problem is more constrained for larger Λ_{1dc}. The parameter satisfies $\Lambda_{1dc} \geq 1$ and $\Lambda_{1dc} = 1$ for a cylindrical envelope (of any cross section).

The relationship between Λ_{1dc} and Ω_{1dc} is discussed by Millane (2017). In summary, the phase problem for a one-dimensional crystal is marginally constrained in general, but a unique solution is expected with minimal *a priori* information.

3. The one-dimensional phase problem

As mentioned in §1, the solution to the one-dimensional (non-crystalline) phase problem is highly non-unique. In §4 we analyse the two-dimensional crystal phase problem using a decomposition into one-dimensional phase problems. We need, therefore, in that analysis, to consider the degree and nature of the ambiguities of the one-dimensional phase problem. These characteristics have been studied extensively (Beinert & Plonka, 2015) and are summarized in this section.

It is convenient here to consider the *discrete* problem as this allows the ambiguities to be counted in a useful way. Consider an N-sample complex-valued signal $f[n] \in \mathbb{C}^N$, for $n \in 0, 1, \ldots, N-1$. We consider complex $f[n]$ since this is pertinent for our analysis in §4. Since we are considering a signal $f[n]$ of finite extent, or compact support, the definition of $f[n]$ is extended to $2N-1$ samples such that $f[n] = 0$ for $N \leq n < 2N-2$. We consider that we measure the amplitudes of the Fourier transform of $f[n]$ continuously in Fourier space. It can be shown that there are $2N-1$ degrees of freedom in the Fourier amplitude, and that a measurement of these is equivalent to measuring the amplitudes $|F[k]|$ of the discrete Fourier transform (DFT) of the length $2N-1$ sample signal (Beinert & Plonka, 2015, 2017). Note that, in the discrete case, for complex $f[n]$, $f[n]$ has $2N$ parameters but there are $2N-1$ data, so there are one fewer data than parameters in this case.

The general one-dimensional phase problem is subject to the usual trivial ambiguities of an unknown shift, an unknown constant phase factor and complex conjugate inversion in the origin. In the case considered here, since the support of $f[n]$ is restricted to $n \in (0, N-1)$, no unknown shift is allowed. The remaining trivial ambiguities are therefore an inversion, which, given the support of $f[n]$, takes the form $f[N-1-n]$ for $0 \leq n \leq N-1$, and an unknown constant phase factor denoted $\exp(i\varphi)$.

The one-dimensional phase problem, in general, has many ambiguities aside from the trivial ambiguities described above (Bruck & Sodin, 1979; Hayes *et al.*, 1980; Beinert & Plonka, 2015). This can be seen by writing the DFT $F[k]$ as a z transform, factorizing the resulting polynomial of order $N-1$ in z into $N-1$ linear factors, and noting that the intensity $|F[k]|^2$ is a polynomial of order $2N-2$ whose zeros occur in conjugate reciprocal pairs in the z plane. Each zero, or its conjugate reciprocal, of $|F[k]|^2$ corresponds to a zero of $F[k]$. Therefore, exchanging a zero of $F[k]$ with its conjugate reciprocal gives a different $f[n]$ with the Fourier amplitude $|F[k]|$ unchanged. Since there are $N-1$ zeros, there are 2^{N-1} possible signals $f[n]$ with the same Fourier amplitude $|F[k]|$. Since exchanging all the zeros with their conjugate reciprocal inverts $f[n]$, the trivial inversion ambiguity described above is included. In total, therefore, the solution set for the one-dimensional phase problem can be described as a set of 2^{N-1} one-dimensional manifolds, where each manifold represents the phase factor $\exp(i\varphi)$, parametrized by φ.

The number 2^{N-1} of one-dimensional manifolds is the *maximum* number, since if any zero pair lies on the unit circle then the zero and its conjugate reciprocal are coincident, and do not contribute an ambiguity. Furthermore, in a practical sense, if any zero pair is *close enough* to the unit circle, then the two ambiguous signals generated are sufficiently close to be essentially the same for practical purposes. This point is relevant since for structured (*i.e.* not noise-like) signals, the zeros tend to approach the unit circle as the corresponding value of k increases (this is analogous to the general decrease of the diffracted amplitude with resolution). Therefore, there can be a number of such zeros that do not contribute ambiguities. In summary, then, the set of solutions to the one-dimensional phase problem corresponds to 2^{N_1-1} one-dimensional manifolds where $N_1 \leq N$ is the number of zeros of the polynomial associated with $f[n]$ that are sufficiently distant from the unit circle. In some practical cases we may have that $N_1 \ll N$. We note that this is a very large number of solutions. In the following, we use the term 'solution' to mean 'one-dimensional manifold of solutions with unspecified phase factor', which should not cause confusion.

The number of solutions to the one-dimensional phase problem can be reduced by the presence of additional *a priori* information. Positivity of the signal can eliminate some of the multiple solutions, but in general the reduction in the number of solutions is not dramatic (Beinert, 2017). Knowledge of one or more of the samples of $f[n]$ can reduce the number of solutions dramatically (Xu *et al.*, 1987; Beinert & Plonka, 2017). In fact, knowledge of one sample, say $f[p] = C$, where $p \in (0, N-1)$ reduces the solution set to a single solution (or to two solutions if $p = 0, N/2, N-1$). Exceptions occur with probability zero, that are not considered here. However, special cases occur if $C = 0$ since then if $p = 0$ or $N-1$, the signal is reduced to length $N-1$ and the number of ambiguities is reduced only by a factor of 2. If $C = 0$ and $p \in (1, N-2)$, sometimes called the case of a *disconnected support*, then, almost always, there is a unique solution, or two solutions if $p = N/2$. Clearly, if there is more than one value of p on the interval $(1, N-2)$ where $f[p] = 0$ then there is a unique solution and the problem is even more constrained. In summary, then, the one-dimensional phase problem has many solutions, except in the case of a disconnected support where a single solution is expected, almost always. These results are used in our analysis of the two-dimensional crystal phase problem in §4.

Note that essentially analogous results apply to the *continuous* one-dimensional phase problem for a signal $f(x)$. In that case, the Fourier transform $F(u)$ can be extended into the complex plane $z = u + iv$ and factorized into linear factors. The intensity $|F(z)|^2$ is characterized by pairs of zeros that are reflected in the real axis, and a solution to the phase problem corresponds to selection of one zero from each pair. Since the zeros tend to approach the real axis with increasing u, there are only a finite number N of zeros with significant

imaginary part, and the number of solutions is again, effectively, 2^{N-1}.

4. The two-dimensional crystal phase problem

Consider now the case of a two-dimensional crystal. We consider, for simplicity, a two-dimensional crystal with a rectangular unit cell in space group $P1$ with unit-cell dimensions a and b (Fig. 1). The results apply straightforwardly to other two-dimensional crystal classes, and Millane & Arnal (2015) show that space-group symmetry does not affect uniqueness of the phase problem, except for the case of centric space groups. Since the specimen is one unit-cell thick, there is strictly no unit-cell dimension in the z direction, but we denote by c the maximum thickness of the protein or molecular assembly in the z direction (Fig. 1).

We denote the electron density in one unit cell of the two-dimensional crystal as $f(x, y, z)$. In order to represent the case of a two-dimensional crystal, and for convenience, we consider the discrete version, denoted $f[m, n, p]$, of $M \times M \times (2N-1)$ samples, with $m,\ n \in (0, M-1)$ and $p \in (0, 2N-2)$, and $f[m, n, p] = 0$ for $p \in (N, 2N-2)$. This allows us to make use of the results of §3. The DFT of $f[m, n, p]$ is denoted $F[r, s, t]$.

Consider the two-dimensional Fourier transform of $f[m, n, p]$ in the m and n directions, which we denote by $\tilde{f}_{rs}[p]$, *i.e.*

$$\tilde{f}_{rs}[p] = \sum_{m=0}^{M-1}\sum_{n=0}^{M-1} f[m, n, p]\exp[i2\pi(rm + sn)/M]. \qquad (5)$$

Note that, in general, $\tilde{f}_{rs}[p]$ will be complex, even if $f[m, n, p]$ is real. The one-dimensional Fourier transform of $\tilde{f}_{rs}[p]$ with respect to p is $F[r, s, t]$. Since $|F[r, s, t]|^2$ is measured at its Nyquist spacing in t, for fixed (r, s) we have a one-dimensional (non-crystalline) phase problem for $\tilde{f}_{rs}[p]$ whose solution is determined within the set of ambiguities described in §3. These one-dimensional phase problems are independent for different (r, s) since $|F[r, s, t]|^2$ is sampled at twice its Nyquist spacing in r and s. We denote the set of the (many) possible solutions of these one-dimensional phase problems by $\{\hat{f}_{rs}[p]\}$.

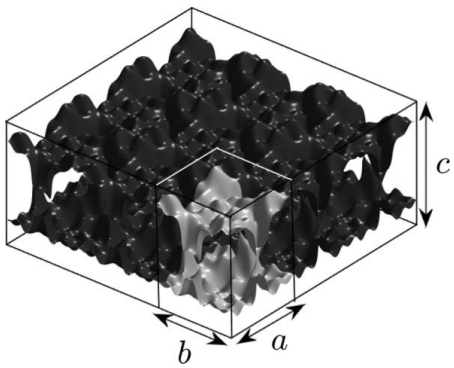


Figure 1
A two-dimensional crystal with unit-cell dimensions a and b, and maximum thickness c.

The possible solutions to the two-dimensional crystal phase problem, denoted $\hat{f}[m, n, p]$, are then given by

$$\hat{f}[m, n, p] = \sum_{r=0}^{M-1}\sum_{s=0}^{M-1} \hat{f}_{rs}[p]\exp[-i2\pi(rm + sn)/M]. \qquad (6)$$

Since there are many solutions $\hat{f}_{rs}[p]$, there are many solutions $\hat{f}[m, n, p]$. The correct solution is obtained only in the (very unlikely) event that $\hat{f}_{rs}[p] = \tilde{f}_{rs}[p]$, for all r and s. The solution to the two-dimensional crystal phase problem is therefore highly non-unique.

The number of solutions $\hat{f}[m, n, p]$ can be counted as follows. Referring to §3, each $\hat{f}_{rs}[p]$ is restricted to a set of 2^{N-1} one-dimensional manifolds. A consideration of the topology of the solution set for $\hat{f}[m, n, p]$ shows that it corresponds to Q (M^2-dimensional manifolds), where

$$Q = 2^{(N-1)M^2}. \qquad (7)$$

Note that the number of manifolds Q is of secondary importance, relative to the manifold dimensionality, but adds to the topological complexity of the solution manifold. The solution manifold can be considered a single, highly complex, M^2-dimensional manifold. Referring to the discussion in §3, the number of manifolds may be reduced to

$$Q = 2^{\left(\sum_{r,s} N_{rs}\right) - M^2}, \qquad (8)$$

where N_{rs} denotes the number of zeros of the polynomial for $f_{rs}[p]$ that are sufficiently distant from the unit circle, although the solution manifold dimensionality remains M^2.

The above results can be extended to the continuous, finite-resolution, case as follows. Letting the resolution of the diffraction data be d, the number of significant zeros is approximately $N_{rs} \simeq c/d$, and M is approximately $M \simeq a/d$, and substitution into equation (7) gives

$$Q \simeq 2^{(c/d-1)(a/d)^2} \simeq 2^{V/d^3}, \qquad (9)$$

since $c/d \gg 1$, and V is the volume of the unit cell. The number of solutions to the two-dimensional crystal phase problem is therefore very large. For example, for a crystal with $a = 100$, $c = 30$ Å and 3 Å resolution data, the solution set corresponds to $\sim 10^{3000}$ ($\sim$ 1000-dimensional manifolds), or a highly complex $\sim$ 1000-dimensional manifold. Keep in mind, however, that at this point we are considering the case with no additional *a priori* information.

In general, $\hat{f}[m, n, p]$ will be complex and so the solutions given by equation (6) belong to an M^2-dimensional manifold in $\mathbb{C}^{M^2N}$. Since the electron density is real, it is useful to consider only the real solutions. Referring to equation (6), if the values $\hat{f}_{rs}[p]$ are restricted such that $\hat{f}_{rs}[p] = \hat{f}^*_{M-1-r,M-1-s}[p]$, then $\hat{f}[m, n, p]$ is real. We assume that $\hat{f}_{rs}[p]$ has been so restricted, and then the solutions $\hat{f}[m, n, p]$ belong to an M^2-dimensional manifold, denoted E_1, in $\mathbb{R}^{M^2N}$.

An example of multiple solutions is generated as follows and shown in Fig. 2. A positive, $7 \times 7 \times 7$ sample object $f[m, n, p]$ was generated with random values uniformly distributed on (0, 1) and is shown in the left column of Fig. 2. An object $\hat{f}[m, n, p]$ is generated from $f[m, n, p]$ by calculating

the $f_{rs}[p]$ and the corresponding polynomial for each (r, s), randomly exchanging each zero of the polynomial with its conjugate reciprocal, recomputing $\hat{f}_{rs}[p]$ and assembling $\hat{f}[m, n, p]$. A constant phase shift was not applied to each $\hat{f}_{rs}[p]$ in this example. Such an object is shown in the second column of Fig. 2. The Fourier amplitudes of $f[m, n, p]$ and $\hat{f}[m, n, p]$ calculated by the DFT are identical, and are shown in the right columns of Fig. 2, illustrating the non-uniqueness.

Since the solution to the generic two-dimensional crystal phase problem is highly non-unique, an immediate question is: are there other constraints available that might reduce the ambiguities to a small number? Various constraints may be available in practice, and we investigate the effect of positivity and molecular envelope constraints in the next two subsections.

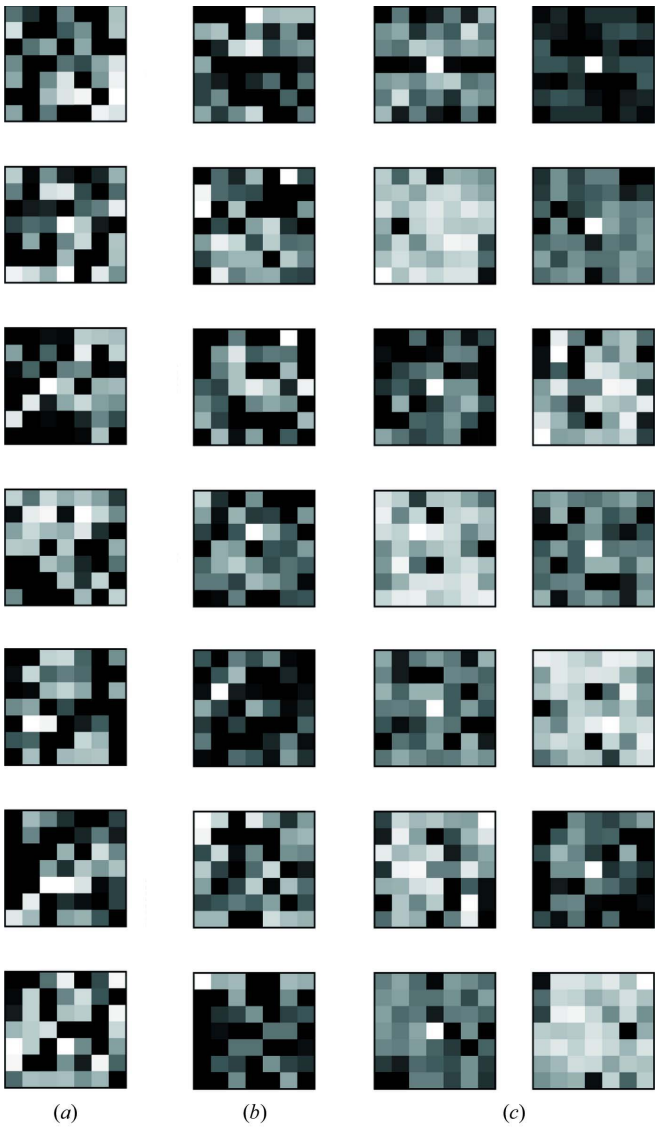


Figure 2
(*a*) A positive $7 \times 7 \times 7$ object $f[m, n, p]$ with each panel representing one value of p. (*b*) An object $\hat{f}[m, n, p]$ generated from $f[m, n, p]$ by exchanging zeros of the z transform as described in the text. (*c*) The amplitudes of the DFT $F[r, s, t]$ of both $f[m, n, p]$ and $\hat{f}[m, n, p]$. Since the two-dimensional crystal has finite thickness in the p direction, the DFT amplitude is shown for 14 values of t.

4.1. Positivity constraint

For a positive electron density, the relevant question is how many of the large number of possible solutions to the phase problem are positive? A positivity constraint restricts the solution to belonging to the M^2N-dimensional positive orthant $f[m, n, p] \geq 0$, which we denote $\mathbb{R}^{M^2N+}$. The solution manifold, denoted E_1^+, is then $E_1^+ = E_1 \cap \mathbb{R}^{M^2N+}$. Although, in general, this reduces the number of solutions by a factor 2^{M^2N}, the dimensionality of E_1^+ remains at M^2. Therefore, although a positivity constraint reduces the number of solutions, it does not reduce the dimensionality of the solution set, and so does not significantly reduce the ambiguity of the solution to the two-dimensional crystal phase problem.

4.2. Envelope constraint

We consider now the case where some information (generally at low resolution) is available on the molecular envelope. For the case of membrane proteins in particular, atomic force microscopy has been used to define molecular boundaries in two-dimensional crystals at resolutions of up to 5–10 Å (Frederix *et al.*, 2009). Considering the discrete case and taking N to be the maximum number of samples spanning the electron density in the p direction, let $\mathcal{U}$ denote the region of the discrete unit cell, *i.e.* $\mathcal{U} = \{0 \leq m \leq M-1, 0 \leq n \leq M-1, 0 \leq p \leq N-1\}$. Let $\mathcal{S}$ be the region of the molecular envelope and Δ the region inside $\mathcal{U}$ but outside $\mathcal{S}$, so that $\mathcal{U} = \mathcal{S} \cup \Delta$ (Fig. 3). The envelope constraint then corresponds to the condition $f[m, n, p] = 0$ for $(m, n, p) \in \Delta$. In general, many of the solutions to the phase problem will not satisfy this condition and thus the envelope constraint reduces the number of valid solutions. In general, the larger is the region Δ, the greater is the restriction on the solution set.

The envelope constrains certain samples of $f[m, n, p]$ to be zero. If there are P such samples (*i.e.* $|\Delta| = P$), then $f[m, n, p]$ is restricted to a manifold (in fact a hyperplane), denoted H_P, of dimension $(M^2N - P)$ in $\mathbb{R}^{M^2N}$. The set (manifold) of solutions to the phase problem that satisfy the envelope constraint, denoted E_2, is then $E_2 = E_1 \cap H_P$, which has dimension $(M^2 - P)$ in $\mathbb{R}^{M^2N}$. Therefore, if $P = M^2$ then the solution manifold is a point set, and if $P > M^2$ then a unique

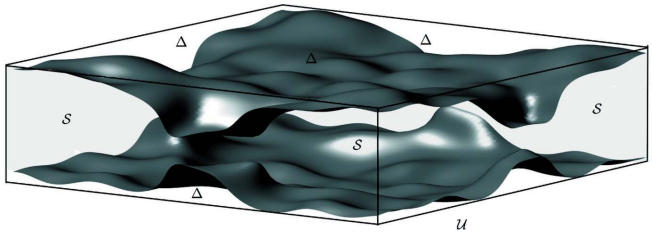


Figure 3
Illustration of the molecular envelope $\mathcal{S}$ and the unit cell $\mathcal{U}$ with the upper and lower surfaces bounding $\mathcal{S}$. The region Δ is the region inside $\mathcal{U}$ but outside $\mathcal{S}$.

solution is expected (*i.e.* it is unlikely, except in contrived cases, that the subspace defined by the additional constraint will pass through more than one of the points). While this is not a rigorous proof of uniqueness, considering in particular the highly complex nature of the manifold involved, it indicates that for $P > M^2$ zero samples of $f[m, n, p]$, the solution to the phase problem is likely to be highly constrained. Although positivity on its own is a weak constraint, it is likely to be more effective in the presence of a strong envelope constraint.

It is worth noting that the improved uniqueness of the one-dimensional phase problem for the case of a disconnected support as described in §3 does not appear to be helpful for the two-dimensional crystal phase problem. There are three reasons for this. First, application to the two-dimensional crystal problem would require that the $\tilde{f}_{rs}[p]$ have a disconnected support, and it is difficult to imagine a molecular envelope that would produce this. The only obvious case where this would occur is if there were a slab parallel to the crystal plane and interior to the unit cell that is devoid of electron density (or is solvent). However, this would imply that the molecule consists of two completely disconnected parts, which is not feasible in practice. Second, such a slab would contain at least M^2 zero samples, which is no fewer than the number of zero samples required for the envelope constraint case described above. Third, although this reduces the number of manifolds, the dimensionality of the solution manifold remains M^2.

For the case of a one-dimensional crystal, the quantity $\Lambda_{1\mathrm{dc}}$ defined by equation (4) is useful for assessing the effect of a molecular envelope on uniqueness for that problem, in the sense that the solution is expected to be highly constrained if $\Lambda_{1\mathrm{dc}} > 1$ (Millane, 2017). We define here an analogous quantity, denoted $\Lambda_{2\mathrm{dc}}$, for the two-dimensional crystal phase problem. The quantity exactly analogous to $\Lambda_{1\mathrm{dc}}$ is $|\tilde{\mathcal{U}}|/|\mathcal{S}|$, where $\tilde{\mathcal{U}}$ is the smallest union of slabs parallel to the crystal plane that bounds the molecule. However, as noted above, because the molecule or assembly must form a connected object in practice, the set $\tilde{\mathcal{U}}$ will always reduce to a single slab that is the unit cell $\mathcal{U}$. We therefore define $\Lambda_{2\mathrm{dc}}$ as

$$\Lambda_{2\mathrm{dc}} = \frac{|\mathcal{U}|}{|\mathcal{S}|}, \tag{10}$$

and it is easily seen that $\Lambda_{2\mathrm{dc}} \geq 1$.

For the discrete case, as described above, uniqueness requires that $P = |\Delta| > M^2$. Using this requirement and equation (10), and noting that $|\mathcal{U}| = M^2N$ and $|\mathcal{S}| = M^2N - |\Delta|$, it is easily shown that the necessary condition for uniqueness is

$$\Lambda_{2\mathrm{dc}} > \left(1 - \frac{1}{N}\right)^{-1}. \tag{11}$$

It is therefore convenient to define the quantity $\Lambda'_{2\mathrm{dc}}$ by

$$\Lambda'_{2\mathrm{dc}} = \left(1 - \frac{1}{N}\right)\frac{|\mathcal{U}|}{|\mathcal{S}|} \tag{12}$$

and the necessary condition for uniqueness is then $\Lambda'_{2\mathrm{dc}} > 1$.

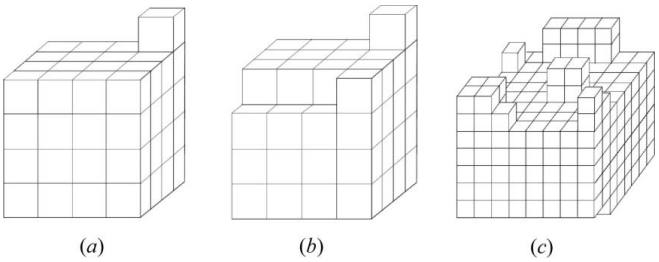


Figure 4
Examples of some simple discrete envelopes with corresponding values of $\Lambda'_{2\mathrm{dc}}$: (*a*) 0.98, (*b*) 1.03 and (*c*) 1.14.

For the continuous case, following the same reasoning as above gives

$$\Lambda'_{2\mathrm{dc}} = \left(1 - \frac{d}{c}\right)\frac{|\mathcal{U}|}{|\mathcal{S}|} \tag{13}$$

and, again, $\Lambda'_{2\mathrm{dc}} > 1$ is a necessary condition for uniqueness. Clearly $\Lambda'_{2\mathrm{dc}} < \Lambda_{2\mathrm{dc}}$, and $\Lambda'_{2\mathrm{dc}} \rightarrow \Lambda_{2\mathrm{dc}}$ as $d \rightarrow 0$, *i.e.* at very high resolution. The condition $\Lambda'_{2\mathrm{dc}} > 1$ means that the molecular envelope must be *sufficiently structured*. Some simple discrete examples and the corresponding value of $\Lambda'_{2\mathrm{dc}}$ are shown in Fig. 4.

A simple interpretation of the condition $\Lambda'_{2\mathrm{dc}} > 1$ can be derived as follows. Let v be the fraction of the unit cell occupied by solvent, *i.e.* $v = \Delta/|\mathcal{U}|$, and let $\alpha = d/c$, *i.e.* the ratio of the resolution to the thickness of the unit cell. Substitution into equation (13) shows that $\Lambda'_{2\mathrm{dc}} = v/\alpha$, so that uniqueness requires that

$$v > \alpha. \tag{14}$$

Equation (14) then gives a necessary condition for uniqueness, *i.e.* the solvent content must be greater than the normalized resolution. Although not a sufficient condition, the solution in this case is highly constrained and uniqueness can be considered likely.

Although in theory $\Lambda'_{2\mathrm{dc}} > 1$ is sufficient for uniqueness, in reality a margin will be necessary in order to successfully reconstruct the electron density from the diffraction amplitudes, particularly in practical cases where there are data missing and the data are noisy. For example, for the examples considered by Millane & Chen (2015), in terms of the constraint ratio, although $\Omega > 1$ is theoretically needed for uniqueness, it was found that $\Omega > 1.2$ is a more realistic requirement, even in the absence of errors in the data. For noisy, incomplete data, $\Omega > 1.5$ might be more realistic (Millane & Lo, 2013). In a similar fashion, a margin over $\Lambda'_{2\mathrm{dc}} > 1$ will be required in practice, which will be influenced by data errors and completeness.

In summary, the phase problem for a two-dimensional crystal is underconstrained in general, but is highly constrained in the presence of additional *a priori* information if $\Lambda'_{2\mathrm{dc}} > 1$. In contrast to the three-dimensional crystal case where greater than 50% solvent is required for uniqueness (Millane & Arnal, 2015), a smaller solvent content is sufficient in the two-dimensional crystal case. The presence of other

constraints such as non-crystallographic symmetry or histogram information will further constrain the solution. Therefore, while the *ab initio* problem for a two-dimensional crystal does not have a unique solution in general, *ab initio* phasing may be feasible in favourable circumstances with fairly modest *a priori* information.

5. The constraint ratio

As described in §2, the effect of a molecular envelope constraint on uniqueness of the phase problem for a single object or a three-dimensional crystal can be usefully quantified by the constraint ratio. It is useful to examine the constraint ratio for the case of a two-dimensional crystal.

The constraint ratio for a two-dimensional crystal is given in equation (1) with $\mathcal{A}$ replaced by one period of the Patterson function of the two-dimensional crystal, denoted $\mathcal{P}_{2dc}$. The constraint ratio for the two-dimensional crystal, denoted Ω_{2dc}, is then given by

$$\Omega_{2dc} = \frac{|\mathcal{P}_{2dc}|}{2|\mathcal{S}|}. \qquad (15)$$

For a two-dimensional crystal, $\mathcal{P}_{2dc}$ takes the form of a Patterson function in the x and y directions and an autocorrelation in the z direction. The Patterson function is equal to the sum of periodically repeated autocorrelations (Millane, 1990), and the calculation of $\mathcal{P}_{2dc}$ is aided by writing it in the form

$$\mathcal{P}_{2dc} = (\mathcal{A}_{00} \cup \mathcal{A}_{01} \cup \mathcal{A}_{10} \cup \mathcal{A}_{11}) \cap [(0, a) \times (0, b)], \qquad (16)$$

where $\mathcal{A}_{mn}$ denotes the support of the autocorrelation of one unit cell shifted by ma and nb in the x and y directions, respectively, and $[(0, a) \times (0, b)]$ denotes the region of the projected unit cell.

The constraint ratio Ω_{2dc} can be bounded as follows. For a single three-dimensional object (§2.1) we have that $\Omega \geq 4$ which implies that $|\mathcal{A}| \geq 8|\mathcal{S}|$. Referring to equation (16) shows that $|\mathcal{P}_{2dc}| \geq |\mathcal{A}|/4$, and substituting into equation (15) shows that $\Omega_{2dc} \geq 1$. We also have that $|\mathcal{P}_{2dc}| \leq 2|\mathcal{U}|$, so that substituting into equation (15) and using equation (10) shows that $\Omega_{2dc} \leq \Lambda_{2dc}$. In summary, then, Ω_{2dc} is bounded as

$$1 \leq \Omega_{2dc} \leq \Lambda_{2dc}. \qquad (17)$$

Referring to equation (17), since $\Lambda_{2dc} \geq \Omega_{2dc}$, an increasing constraint ratio is helpful in terms of uniqueness, but $\Omega_{2dc} > 1$ is not a useful condition as uniqueness requires $\Lambda'_{2dc} \geq 1$, and $\Lambda'_{2dc} < \Lambda_{2dc}$, *i.e.* it is possible that $\Omega_{2dc} > 1$ but $\Lambda'_{2dc} < 1$. In general, then, Λ'_{2dc} is a more useful metric than Ω_{2dc}.

We note, for completeness, that the bound $\Omega_{2dc} \leq \Lambda_{2dc}$ is analogous to the bound $\Lambda_{1dc} \leq |\mathcal{C}_c|/|\mathcal{S}|$, described by Millane (2017) for the one-dimensional crystal case, where $\mathcal{C}_c$ is the smallest, convex, centrosymmetric cylinder that circumscribes the molecule. Note that, referring to §4.2, and the definition of Λ_{2dc} in equation (10), the unit cell $\mathcal{U}$ in the two-dimensional crystal case is convex and centrosymmetric and takes the place of $\mathcal{C}_c$ in the one-dimensional crystal case.

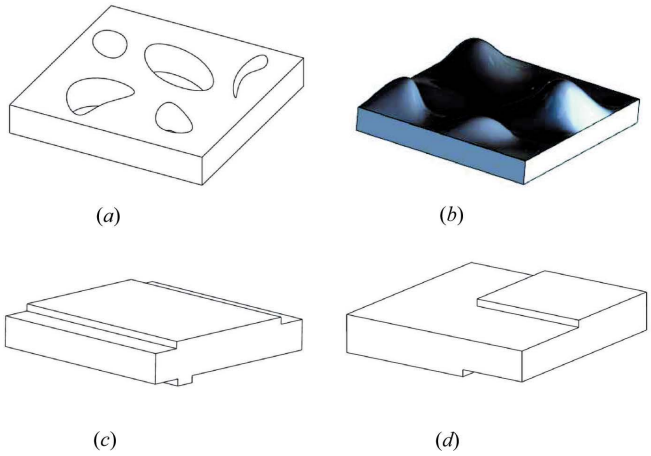


Figure 5
(*a*) A unit cell with flat surfaces and a number of holes (pores). (*b*), (*c*), (*d*) Examples of envelopes for which $\Omega_{2dc} = \Lambda_{2dc}$.

6. Examples

The theory of uniqueness for the two-dimensional crystal phase problem described above, and the quantities Λ_{2dc}, Λ'_{2dc} and Ω_{2dc} are illustrated here for a number of simple example envelopes.

First, if the unit cell is 'full', or we have no information on the protein envelope, then $\Lambda_{2dc} = \Omega_{2dc} = 1$, and multiple solutions are expected.

Second, consider the case where the upper and lower molecular surfaces are flat except for a 'hole' (or number of holes) that passes through the crystal, as shown in Fig. 5(*a*). This could correspond, for example, to pores in a membrane protein. In this case, it is easily seen that $|\mathcal{P}_{2dc}| = 2|\mathcal{U}|$, so that $\Omega_{2dc} = \Lambda_{2dc}$. Uniqueness then requires that the volume of the pore satisfies equation (14), *i.e.* a sufficiently large pore will force uniqueness.

Third, consider the case of an envelope with one flat surface and an unrestricted variation on the other surface, as shown in Fig. 5(*b*). Using equation (16), it can be shown that, in this case also, $|\mathcal{P}_{2dc}| = 2|\mathcal{U}|$ so that, again, $\Omega_{2dc} = \Lambda_{2dc}$. For a specific example of this case, consider a flat molecular envelope of uniform thickness t, with a square unit cell and a small cuboidal feature on one surface as shown in Fig. 6(*a*). The surface feature has a width that is a fraction w of the lateral unit-cell dimension a and a height that is a fraction h of the thickness t. In this case, then, $c = t(1 + h)$. Calculation of Λ_{2dc} gives

$$\Lambda_{2dc}(h, w) = \Omega_{2dc}(h, w) = \frac{1 + h}{1 + hw^2}, \quad 0 < w < 1, \qquad (18)$$

which is plotted *versus* h and w in Fig. 6(*b*). The surface feature increases Λ_{2dc} and Ω_{2dc} to a value greater than unity. Inspection of the figure shows that the effect of the feature increases as it becomes narrower (small, but nonzero w) or taller (large h). Note that $h = 0$ or $w = 0$ or $w = 1$ corresponds to no feature and $\Lambda_{2dc}(0, w) = \Lambda_{2dc}(h, 0) = \Lambda_{2dc}(h, 1) = 1$. There is however a discontinuity at $w = 0$.

The figure shows that the effect of the feature on Λ_{2dc} can be significant, although uniqueness depends on the size of the feature and the resolution of the data.

Numerous other examples can be generated that give $\Omega_{2dc} = \Lambda_{2dc}$. The key requirement is that the autocorrelation of one unit cell has a flat region of maximum thickness that has extent $a \times b$. Two other such examples are shown in Fig. 5.

Consider now an analogous case to that above except that the cuboid feature is on both sides of the slab, with each feature having dimensions $w \times w \times h/2$, as shown in Fig. 6(*c*). Using equation (16) shows that $|\mathcal{P}_{2dc}| < 2|\mathcal{U}|$, so that in this case $\Omega_{2dc} < \Lambda_{2dc}$. The value of Λ_{2dc} is still given by equation (18), but analysis of this case shows that the constraint ratio is given by

$$\Omega_{2dc} = \frac{1 + h/2 + 2hw^2}{1 + hw^2}, \quad 0 < w < 1/2, \tag{19}$$

so that, indeed, $\Omega_{2dc} < \Lambda_{2dc}$. The constraint ratio is plotted in Fig. 6(*d*) *versus* h and w, and comparison with Λ_{2dc}, which is shown in Fig. 6(*b*), shows that it is smaller. In general, large, narrow excursions from a flat molecular envelope are required to reduce Ω_{2dc} significantly from Λ_{2dc}. This is unlikely in general for membrane systems, so that in practice Ω_{2dc} and Λ_{2dc} are not expected to be too different.

We now consider a more realistic protein envelope example, for which we use the envelope of the membrane protein aquaporin 1 (AQP1) (Ren *et al.*, 2000). The molecular envelope is essentially a thresholded, low-resolution version of the electron density that identifies the outer boundary of the molecule. A molecular boundary would typically be defined at a resolution of about 10 Å. As the resolution of the envelope is increased, it becomes more intricate and will tend to generate more solvent regions and will thus modestly increase Λ_{2dc}. Because of the smoothing effect of autocorrelation, $|\mathcal{S}|$ will tend to decrease more than $|\mathcal{A}|$ and Ω_{2dc}

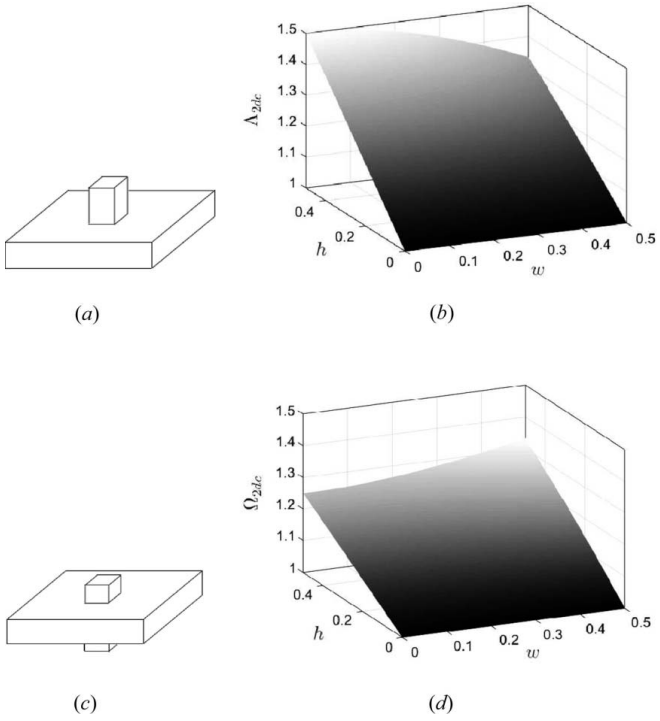


Figure 6
(*a*) A unit cell with a cuboidal surface feature and (*b*) the corresponding values of $\Lambda_{2dc} = \Omega_{2dc}$. (*c*) A unit cell with two cuboidal surface features and (*d*) the corresponding values of Ω_{2dc}. The corresponding values of Λ_{2dc} are the same as in (*b*).

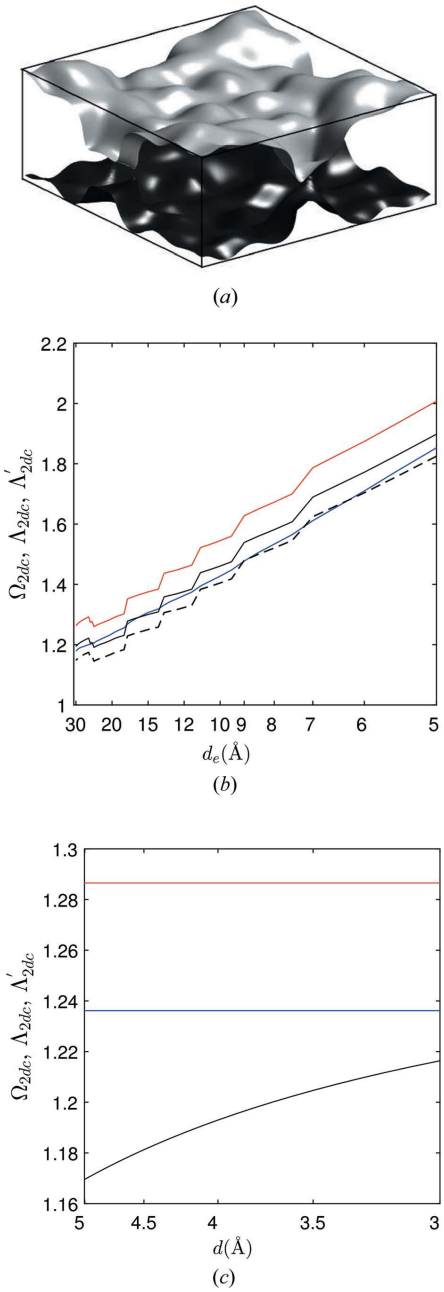


Figure 7
(*a*) The aquaporin 1 envelope at a resolution $d_e = 10$ Å, calculated as described in the text. (*b*) The parameters Λ_{2dc} (red) and Ω_{2dc} (blue), and Λ'_{2dc} for resolution $d = 3$ Å (black) and $d = 5$ Å (black dashed), *versus* the envelope resolution d_e. (*c*) The parameters Λ_{2dc} (red), Ω_{2dc} (blue) and Λ'_{2dc} (black) *versus* the resolution of the data d for an envelope resolution $d_e = 20$ Å.

will tend to increase more than Λ_{2dc}. The parameters Λ_{2dc} and Ω_{2dc} therefore depend on the resolution of the envelope. Note that, at high enough resolution, the envelope tends to define the shape of the molecule and the phase problem is of less significance. Furthermore, as described above, $\Lambda'_{2dc} > 1$ also depends on the resolution of the data which will generally be higher than the resolution of the envelope.

The following procedure was used to calculate an envelope of AQP1 at a variable resolution. First, the electron density was thresholded at its mean value and values below the mean set to zero. This has the effect of removing spurious noise peaks and small disconnected regions of the density. Second, the resulting density was convolved with a spherically symmetric three-dimensional Gaussian function with a full width at half-maximum (FWHM) set to a value, denoted d_e, that we refer to as the resolution of the envelope. Third, the smoothed density was thresholded at its mean plus one standard deviation, and the grid points below the threshold value set to zero and those above to unity. The resulting binary function then represents the envelope at that resolution. The envelope calculated in this way at a resolution of 10 Å is shown in Fig. 7(*a*). This is a reasonable representation of the envelope and the undulations on the surface due to the molecular structure are evident.

The aquaporin unit cell has dimensions 100 × 100 Å and a thickness of 55 Å. The envelope was first calculated for resolutions 30 Å $< d_e <$ 5 Å, and Λ_{2dc} and Ω_{2dc} calculated numerically as a function of d_e. The results are shown in Fig. 7(*b*). Inspection of the figure shows that both parameters increase as the resolution of the envelope increases, as anticipated, and $\Omega_{2dc} < \Lambda_{2dc}$. The parameter Λ'_{2dc} is also plotted in Fig. 7(*b*) for resolutions of the data, d, of 5 and 3 Å. Note that $\Lambda'_{2dc} < \Lambda_{2dc}$ and that Λ'_{2dc} is larger for higher resolution. Also, Λ'_{2dc} falls below Ω_{2dc} at low resolution. In this case, $\Lambda'_{2dc} > 1$, but successful phasing would depend on the resolution and the accuracy of the data.

To show the effect of the resolution of the data, for an envelope of resolution $d_e = 10$ Å, Λ'_{2dc} is calculated for 3 Å $< d <$ 5 Å and shown in Fig. 7(*c*) as a function of d, and compared with Λ_{2dc} and Ω_{2dc}. Whereas Λ_{2dc} and Ω_{2dc} are constant, Λ'_{2dc} increases with increasing resolution. For a resolution greater than 3.75 Å, $\Lambda'_{2dc} > 1.2$, which might be sufficient for a unique reconstruction in practice, for example.

7. Incomplete data

In this section we consider briefly the expected effects of incomplete data on uniqueness.

In cryo-electron crystallography, diffraction data can be collected for specimen tilts θ only up to a maximum value of typically about 60°. This results in a cone of missing data in reciprocal space, corresponding to the inaccessible tilts. An identical situation will occur in XFEL diffraction by two-dimensional crystals, although the maximum tilt attainable will probably be larger because of the absence of multiple scattering for X-rays. If the maximum tilt is θ_{max}, then it is easily seen that the data completeness (*i.e.* the fraction of the full data set that can be measured) is $\sin(\theta_{max})$. We assume that Λ'_{2dc} is reduced by this factor. Let $\Lambda'^{(min)}_{2dc}$ be the minimum value of Λ'_{2dc} needed for successful reconstruction with complete data in practice [we might take, for example, $\Lambda'^{(min)}_{2dc} = 1.3$]. Successful *ab initio* phasing with a cone of missing data then requires that

$$\Lambda'_{2dc} > \frac{\Lambda'^{(min)}_{2dc}}{\sin(\theta_{max})}. \qquad (20)$$

For example, for $\Lambda'^{(min)}_{2dc} = 1.3$ and $\theta_{max} = 60°$, this requires that $\Lambda'_{2dc} > 1.5$.

Spence *et al.* (2003) considered the case of supplementing electron diffraction amplitude data with electron micrographs for a small range of tilts to provide some initial phase information. The images effectively provide phase information in the corresponding region of reciprocal space. If image data are available for tilts $0 < \theta < \theta_{images}$, they provide additional data (phases) that are a fraction $\sin(\theta_{images})$ of a full amplitude data set. Therefore, with this supplemental information the value of Λ'_{2dc}, denoted $\Lambda'^{(images)}_{2dc}$, can be approximated as

$$\Lambda'^{(images)}_{2dc} \simeq \Lambda'_{2dc}\,[\sin(\theta_{max}) + \sin(\theta_{images})]. \qquad (21)$$

Spence *et al.* (2003) describe simulations reconstructing lysozyme using data of this kind and the HIO algorithm. They found that reconstruction was successful using a flat support constraint (*i.e.* $\Lambda_{2dc} = 1$), a resolution of 3 Å and a unit-cell thickness of 40 Å (which gives $\Lambda'_{2dc} = 0.92$), complete diffraction amplitude data ($\theta_{max} = 90°$), and images for tilts up to 15° ($\theta_{images} = 15°$). Substituting these values into equation (21) gives $\Lambda'^{(images)}_{2dc} = 1.2$, a value that is consistent with our expectations for successful phasing, at least in the absence of noise.

8. Summary

The phase problem for a two-dimensional crystal is better determined than for a three-dimensional crystal, since the data give access to the continuous Fourier amplitude along lines in reciprocal space normal to the crystal plane. However, the allowed solutions still belong to a very high dimensional set and the solution is highly non-unique in general. The parameter Λ'_{2dc}, which depends on the shape of the molecular envelope and the resolution, is useful for defining uniqueness of the solution. This parameter is more useful than the usual constraint ratio Ω_{2dc} in this case. With sufficiently detailed molecular envelope information, and sufficient resolution, a unique solution and successful *ab initio* phasing may be feasible. Other information, such as non-crystallographic symmetry, histogram information, will further help constrain the solution. The results may also have application to the case of stacks of two-dimensional crystals in which there is lateral translational disorder between adjacent crystal sheets in the stack.

Cryo-electron crystallography of two-dimensional crystals has been an important technique in protein structure determination, particularly of membrane proteins. The recent

availability of XFEL sources offers the potential for X-ray crystallography of two-dimensional crystals, and avoids the necessity for cryo-freezing. The results therefore offer some optimism for *ab initio* phasing for XFEL data from two-dimensional crystals.

Recent work has highlighted the potential of iterative projection algorithms that have a large radius of convergence (Elser, 2003; Marchesini, 2007; Millane & Lo, 2013) for *ab initio* phasing in conventional three-dimensional protein crystallography with suitable constraints (Liu *et al.*, 2012; He & Su, 2015; Lo *et al.*, 2015). The results presented here therefore show that although *ab initio* phasing in two-dimensional crystallography may be difficult in general, it may be feasible in favourable circumstances. In other cases, successful phasing may be possible using much less initial phase information than is necessary in conventional crystallography with three-dimensional crystals.

Millane (2017) discusses the nature of the solution manifold in the case of the phase problem for one-dimensional crystals. In that case, the constraint ratio $\Omega_{1dc} \geq 2$ but, despite this large value, as a result of the specific form of the sampling of the Fourier amplitude in that case, the solution is not unique, but belongs to a fairly low dimensional set. Minimal additional information however is expected to be sufficient to restore uniqueness. In the two-dimensional crystal case considered here, however, we have only that $\Omega_{2dc} \geq 1$, so that a lack of data is problematic before the compounding effect of the regular sampling in the transverse plane in reciprocal space is considered. The net result is a very high dimensional solution manifold in general in the two-dimensional crystal case. Hence, significantly more *a priori* information is needed in this case for a unique solution. This requirement is characterized by the parameter Λ'_{2dc}.

Acknowledgements

The authors are grateful to John Spence and Henry Chapman for discussion and comments, and to Alok Mitra for discussion and for provision of the AQP1 electron density. This work was supported by a James Cook Research Fellowship and a Marsden grant to RPM, and a University of Canterbury College of Engineering Doctoral Scholarship to RDA.

References

Agard, D. A. & Stroud, R. M. (1982). *Biophys. J.* **37**, 589–602.
Arnal, R. D. *et al.* (2018). In preparation.
Barakat, R. & Newsam, G. (1984). *J. Math. Phys.* **25**, 3190–3193.
Bates, R. H. T. (1982). *Optik*, **61**, 247–262.
Bates, R. H. T. (1984). *Comput. Vis. Graph. Image Process.* **25**, 205–217.
Beinert, R. (2017). *Inf. Inference*, **6**, 213–224.
Beinert, R. & Plonka, G. (2015). *J. Fourier Anal. Appl.* **21**, 1169–1198.
Beinert, R. & Plonka, G. (2017). *Appl. Comput. Harmonic Anal.* In the press.
Bruck, Y. M. & Sodin, L. G. (1979). *Opt. Commun.* **30**, 304–308.
Elser, V. (2003). *J. Opt. Soc. Am. A*, **20**, 40–55.
Elser, V. & Millane, R. P. (2008). *Acta Cryst.* A**64**, 273–279.
Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.
Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press.
Frank, M. *et al.* (2014). *IUCrJ*, **1**, 95–100.
Frederix, P. L., Bosshart, P. D. & Engel, A. (2009). *Biophys. J.* **96**, 329–338.
Gipson, B. R., Masiel, D. J., Browning, N. D., Spence, J., Mitsuoka, K. & Stahlberg, H. (2011). *Phys. Rev. E*, **84**, 011916.
Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. (1996). *J. Mol. Biol.* **259**, 393–421.
Hayes, M. H., Jae Lim & Oppenheim, A. V. (1980). *IEEE Trans. Acoust. Speech, Signal. Process.* **28**, 672–680.
He, H. & Su, W.-P. (2015). *Acta Cryst.* A**71**, 92–98.
Kühlbrandt, W. & Wang, D. N. (1991). *Nature*, **350**, 130–134.
Liu, Z.-C., Xu, R. & Dong, Y.-H. (2012). *Acta Cryst.* A**68**, 256–265.
Lo, V. L., Kingston, R. L. & Millane, R. P. (2015). *Acta Cryst.* A**71**, 451–459.
Marchesini, S. (2007). *Rev. Sci. Instrum.* **78**, 011301.
Miao, J., Sayre, D. & Chapman, H. N. (1998). *J. Opt. Soc. Am. A*, **15**, 1662–1669.
Millane, R. P. (1990). *J. Opt. Soc. Am. A*, **7**, 394–411.
Millane, R. P. (1996). *J. Opt. Soc. Am. A*, **13**, 725–734.
Millane, R. P. (2017). *Acta Cryst.* A**73**, 140–150.
Millane, R. P. & Arnal, R. D. (2015). *Acta Cryst.* A**71**, 592–598.
Millane, R. P. & Chen, J. P. J. (2015). *J. Opt. Soc. Am. A*, **32**, 1317–1329.
Millane, R. P. & Lo, V. L. (2013). *Acta Cryst.* A**69**, 517–527.
Pedrini, B. *et al.* (2014). *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130500.
Ren, G., Cheng, A., Reddy, V., Melnyk, P. & Mitra, A. K. (2000). *J. Mol. Biol.* **301**, 369–387.
Spence, J. C. H., Weierstall, U., Fricke, T. T., Glaeser, R. M. & Downing, K. (2003). *J. Struct. Biol.* **144**, 209–218.
Stroud, R. M. & Agard, D. A. (1979). *Biophys. J.* **25**, 495–512.
Xu, L., Yan, P. & Chang, T. (1987). *Proc. IEEE Int. Symp. Circuits Syst.* pp. 851–854.

4 THE PHASE PROBLEM FOR TWO-DIMENSIONAL CRYSTALS. II. SIMULATIONS

Arnal, R. D., Zhao, Y. , Mitra, A. K., Spence, J. C. and Millane, R. P. (2018). "The phase problem for two-dimensional crystals. II. Simulations," *Acta Crystallographica A*, vol. 74, pp. 537-544.

FOUNDATIONS ADVANCES

ISSN 2053-2733

The phase problem for two-dimensional crystals. II. Simulations

Romain D. Arnal,[a] Yun Zhao,[b] Alok K. Mitra,[c] John C. H. Spence[b] and Rick P. Millane[a]*

[a]Computational Imaging Group, Department of Electrical and Computer Engineering, University of Canterbury, Christchurch, New Zealand, [b]Department of Physics, Arizona State University, Tempe, USA, and [c]School of Biological Sciences, University of Auckland, Auckland, New Zealand. *Correspondence e-mail: rick.millane@canterbury.ac.nz

Received 1 February 2018
Accepted 13 June 2018

Edited by H. Schenk, University of Amsterdam, The Netherlands

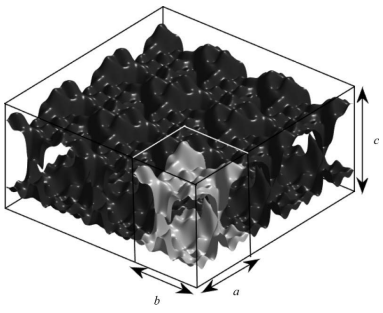
Keywords: phase problem; phase retrieval; two-dimensional crystals; XFELs; *ab initio* phasing; iterative projection algorithms; membrane proteins.

Phasing of diffraction data from two-dimensional crystals using only minimal molecular envelope information is investigated by simulation. Two-dimensional crystals are an attractive target for studying membrane proteins using X-ray free-electron lasers, particularly for dynamic studies at room temperature. Simulations using an iterative projection algorithm show that phasing is feasible with fairly minimal molecular envelope information, supporting recent uniqueness results for this problem [Arnal & Millane (2017). *Acta Cryst.* A**73**, 438–448]. The effects of noise and likely requirements for structure determination using X-ray free-electron laser sources are investigated.

1. Introduction

Successful phasing in macromolecular crystallography requires starting phase information such as from using isomorphous replacement or anomalous dispersion, or molecular replacement. The requirement for initial starting phases is a result of the unavoidable Bragg sampling of the diffraction by a three-dimensional crystal, which renders the solution to the *ab initio* phase problem highly nonunique in the absence of additional information (*e.g.* Millane, 1990; Millane & Arnal, 2015).

It has been known for some time however that for a two-dimensional crystal, the lack of Bragg sampling along one axis in reciprocal space provides additional constraints on the phases, thus easing the phase problem (Stroud & Agard, 1979; Agard & Stroud, 1982). Some macromolecular systems such as membrane proteins can form two-dimensional crystals grown in a lipid bilayer that mimics their native environment. These small crystals, typically 0.5–2 µm across, are not suitable for conventional crystallography with synchrotron sources, but have been studied using cryo-electron crystallography (Kühlbrandt *et al.*, 1994; Grigorieff *et al.*, 1996; Murata *et al.*, 2000; Ren *et al.*, 2001; Frank, 2006). Recently, the use of X-ray free-electron laser (XFEL) sources has been proposed as a new approach to structure determination using two-dimensional crystals (Frank *et al.*, 2014). The high intensity, small beam focus and short X-ray pulse length of XFELs can potentially overcome the difficulties of weak scattering, small grain size and radiation damage that are associated with two-dimensional crystallography using synchrotron sources (Frank *et al.*, 2014). An advantage of XFEL studies is the possibility of obtaining dynamic information at room temperature and under physiological conditions. Preliminary experiments have demonstrated that good data in one projection can be obtained to 4 Å resolution (Frank *et al.*, 2014; Pedrini *et al.*,



2014; Casadei *et al.*, 2018). It is likely that future extensions of this approach will allow collection of full three-dimensional data sets.

An alternative approach to two-dimensional crystallography using XFELs has been proposed by Kewish *et al.* (2010). In this approach, using a small X-ray focus, diffraction patterns are collected and classified according to the position of the X-ray beam relative to the crystal lattice. A reconstruction from these patterns can then be conducted using the method of ptychography (Rodenburg & Faulkner, 2004). The feasibility of such an approach was demonstrated by simulation.

The phase problem for two-dimensional crystals is alleviated to some degree, relative to that for three-dimensional crystals. Since the specimen is only one molecule or assembly thick, the Fourier amplitude can be measured effectively continuously in the direction in reciprocal space normal to the crystal surface, as opposed to only at the reciprocal-lattice points. Because the two-dimensional crystal is periodic in the plane parallel to the crystal surface, the Fourier amplitude is Bragg sampled in the corresponding two directions in reciprocal space. The Fourier amplitude is therefore measured along reciprocal-lattice lines, or Bragg rods, in reciprocal space. This increased sampling of the Fourier amplitude further constrains the phases compared with the three-dimensional crystal case. A similar situation occurs with one-dimensional crystals, in which case there is continuous sampling of the Fourier amplitudes on lattice planes in reciprocal space (Millane, 2017).

In the first article in this series (Arnal & Millane, 2017), which we refer to here as Paper I, we studied in detail the effect of this increased sampling on the expected uniqueness of the *ab initio* phase problem. We showed that although a unique solution is not expected in general, a unique solution may be feasible with fairly minimal *a priori* information. In particular, if low-resolution information is available on the molecular envelope, from atomic force microscopy (AFM) for example (Frederix *et al.*, 2009), there may be a unique solution. In this second article we demonstrate the implications of these uniqueness results by simulation of phase retrieval for data from two-dimensional crystals.

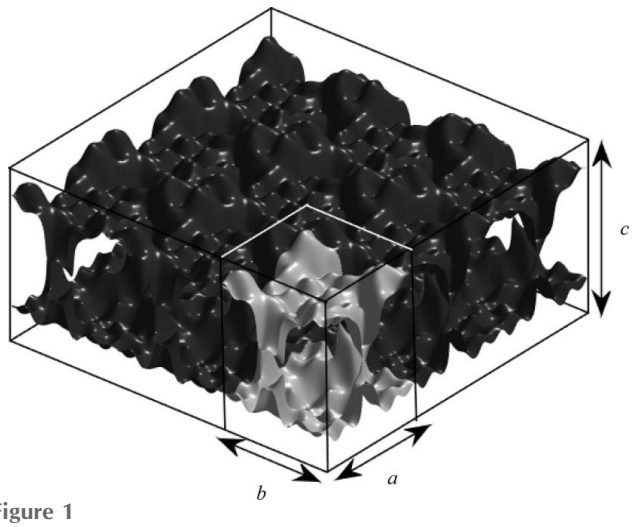


Figure 1
A two-dimensional crystal with a unit cell of dimensions $a \times b \times c$, where c is the maximum thickness of the molecular assembly, as described in the text.

In cryo-electron crystallography using two-dimensional crystals, phases are usually obtained from micrographs (images), and in some cases may be refined using a simple density-modification type of algorithm. A more sophisticated reconstruction algorithm has also been investigated (Gipson *et al.*, 2011). Phasing algorithms used for single-particle imaging (Bates, 1984; Fienup, 1982; Elser, 2003; Marchesini, 2007) have been investigated by Spence *et al.* (2003) for *ab initio* phasing for two-dimensional crystals. Using simulated reconstructions of lysozyme from two-dimensional crystal diffraction data, they found that *ab initio* phasing was not successful using the diffraction data alone, but it was successful if the diffraction data were supplemented by phases from sufficient images to fill in the reciprocal lattice with tilts between 0 and 15°. However, since images are not available to provide initial phase estimates for two-dimensional X-ray crystallography using XFEL sources, *ab initio* phasing takes on more importance in this case.

Uniqueness properties for the two-dimensional crystal phase problem as derived in Paper I are briefly reviewed in the next section. The phase retrieval algorithm that we used is briefly outlined in §3 and the results of simulations are described in §4. The effects of noise and practical aspects related to the use of XFEL data are discussed in §5. Concluding remarks are made in §6.

2. The two-dimensional crystal phase problem

Consider a two-dimensional crystal which, for simplicity, has a rectangular unit cell in plane group $P1$ with unit-cell dimensions a and b (Fig. 1). The results apply straightforwardly to other two-dimensional crystal classes, and Millane & Arnal (2015) show that space-group symmetry does not affect uniqueness of the phase problem, except for the case of centric space groups. Since the specimen is one molecular assembly thick, there is strictly no unit-cell dimension in the direction normal to the crystal plane, but we denote by c the maximum thickness of the monolayer in this direction (Fig. 1). For convenience, we refer to the cuboid of dimensions $a \times b \times c$ as the unit cell.

We denote the electron density in one unit cell of the two-dimensional crystal as $f(\mathbf{x}) = f(x, y, z)$, where $\mathbf{x} = (x, y, z)$ denotes position in real space, or in the discrete (sampled) case by $f[m, n, p]$ with $M \times M \times N$ samples (for which we assume a square unit cell, *i.e.* $a = b$). In the sampled case, the real electron density can be considered to belong to the vector space $\mathbb{R}^{M^2N}$, where each coordinate value represents the electron density at one sample point (Paper I).

We show in Paper I that the solution to the two-dimensional crystal phase problem is not unique if there is no additional *a priori* information, and it belongs to a high-dimensional manifold in $\mathbb{R}^{M^2N}$. A positivity constraint reduces the size of

Table 1
Values of Λ'_{2dc} for various experimental parameters.

$\lvert\mathcal{S}\rvert/\lvert\mathcal{U}\rvert$	v	c (Å)	d (Å)	Λ'_{2dc}	Λ'_{2dc} ($\theta_{max} = 70°$)	Λ'_{2dc} ($\theta_{max} = 60°$)
0.7	0.3	50	5	1.29	1.21	1.11
0.7	0.3	50	3	1.34	1.26	1.16
0.8	0.2	50	5	1.13	1.06	0.97
0.8	0.2	50	3	1.18	1.10	1.02

the solution set, but not significantly. A molecular envelope constraint, however, can significantly reduce the size of the solution set, and a unique solution is expected if the envelope deviates sufficiently from the unit cell.

Three parameters are derived in Paper I that can be used to characterize uniqueness of the phase problem. The most important parameter, denoted Λ'_{2dc}, is given by

$$\Lambda'_{2dc} = \left(1 - \frac{d}{c}\right)\frac{|\mathcal{U}|}{|\mathcal{S}|}, \tag{1}$$

where d is the resolution of the available diffraction data, $\mathcal{S}$ and $\mathcal{U}$ denote the region of the molecule (envelope) and the unit cell, respectively, and $|\cdot|$ denotes the size (volume) (Fig. 2). We show in Paper I that the solution to the phase problem for a two-dimensional crystal is highly constrained if

$$\Lambda'_{2dc} > 1 \tag{2}$$

and a unique solution is then likely. The problem is therefore favoured by a small envelope region, high resolution (small d) and a thick monolayer. In practice, in the presence of noise and missing data, a margin will be required, *i.e.*

$$\Lambda'_{2dc} > \Lambda'^{(min)}_{2dc} > 1, \tag{3}$$

where $\Lambda'^{(min)}_{2dc}$ is likely to be between, say, 1.2 and 1.5 (Millane & Lo, 2013). Some values of Λ'_{2dc} are shown in Table 1 for realistic experimental parameters, which show that phasing is feasible in favourable circumstances.

We also show in Paper I that the necessary condition for uniqueness $\Lambda'_{2dc} > 1$ can be rewritten as

$$v > \alpha, \tag{4}$$

where $v = 1 - |\mathcal{S}|/|\mathcal{U}|$ is the fractional volume of the unit cell (as defined above) that is outside the envelope and $\alpha = d/c$ is the resolution normalized to the thickness of the monolayer. Note that v will generally be smaller than the solvent content of the two-dimensional crystal since the latter includes solvent that is inside the envelope. Keep in mind, however, that equation (4) corresponds to equation (2) and an extra margin will be required in practice [equation (3)].

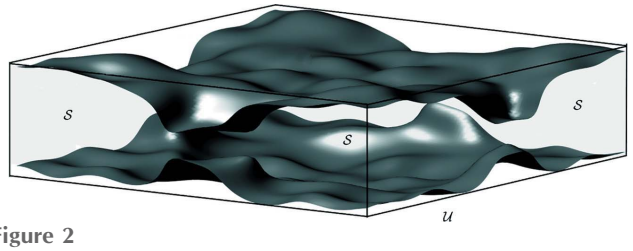


Figure 2
Illustration of the molecular envelope $\mathcal{S}$ and the unit cell $\mathcal{U}$ with upper and lower surfaces that bound $\mathcal{S}$.

In practice, there will be a missing cone of diffraction data around the z^* axis due to a maximum possible tilt of the two-dimensional crystal relative to the incident X-ray beam. We show in Paper I that the effect of the missing cone on Λ'_{2dc} is to multiply it by the factor $\sin(\theta_{max})$, where θ_{max} is the maximum tilt of the crystal to the beam. The effect of maximum tilts of 70° and 60° on Λ'_{2dc} is shown in Table 1. The effect is small for $\theta_{max} = 70°$, but is significant for $\theta_{max} = 60°$.

The second and third parameters described in Paper I are Λ_{2dc} given by

$$\Lambda_{2dc} = \frac{|\mathcal{U}|}{|\mathcal{S}|}, \tag{5}$$

which is equal to the inverse of the protein content of the crystal, and the constraint ratio Ω_{2dc} given by

$$\Omega_{2dc} = \frac{|\mathcal{P}_{2dc}|}{2|\mathcal{S}|}, \tag{6}$$

where $\mathcal{P}_{2dc}$ is the support of the Patterson function of the two-dimensional crystal. The parameters Λ_{2dc}, Λ'_{2dc} and Ω_{2dc} satisfy the inequalities

$$\begin{aligned} &\Lambda'_{2dc} < \Lambda_{2dc} \\ &1 \leq \Omega_{2dc} \leq \Lambda_{2dc}. \end{aligned} \tag{7}$$

The solution to the phase problem for an isolated object is unique if the constraint ratio Ω satisfies $\Omega > 1$ (Elser & Millane, 2008). However, we show in Paper I that, as a result of the particular sampling of the Fourier amplitude for a two-dimensional crystal, $\Omega_{2dc} > 1$ is not sufficient for uniqueness in the two-dimensional crystal case, and it is possible for $\Omega_{2dc} > 1$ but $\Lambda'_{2dc} < 1$.

In summary, the phase problem for a two-dimensional crystal is underconstrained in general, but is highly constrained if $\Lambda'_{2dc} > 1$. In contrast to the three-dimensional crystal case where greater than 50% solvent is required for uniqueness (Millane & Arnal, 2015), a smaller solvent content is sufficient in the two-dimensional crystal case. The presence of other constraints such as non-crystallographic symmetry or histogram information will further constrain the solution. Therefore, while the *ab initio* problem for a two-dimensional crystal does not have a unique solution in general, phasing may be feasible in favourable circumstances with fairly modest *a priori* information.

3. Phase retrieval

Most practical approaches to phase retrieval in the absence of initial phase information are based on iterative projection algorithms (Fienup, 1982; Elser, 2003; Marchesini, 2007; Millane & Lo, 2013; He & Su, 2015). Other approaches are also in use such as those of Lunin *et al.* (2000) and charge flipping (Oszlányi & Sütő, 2008), although these are in general effective only at low and high resolution, respectively. Iterative projection algorithms, on the other hand, are general-purpose global optimization procedures which are resolution

independent. We describe here the iterative projection algorithm that we used for phase retrieval for two-dimensional crystal data. We assume that the molecular support (envelope) is known. As described in the previous sections, the electron density in the unit cell is represented as a vector (or point), denoted $\mathbf{f}$, in the vector space $\mathbb{R}^{M^2N}$. The phase retrieval problem is formulated as finding a point in this vector space that is in the intersection of two constraint sets. One constraint set, denoted S, contains all electron densities (points) that satisfy the real-space constraints, which in our case are electron densities that are zero outside the molecular envelope. The other constraint set, denoted M, contains all electron densities whose Fourier amplitude is equal to the measured amplitude, denoted $|F(\mathbf{u})|$, on the reciprocal-lattice lines $\mathbf{u}$. A point in the intersection of the two sets thus satisfies both constraints and represents a solution to the problem. An intersection that contains a single point represents a unique solution, although a sufficiently small region of intersection represents a unique solution for practical purposes.

Iterative projection algorithms make use of projections onto the constraint sets. The projection of a point $\mathbf{f}$ in the vector space onto a constraint set A, denoted $P_A\mathbf{f}$, is the point in the set A that is closest to $\mathbf{f}$, *i.e.*

$$P_A\mathbf{f} = \operatorname*{argmin}_{\mathbf{f}'} ||\mathbf{f}' - \mathbf{f}||, \tag{8}$$

where $\operatorname{argmin}_{\mathbf{x}} f(\mathbf{x})$ denotes the value of $\mathbf{x}$ that minimizes $f(\mathbf{x})$ and $||\cdot||$ denotes the Euclidean norm. Iterative projection algorithms generate a sequence of 'iterates' $\mathbf{f}_i$ that ideally converge to a point in the intersection $S \cap M$ of the two constraint sets, thereby locating a solution that satisfies both the real-space constraints and the Fourier amplitude data. The sequence of iterates is generated by applying an update rule to the iterate $\mathbf{f}_i$ to generate the next iterate $\mathbf{f}_{i+1}$. For constraint sets that are nonconvex (as is the constraint set M in the case at hand), a variety of iterative projection algorithms have been used. Here we use the difference map algorithm (Elser, 2003) for which the update rule is given by

$$\mathbf{f}_{i+1} = \mathbf{f}_i + \beta(P_S L_M \mathbf{f}_i - P_M L_S \mathbf{f}_i), \tag{9}$$

where L_S and L_M are the relaxed projections given by

$$\begin{aligned} L_S\mathbf{f}_i &= (1+\gamma_S)P_S\mathbf{f}_i - \gamma_S\mathbf{f}_i \\ L_M\mathbf{f}_i &= (1+\gamma_M)P_M\mathbf{f}_i - \gamma_M\mathbf{f}_i, \end{aligned} \tag{10}$$

where γ_S and γ_M are relaxation parameters, and $-1 < \beta < 1$ is a parameter. Following Elser (2003), we used the values $\gamma_S = -1/\beta$ and $\gamma_M = 1/\beta$, and the algorithm has the single parameter β. Note that the iterate is not itself an estimate of the solution, but that once the algorithm has converged, or reached a fixed point, *i.e.* $\mathbf{f}_{i+1} = \mathbf{f}_i = \mathbf{f}'$, the solution $\hat{\mathbf{f}}$ (that satisfies both constraints) is given by (Elser, 2003; Millane & Lo, 2013)

$$\hat{\mathbf{f}} = P_S L_M \mathbf{f}' = P_M L_S \mathbf{f}'. \tag{11}$$

The support projection is implemented in the usual way by setting sample values that are outside the envelope to zero. Similarly, the positivity projection, if it is applied, is implemented by setting negative sample values to zero.

The Fourier space projection corresponds to making the smallest change to the current iterate such that its Fourier amplitude is equal to the measured value $|F(\mathbf{u})|$. It is easily shown that this corresponds to setting the Fourier amplitude of the iterate to the measured value and leaving the phase unchanged.

4. Simulations

Reconstruction of two-dimensional crystals was used to investigate the uniqueness results derived in Paper I, and their implications for phase retrieval, using simulated data. Two kinds of crystals were used. The first are simple synthetic objects designed to study the effect of different envelope shapes and the parameters described in §2. The second is the electron density of a membrane protein that forms two-dimensional crystals. Fourier amplitude data were calculated as for a two-dimensional crystal, *i.e.* Bragg sampled in two dimensions, and oversampled by a factor 4 in the third dimension. Noise was not added to the data since the objective here is to investigate uniqueness properties and determination of the solution under ideal conditions. The effects of noise are considered in the next section. An envelope for each object was defined as described below and an envelope constraint applied in real space. A positivity constraint was applied in some cases. The Fourier amplitude constraint was applied along the reciprocal-lattice lines (Bragg rods) where data are measured for the two-dimensional crystal. Phase retrieval was conducted using the difference map algorithm, as described in the previous section, with $\beta = 0.9$, which was started with random electron densities within the envelope. Since iterative projection algorithms do not always converge, for each example the algorithm was run a number of times starting with different random initial electron densities. Although failure of an iterative projection algorithm to converge in multiple runs does not prove that a solution cannot be found, these algorithms are quite effective if multiple starts are used and so the results obtained are quite suggestive of the feasibility of finding a solution.

Two error metrics were calculated to assess convergence of the algorithm and the quality of reconstructions. The first error metric, E_n, measures the difference between the amplitude data $|F(\mathbf{u})|$ and the Fourier amplitude of the iterate $|\hat{F}_n(\mathbf{u})|$, calculated based on $\hat{\mathbf{f}}$ given in equation (11), at iteration n, and is given by

$$E_n^2 = \frac{\sum_{\mathbf{u}}[|F(\mathbf{u})| - |\hat{F}_n(\mathbf{u})|]^2}{\sum_{\mathbf{u}} |F(\mathbf{u})|^2}. \tag{12}$$

The metric E_n monitors convergence to the diffraction amplitude data, *i.e.* it is small if the reconstructed electron density gives diffraction amplitudes that are close to data. The second error metric, e_n, measures the accuracy of the reconstruction, and is given by

$$e_n^2 = \frac{\sum_{\mathbf{x}}[f(\mathbf{x}) - \hat{f}_n(\mathbf{x})]^2}{\sum_{\mathbf{x}} f^2(\mathbf{x})}, \qquad (13)$$

where $f(\mathbf{x})$ is the true electron density and $\hat{f}_n(\mathbf{x})$ is the estimate of the solution at iteration n, calculated using equation (11). The metric e_n is small if the correct solution has been found. Multiple converged runs (small E) that give a small e suggest a unique solution. Runs that give a small E and a large e indicate nonunique solutions.

The first set of simulations used $8 \times 8 \times 8$ sample objects, with various defined envelopes, with the sample values selected randomly from a uniform distribution on (0, 1). Eight different envelopes were used, labelled 1 through 8. Envelope 1 is the full unit cell. Envelopes 2 and 3 have one flat surface and one structured surface, and envelope 4 has two structured surfaces. Envelopes 5 and 6 have 15% and 20%, respectively, of the samples removed at random positions, from the unit cell. Envelopes 7 and 8 have a pore (channel) through the object, with cross sections of one and two samples, respectively. The parameters Λ'_{2dc}, Λ_{2dc} and Ω_{2dc} are calculated for each envelope and are listed in Table 2. The algorithm was run ten times for each of these envelopes, both with and without positivity applied. Convergence of the algorithm is defined by $E_n < 10^{-3}$ and the solution is taken as that where the error metric E_n is a minimum. A correct solution is defined as one for which $e_n < 10^{-2}$.

The results are summarized in Table 2, which shows the number of runs that converged and the number of converged runs that gave the correct solution. Inspection of the table shows that the algorithm converged and the correct solution found for envelopes for which $\Lambda'_{2dc} \geq 1.06$. This is consistent with our expectations for data with no noise. Positivity did not have a dramatic effect, but increased the proportion of converged runs in some cases. The only cases of incorrect (nonunique) solutions were for envelope 1, where $\Lambda'_{2dc} = 0.87$. For intermediate values of Λ'_{2dc} $(0.99 < \Lambda'_{2dc} < 1.03)$ non-convergence, rather than convergence to an alternative solution, occurred. The reasons for this are not clear, but it is likely that more iterations are required to find a solution in these marginally constrained cases. Note that larger values of Ω_{2dc} (≥ 1.18) are required for convergence to a correct solution. Two examples of electron densities and their reconstructions are shown in Fig. 3. The reconstruction results obtained are consistent with the uniqueness theory described in Paper I.

Table 2
Summary of reconstruction results for the first set of examples described in the text.

Object	Λ'_{2dc}	Λ_{2dc}	Ω_{2dc}	Positivity	Runs converged	Correct solutions
1	0.87	1.00	0.94†	Y	8/10	0/8
				N	9/10	0/9
2	0.99	1.13	1.06	Y	0/10	
				N	0/10	
3	1.16	1.32	1.24	Y	10/10	10/10
				N	2/10	2/2
4	1.19	1.36	1.28	Y	8/10	8/8
				N	5/10	5/5
5	1.03	1.12	1.10	Y	0/10	
				N	0/10	
6	1.06	1.22	1.14	Y	10/10	10/10
				N	10/10	10/10
7	1.02	1.16	1.10	Y	0/10	
				N	0/10	
8	1.44	1.64	1.54	Y	10/10	10/10
				N	10/10	10/10

† Note that although $\Omega_{2dc} \geq 1$, it can be slightly less than unity for a discrete object.

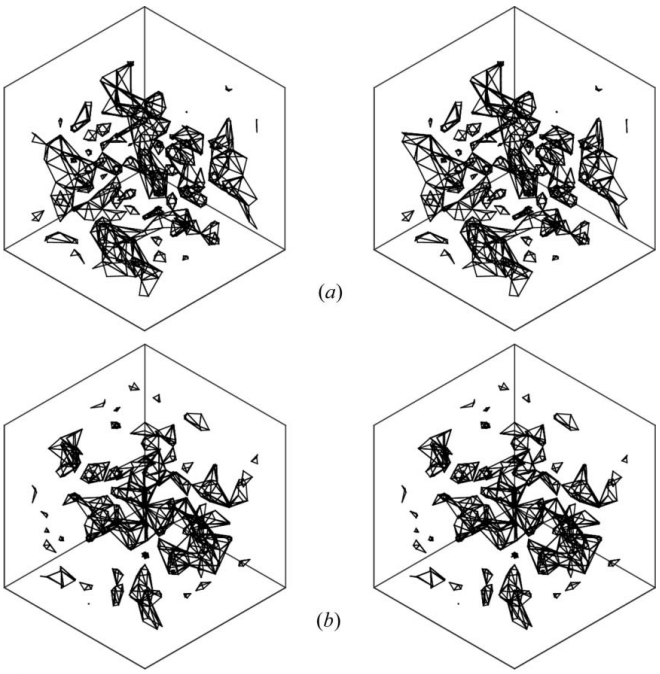


Figure 3
Example objects (left) and successful reconstructions (right) for (*a*) object 3 and (*b*) object 6, for the first set of examples described in the text. Contours are shown at 1σ and the object has been upsampled for display purposes.

For the second set of simulations we used the electron density of the membrane protein aquaporin 1 (AQP1) (Ren *et al.*, 2000). This crystal structure has a tetragonal unit cell of dimensions 100×100 Å, a thickness of 60 Å and space group $P42_12_1$. For our purposes, molecular envelopes of variable detail (resolution), as might be obtained by AFM, were constructed as described in Paper I. This involves convolving the electron density with a spherically symmetric three-dimensional Gaussian function with a full width at half-maximum (FWHM) set to a value, denoted d_e, which we refer to as the resolution of the envelope, followed by thresholding, to define the envelope at that resolution. As the envelope detail (resolution) increases (smaller d_e), the parameters Λ'_{2dc}, Λ_{2dc} and Ω_{2dc} all increase.

For the simulations described here, the AQP1 electron density was sampled on a $100 \times 100 \times 60$ grid, the envelope

determined as described above, and the parameters Λ_{2dc} and Ω_{2dc} calculated. The density was then subsampled onto a grid of $20 \times 20 \times 12$ samples for the purposes of phase retrieval and calculation of Λ'_{2dc}. Ten different envelope resolutions d_e were used, and for each, ten runs of the difference map were conducted, using the same protocol as for the first set of simulations described above with positivity applied. In this case it was found that the convergence criteria $E_n < 10^{-2}$ and $e_n < 0.20$ produced a sufficiently accurate electron density. Computational times were $\sim 10^4$ s for 10^5 iterations on an Intel Core i7-4700MQ CPU @2.4 GHz. The computational cost will be higher at higher resolution, but appears to be practical, particularly if some parallelization is employed.

The results are summarized in Table 3. Inspection of the table shows that good reconstructions are obtained for $\Lambda'_{2dc} \geq 1.24$, which corresponds to envelope resolutions greater than 17 Å. The proportion of algorithm runs that converge increases, and the average number of iterations required for convergence decreases, as Λ'_{2dc} increases, as expected. A correct solution was not obtained with no envelope information, as expected. These results are, again, consistent with the uniqueness theory described in Paper I, and illustrate the feasibility of phasing for two-dimensional crystals with only modest envelope information. Examples of the envelope and the true and reconstructed electron densities for $d_e = 16$ Å are shown in Fig. 4.

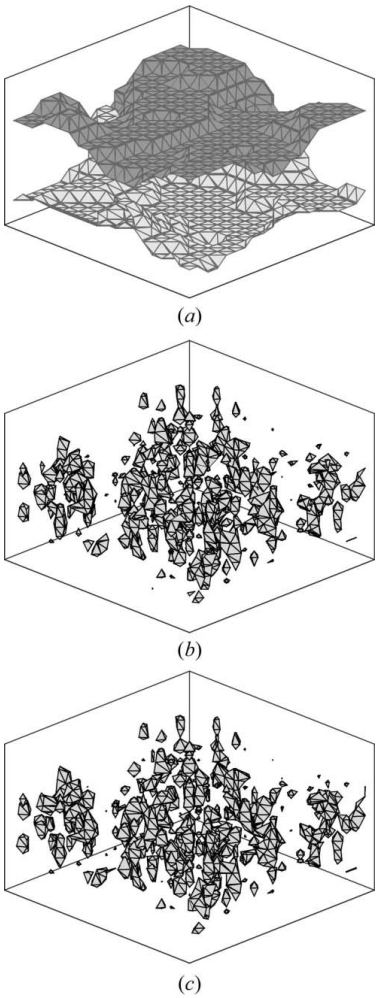


Figure 4
Reconstruction of APQ1 with noise-free amplitude data and an envelope resolution of 16 Å, as described in the text. (*a*) The envelope, and (*b*) and (*c*) the true and reconstructed electron densities, respectively, contoured at 1.5σ.

Table 3
Summary of reconstruction results for AQP1 with noise-free amplitude data.

d_e (Å)	Λ'_{2dc}	Λ_{2dc}	Ω_{2dc}	Runs converged	Correct solutions	Average number of iterations
∞	0.91	1.00	1.00	0/10		
18.0	1.19	1.30	1.25	0/10		
17.0	1.24	1.36	1.28	1/10	1/1	367000
16.0	1.25	1.37	1.29	4/10	4/4	367000
15.0	1.26	1.38	1.31	3/10	3/3	332000
14.0	1.27	1.39	1.32	5/10	5/5	430330
13.0	1.32	1.44	1.35	10/10	10/10	122080
12.0	1.34	1.46	1.37	10/10	10/10	13627
11.0	1.40	1.52	1.40	10/10	10/10	3485
10.0	1.42	1.55	1.43	10/10	10/10	450

5. Effects of noise

For successful phasing of XFEL data from two-dimensional crystals using only minimal envelope information, there are two practical considerations, both related to noise. The first is, what is the minimal signal-to-noise level in the measured diffraction amplitude data that is needed for successful phasing? The second is, how many single-shot diffraction patterns are needed to obtain this required signal-to-noise ratio? Both of these questions are considered in this section.

Even in cases where a unique solution to the phase problem is expected for noise-free data, in terms of $\Lambda'_{2dc} > \Lambda'^{(min)}_{2dc} > 1$, successful phase retrieval will inevitably depend on the precision of the diffraction amplitude data. Such is the case for any reconstruction problem, including conventional protein crystallography using three-dimensional crystals. In the latter case, the accuracy of the diffraction data is frequently measured by the resolution-dependent signal-to-noise ratio (SNR) $I/\sigma(I)$, where I and $\sigma(I)$ are the mean intensity and its standard deviation, respectively, in a resolution shell. The SNR decreases with increasing resolution, since I falls with increasing resolution whereas $\sigma(I)$ tends to remain relatively constant. In the case of three-dimensional crystal crystallography, an interpretable electron-density map can often be obtained, assuming that good molecular replacement phases are available, if $I/\sigma(I)$ is greater than about 1–2 at the highest resolution of the data (*e.g.* Gati *et al.*, 2017; Dods *et al.*, 2017). However, for phasing from two-dimensional crystal data in the absence of molecular replacement phase information, a larger SNR is likely to be needed.

To determine the minimum SNR needed for phasing of two-dimensional crystal data, reconstructions were performed for AQP1, as described in the previous section but with noise added to the amplitude data. For XFEL crystallography with

two-dimensional crystals, the individual, weak patterns will be dominated by photon noise, but on averaging many patterns, the noise in the merged patterns will be approximately Gaussian. Therefore, Gaussian distributed noise was added to the simulated intensity data, the variance of the noise being adjusted to fix the SNR at the highest resolution to desired values. Phase retrieval was conducted with various SNRs, with a $20 \times 20 \times 12$ sample grid and an envelope resolution of 10 Å. It was found that a good reconstructed electron-density map could be obtained for $I/\sigma(I) \geq 5$ at the highest resolution. An example reconstructed electron density for this case is shown in Fig. 5. This indicates that the SNR needed for phasing with two-dimensional crystals and only molecular envelope information is about three to five times greater than that needed in conventional crystallography starting with molecular replacement phases.

Each XFEL diffraction pattern from a two-dimensional crystal represents a spherical section through the reciprocal-lattice lines and thus consists of sharp spots. The spots allow the individual patterns to be oriented in reciprocal space, and then many patterns averaged to increase the SNR. Casadei *et al.* (2018) demonstrate this approach by orienting approximately 400 indexable XFEL diffraction patterns from bacteriorhodopsin two-dimensional crystals in a single section through reciprocal space. They subsequently average these patterns to obtain good estimates of the diffraction amplitudes on this section.

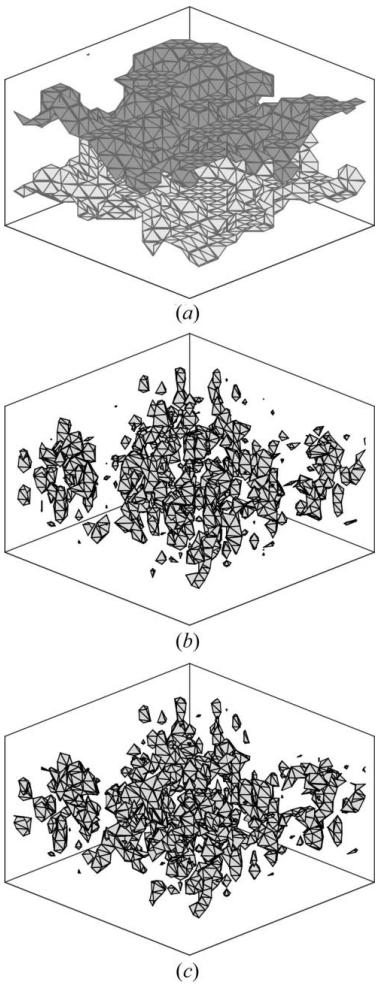


Figure 5
Reconstruction of AQP1 with noisy amplitude data with $I/\sigma(I) = 5$ and an envelope resolution of 10 Å, as described in the text. (*a*) The envelope, and (*b*) and (*c*) the true and reconstructed electron densities, respectively, contoured at 1.5σ.

We now consider the number of diffraction patterns likely to be needed for successful phasing with two-dimensional crystal data. For a three-dimensional crystal with P unit cells, the intensity of the Bragg reflections is proportional to P^2. For a two-dimensional crystal with P unit cells, the intensity on the lattice lines is also proportional to P^2. Therefore, data of a quality comparable with that obtained from three-dimensional crystals should be obtainable from two-dimensional crystals that have a similar number of unit cells. The number of unit cells intersected by the XFEL pulse is therefore of key importance. Structure determination has been successful, for example using serial femtosecond crystallography (SFX), with as few as about 10^4 unit cells in the XFEL focus and an SNR of 1–3 (*e.g.* Boutet *et al.*, 2012; Conrad *et al.*, 2015). Typical bacteriorhodopsin two-dimensional crystals, for example, of dimensions 0.5×0.5 µm contain $\sim 10^4$ unit cells, and should therefore give intensity data of comparable quality to such SFX experiments if the XFEL pulse intersects a full two-dimensional crystal grain. Boosting the SNR by a factor of about three, as described above, for phasing of two-dimensional crystal data, indicates that about a tenfold increase in the number of indexed patterns may be sufficient. Structure determination by SFX typically requires 10^4–10^5 indexed patterns, indicating that phasing of two-dimensional crystal data with only molecular envelope information may require of the order of 10^5–10^6 indexed patterns.

The number of patterns needed for phasing of two-dimensional crystal data can also be estimated using the results of Casadei *et al.* (2018). They obtained an SNR of about 5 at 4 Å resolution using 400 patterns, and their analysis indicates that to obtain the same SNR at 3 Å resolution would require about 4000 patterns. This is for a single section through reciprocal space however, and for a unit-cell thickness of 100 Å and fourfold oversampling along the Bragg rods, consideration of the relative volumes in reciprocal space shows that approximately 100 times as many patterns would be required to obtain a full three-dimensional data set at 3 Å resolution. This indicates a requirement of approximately 4×10^5 indexed patterns in this case, similar to the estimate obtained above. Obtaining this number of patterns would appear to be feasible with current instrumentation, and particularly so with likely improvements in sample scan rates.

6. Summary

XFEL sources offer the potential for X-ray crystallography of two-dimensional crystals at room temperature and for dynamic studies. Although the solution to the *ab initio* phase problem for a two-dimensional crystal is not unique in general, a unique solution and successful phasing are feasible with

rather modest molecular envelope information, less than that required for three-dimensional crystals. While for three-dimensional crystals a solvent content greater than 50% is required for a unique solution, a considerably smaller solvent content can give a unique solution in the two-dimensional crystal case. The utility of molecular envelope information in a specific case can be assessed using the parameter Λ'_{2dc}.

For membrane proteins, if moderately detailed molecular surface information is available, from AFM for example, then iterative projection algorithms appear to offer an effective tool for phasing in the absence of molecular replacement phase information. The results presented here show the potential for this approach, indicating that phasing is feasible with surface topography information at fairly modest resolution and realistic experimental parameters. This approach may be useful for XFEL diffraction imaging of membrane proteins using two-dimensional crystals where independent phase information is difficult to obtain.

The results suggest that true *ab initio* phasing, *i.e.* without molecular envelope information, may also be feasible with two-dimensional crystal data, as long as the volume of the envelope is known. Millane & Arnal (2015) show that, in principle, replacing a molecular envelope constraint by a molecular volume constraint does not alter uniqueness of the solution, although it does make finding the solution more difficult. Given that the solution to the two-dimensional crystal phase problem is better determined than for the three-dimensional crystal case, a molecular volume constraint, together with an algorithm such as shrink-wrap (Marchesini *et al.*, 2003) to refine the envelope, may make *ab initio* phasing feasible in the two-dimensional crystal case. Investigation in this direction would be fruitful.

Funding information

This work was supported by a James Cook Research Fellowship and a Marsden grant to RPM, a University of Canterbury College of Engineering Doctoral Scholarship to RDA and NSF BioXFEL STC award 1231306.

References

Agard, D. A. & Stroud, R. M. (1982). *Biophys. J.* **37**, 589–602.
Arnal, R. D. & Millane, R. P. (2017). *Acta Cryst.* A**73**, 438–448.
Bates, R. H. T. (1984). *Comput. Vis. Graph. Image Process.* **25**, 205–217.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Casadei, C. M. *et al.* (2018). *IUCrJ*, **5**, 103–117.
Conrad, C. E. *et al.* (2015). *IUCrJ*, **2**, 421–430.
Dods, R. *et al.* (2017). *Structure*, **25**, 1461–1468.
Elser, V. (2003). *J. Opt. Soc. Am. A*, **20**, 40–55.
Elser, V. & Millane, R. P. (2008). *Acta Cryst.* A**64**, 273–279.
Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.
Frank, J. (2006). *Three-dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press.
Frank, M. *et al.* (2014). *IUCrJ*, **1**, 95–100.
Frederix, P. L. T. M., Bosshart, P. D. & Engel, A. (2009). *Biophys. J.* **96**, 329–338.
Gati, C. *et al.* (2017). *Proc. Natl Acad. Sci. USA*, **114**, 2247–2252.
Gipson, B. R., Masiel, D. J., Browning, N. D., Spence, J., Mitsuoka, K. & Stahlberg, H. (2011). *Phys. Rev. E*, **84**, 011916.
Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. (1996). *J. Mol. Biol.* **259**, 393–421.
He, H. & Su, W.-P. (2015). *Acta Cryst.* A**71**, 92–98.
Kewish, C. M., Thibault, P., Bunk, O. & Pfeiffer, F. (2010). *New J. Phys.* **12**, 035005.
Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. (1994). *Nature*, **367**, 614–621.
Lunin, V. Y., Lunina, N. L., Petrova, T. E., Skovoroda, T. P., Urzhumtsev, A. G. & Podjarny, A. D. (2000). *Acta Cryst.* D**56**, 1223–1232.
Marchesini, S. (2007). *Rev. Sci. Instrum.* **78**, 011301.
Marchesini, S., He, H., Chapman, H. N., Hau-Riege, S. P., Noy, A., Howells, M. R., Weierstall, U. & Spence, J. C. H. (2003). *Phys. Rev. B*, **68**, 140101.
Millane, R. P. (1990). *J. Opt. Soc. Am. A*, **7**, 394–411.
Millane, R. P. (2017). *Acta Cryst.* A**73**, 140–150.
Millane, R. P. & Arnal, R. D. (2015). *Acta Cryst.* A**71**, 592–598.
Millane, R. P. & Lo, V. L. (2013). *Acta Cryst.* A**69**, 517–527.
Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J. B., Engel, A. & Fujiyoshi, Y. (2000). *Nature*, **407**, 599–605.
Oszlányi, G. & Sütő, A. (2008). *Acta Cryst.* A**64**, 123–134.
Pedrini, B. *et al.* (2014). *Philos. Trans. R. Soc. London Ser. B*, **369**, 20130500.
Ren, G., Cheng, A., Reddy, V. S., Melnyk, P. & Mitra, A. K. (2000). *J. Mol. Biol.* **301**, 369–387.
Ren, G., Reddy, V. S., Cheng, A., Melnyk, P. & Mitra, A. K. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 1398–1403.
Rodenburg, J. M. & Faulkner, H. M. (2004). *Appl. Phys. Lett.* **85**, 4795–4797.
Spence, J. C. H., Weierstall, U., Fricke, T. T., Glaeser, R. M. & Downing, K. (2003). *J. Struct. Biol.* **144**, 209–218.
Stroud, R. M. & Agard, D. A. (1979). *Biophys. J.* **25**, 495–512.

5 *AB INITIO* MOLECULAR REPLACEMENT PHASING

5.1 INTRODUCTION

In this chapter, a novel *ab initio* phase retrieval technique is presented that has strong similarities to conventional molecular replacement (MR) phasing and, for this reason, is referred to here as *ab initio* molecular replacement (*ai*MR) phasing. Recalling from Section 1.3.2.4, in conventional MR, an homologous protein of known structure is positioned in the target unit cell and used to calculate approximate phases. If these phases are sufficiently accurate, they can be used together with the measured structure amplitudes to calculate an interpretable electron density map. Unfortunately, the use of an *a priori* model renders MR ill-equipped for finding new protein folds and is prone to model bias [Evans and McCoy, 2008].

The proposed *ai*MR technique overcomes these issues by collecting experimental data from multiple crystal forms of the target protein. Contrary to MR, no model structure is used. In effect, in *ai*MR, the model and target are the same. It will be shown in Section 5.3 that the diffraction data collected from different crystal forms are mostly independent under fairly general conditions. As the different crystal forms are built with the same building block (the same protein), the independent diffraction data constitute a source of additional information to solve the phase problem.

In practice, crystal forms are generally encountered with two main modifications of the unit cell – swelling/shrinking of the unit cell (which mainly changes the sampling positions $\mathbf{h}$) and space group transformations (which fundamentally changes the way the molecular transform is sampled). The former can be caused by physical or chemical changes during or after crystallisation, for instance changes in temperature, humidity or pressure. The latter is generally accompanied by significant changes in unit cell parameters and occurs generally because of different chemical composition of the mother solution or crystal. Both cases are covered by the *ai*MR theory presented in Section 5.4, but the outcomes may be different in practice.

Fortunately, for many proteins, different crystal forms can easily be formed using controlled crystal hydration/dehydration. Furthermore crystallisation techniques, such as high throughput crystallisation screening (HTCS) are a good source of different

crystal forms due to the automatic and systematic coverage of many crystallisation parameters.

The *ai*MR technique depends on the similar tertiary structures of identical primary structures in different crystal environments. Clearly, this is more likely to be the case than in conventional MR where the primary structures are different. In practice, however, there will be at least small differences between the structures, and this will affect the resolution of the reconstruction. If the proteins are similar enough for an IPA to converge, the reconstruction would most likely correspond to an average of the protein electron densities of the different structures. A representative study of the structural homology that exists between two crystal forms is given in Section 5.5.2.

5.2 DIFFRACTION BY MULTIPLE CRYSTAL FORMS

Denoting by $g(\mathbf{x})$ the electron density of the protein, the electron density of the unit cell in crystal form n with $n = (1, 2, \ldots, N)$, denoted by $f_n(\mathbf{x})$, is given by

$$f_n(\mathbf{x}) = \sum_{k=1}^{K_n} g(R_{nk}\mathbf{x} + \mathbf{t}_{nk}) = \sum_{k=1}^{K_n} g_{nk}(\mathbf{x}), \tag{5.1}$$

where the sum is over the K_n protein copies (asymmetric units) in crystal form n. The rotation matrices R_{nk} and translation vectors $\mathbf{t}_{nk}$ describe the space group symmetry operators for crystal form n, with $g_{nk}(\mathbf{x})$ the copies positioned in the unit cell. The operators R_{nk} and $\mathbf{t}_{nk}$ are assumed known in the *ai*MR technique described here. The Fourier intensity measured at position $\mathbf{u}$ in reciprocal space for crystal form n, denoted $I_n(\mathbf{u})$, is given by

$$I_n(\mathbf{u}) = |\mathcal{F}\left[f_n(\mathbf{x})\right]|^2 = \left| \sum_{k=1}^{K_n} G(R_{nk}^T\mathbf{u}) \exp\left(i2\pi\mathbf{u}\cdot\mathbf{t}_{nk}\right) \right|^2, \tag{5.2}$$

where $G(\mathbf{u})$ is the Fourier transform of $g(\mathbf{x})$.

Equation (5.2) relates the diffraction intensities I_n of the crystal forms to the electron density of the common building block $g(\mathbf{x})$. Only the Bragg diffraction data, $I_n(\mathbf{h})$, where the $\mathbf{h}$ depend on the cell constants, are measured during experiments. The data from different crystal forms are independent if either, the sampling positions $\mathbf{h}$ between crystal forms are different or, the operators R_{nk} and $\mathbf{t}_{nk}$ are different for all crystal forms n. In either case, the equivalent sampling of reciprocal space obtained from multiple crystal forms is denser than the sampling from a single crystal form, providing additional information.

5.3 UNIQUENESS

For *ab initio* phasing, an immediate question of fundamental importance is, does the data from additional crystal forms provide sufficient information to provide a unique solution to the phase problem? In this section, uniqueness of the *ai*MR phase problem is examined using the constraint ratio.

Consider first the case where the data from the N crystal forms are all independent. The total number of data is then the sum of the number of data from each crystal form. Referring to Section 2.2, the number of data for crystal form n is $|\mathcal{P}_n|/2$, where $\mathcal{P}_n$ is the region of the Patterson function of crystal form n. The constraint ratio for N crystal forms is then given by

$$\Omega_{aiMR} = \frac{\sum_{n=1}^{N} |\mathcal{P}_n|/2}{|U|} = \sum_{n=1}^{N} \frac{1}{2p_n}, \tag{5.3}$$

where p_n is the protein content of crystal form n, and the simplification to equation (2.5) has been made. Equation (5.3) shows that the constraint ratio increases dramatically with the number of crystal forms if all the data are independent.

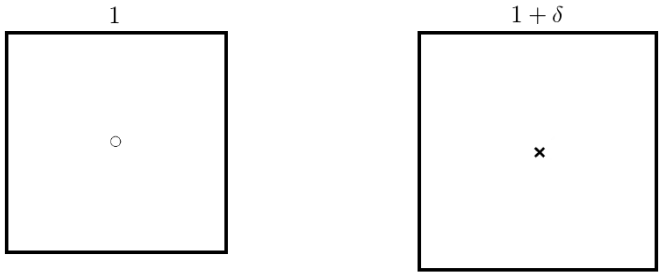
It is possible that if the unit cells are very similar, then the sampling of reciprocal space $\mathbf{h}$ will be similar and not all the structure amplitudes will be independent. This is illustrated by considering two $p1$ square crystal forms whose cell constants differ by a small relative proportion δ, as shown in Fig. 5.1. Denoting, loosely, by ε, the overall proportion of the total data that are independent, then equation (5.3) is replaced by

$$\Omega_{aiMR} = \varepsilon \sum_{n=1}^{N} \frac{1}{2p_n}. \tag{5.4}$$

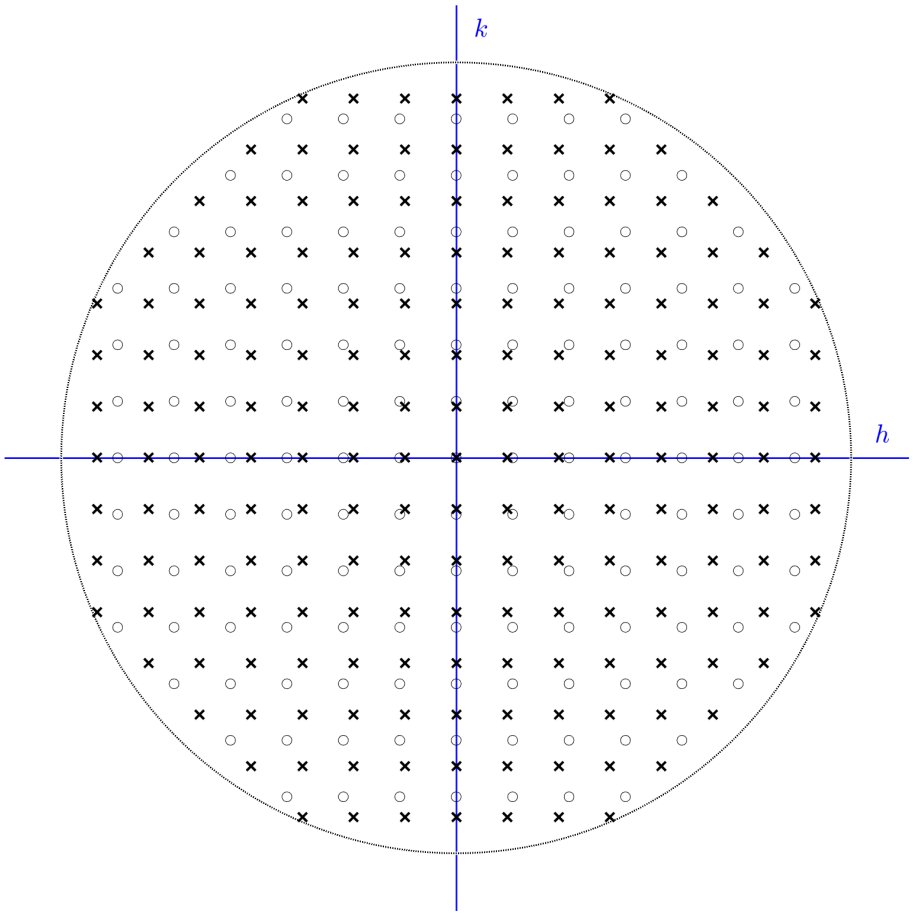
Likely values of ε can be assessed as follows. Assuming, again loosely, that the amplitudes at two sample locations are independent if they are spaced by greater than a fraction Δ of the reciprocal lattice spacing, the quantity ε is then a function of resolution d and is given by

$$\varepsilon(\Delta) = \frac{N(\Delta, d)}{N(d)}, \tag{5.5}$$

where $N(\Delta, d)$ is the number of independent reflections up to resolution d, and $N(d)$ is the total number of reflections. Simple calculations to evaluate likely values of ε are made using the square unit cell shown in Fig. 5.1(a). With a unit cell of dimensions unity, d can be treated as a normalised resolution with the true resolution given by ad, where a is the actual cell constant. The quantity $N(d)$ is known as the Gauss circle



(a)



(b)

Figure 5.1 (a) Two $p1$ square crystal forms whose unit cell differ by a small fraction δ and (b) the corresponding sampling in reciprocal space with a fraction $\delta = 0.05$.

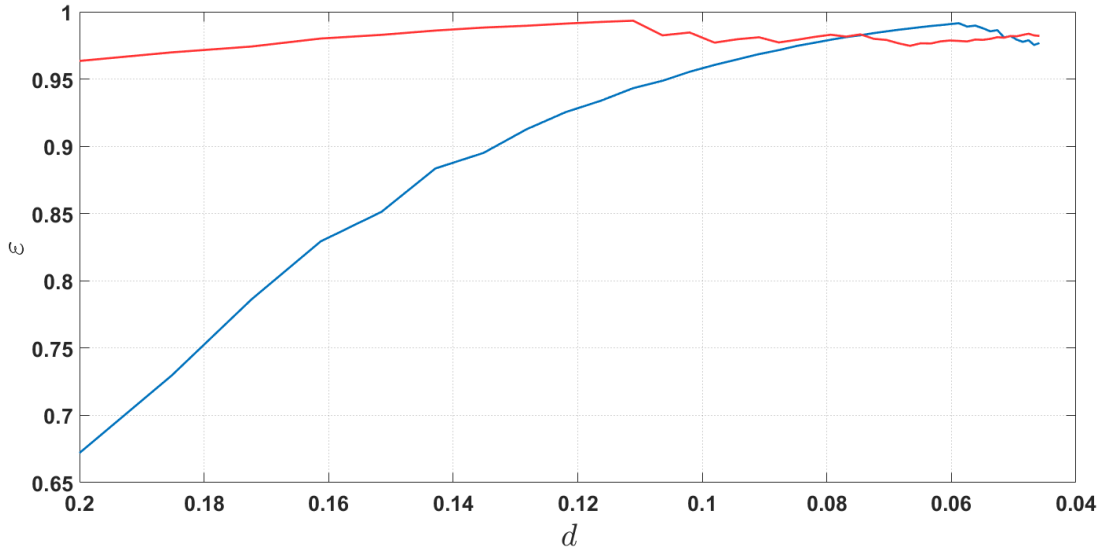


Figure 5.2 Value of ε as a function of resolution d for $\delta = 0.05$ (blue curve) and $\delta = 0.10$ (red curve).

problem [Hilbert and Cohn-Vossen, 1999] and given by

$$N(d) = 1 + 4\sum_{i=0}^{\infty}\left(\left\lfloor\frac{1}{d^2(4i+1)}\right\rfloor - \left\lfloor\frac{1}{d^2(4i+3)}\right\rfloor\right). \tag{5.6}$$

The quantity $N(\Delta, d)$ is now a function of δ, denoted $N(\Delta, d, \delta)$, and $\varepsilon(\Delta, \delta, d)$ is calculated by simulation for $\Delta = 0.1$ and shown as a function of δ and d in Fig. 5.2. It can be seen that ε approaches 1 in most practical cases, even for small changes in unit cell dimensions, as long as the normalised resolution of the data is better than about 0.1. Even for a relatively small unit cell of dimensions 50Å, this corresponds to a resolution greater than 5Å. In this case, $> 95\%$ of the data are independent if the unit cell variations are no less than 5%.

5.4 IMPLEMENTATIONS OF *AI*MR

The phase retrieval problem for *ai*MR consists of synthesising the N diffraction data sets with the real space constraints to achieve a solution. IPAs are used for this purpose.

The following subsections present two distinct IPA implementations of *ai*MR corresponding to two different perspectives. For each implementation, the real space and reciprocal space projections are derived. In the first implementation (A), the IPA attempts to reconstruct the full unit cell of each crystal form, $f_n(\mathbf{x})$, whereas, the second

implementation (B), reconstructs the proteins in each crystal form, $g_{nk}(\mathbf{x})$, individually.

5.4.1 Implementation A

In implementation A, the algorithm operates on the set of the complete unit cell densities $f_n(\mathbf{x})$.

5.4.1.1 Reciprocal space projection

The reciprocal space projection, denoted P_M, makes the smallest change in the current estimate of $f_n(\mathbf{x})$, such that the Fourier intensities are equal to the experimental intensities I_n. This projection was defined in Chapter 1 (equation (1.26)) and is the usual Fourier space projection for reconstruction of a single object and is repeated here for convenience:

$$P_M f_n(\mathbf{x}) = \mathcal{F}^{-1}\left\{\frac{\sqrt{I_n(\mathbf{h})}}{|F_n(\mathbf{h})|}F_n(\mathbf{h})\right\}, \tag{5.7}$$

where $F_n(\mathbf{h})$ is the Fourier transform of $f_n(\mathbf{x})$.

5.4.1.2 Real space projection

The real space projection, denoted P_S, enforces the following two constraints: (i) the iterate $f_n(\mathbf{x})$ is such that the component parts $g_{nk}(\mathbf{x})$ from which it is built are identical for each unit cell n, i.e. the $g_{nk}(R_{nk}^{-1}(\mathbf{x}-\mathbf{t}_{nk}))$ are identical for all n and k, and (ii) $f_n(\mathbf{x})$ is restricted to the support region $s(\mathbf{x})$. Both of these requirements can be satisfied, in the least distance sense, by averaging the components, applying a support constraint $s(\mathbf{x})$, and then rebuilding the unit cells. The real space projection can then be written in two steps as

$$g'(\mathbf{x}) = \frac{s(\mathbf{x})}{P}\sum_{n=1}^{N}\sum_{k=1}^{K_n} S_{nk} f_n(\mathbf{x}), \tag{5.8}$$

$$P_S f_n(\mathbf{x}) = \sum_{k=1}^{K_n} g'(R_{nk}\mathbf{x} + \mathbf{t}_{nk}), \tag{5.9}$$

where $P = \sum_n K_n$ is the total number of asymmetric units in all the crystals, and the operator S_{nk} extracts an estimate of $g(\mathbf{x})$ from the k^{th} asymmetric unit in $f_n(\mathbf{x})$.

If the support regions in a unit cell do not overlap, then the operation S_{nk} can be achieved by simply repositioning $f_n(\mathbf{x})$ within the corresponding support region, and S_{nk} is given by

$$S_{nk} f_n(\mathbf{x}) = f_n(R_{nk}^{-1}(\mathbf{x} - \mathbf{t}_{nk})). \tag{5.10}$$

If the support regions overlap, then one can in principle still develop a distance minimising projection, but the implementation depends on the nature of all the overlaps in the unit cell, and its effectiveness depends on the degree of overlap. Such an implementation would be computationally intensive in practice.

5.4.1.3 Two-dimensional simulations of approach A

Simulations were conducted to illustrate implementation A of *ai*MR described in the previous sections. For each simulation, a pair of different crystal forms of a 32×32 Lena electron density, the "protein", were generated. The support, $s(\mathbf{x})$ was defined by the 32×32 logical mask and the Fourier intensity data $I_n(\mathbf{h})$ calculated using the DFT. The *ai*MR real and reciprocal space projections were implemented and incorporated into the difference map algorithm.

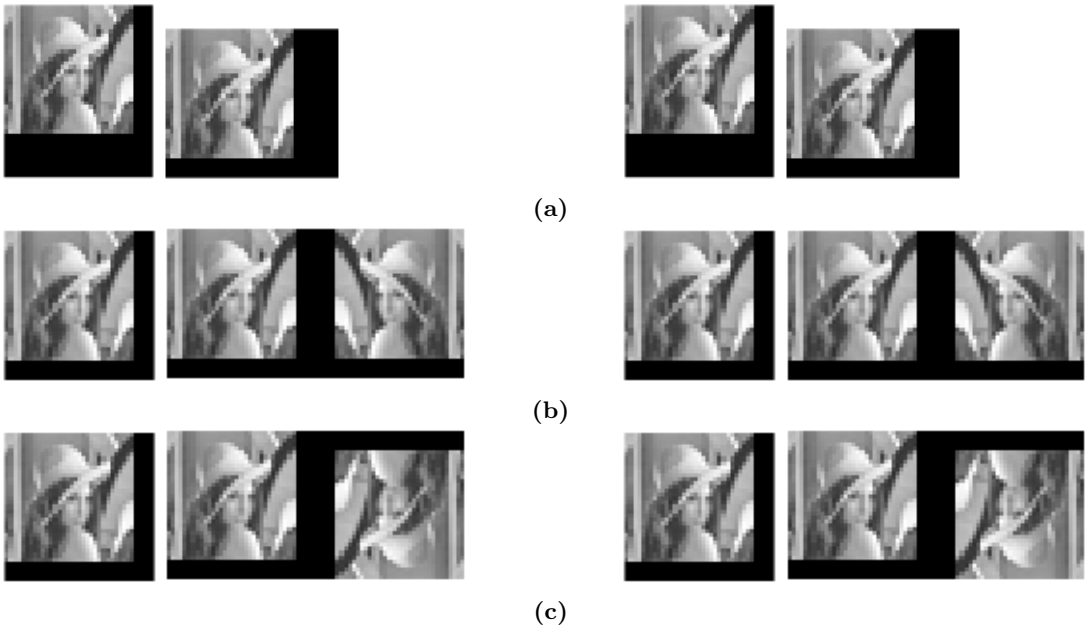


Figure 5.3 Reconstruction results for *ai*MR phase retrieval using approach A. Original pairs of crystal forms (left) and corresponding reconstructions (right). (a) Two $p1$ crystal forms with different shape of the same volume, (b) a $p1$ crystal and a pm crystal, and (c) a $p1$ crystal and a $p2$ crystal.

The three unit cell pairs used in simulation are shown in Fig. 5.3. The first pair consists of two $p1$ unit cells with different shapes of the same volume (Fig. 5.3(a)). The unit cells have dimensions 37×43 and 43×37 pixels. The corresponding solvent content is 36% in both crystals. The second pair is composed of a $p1$ unit cell of dimensions 37×37 and a pm unit cell of 74×37 pixels (Fig. 5.3(b)). Both have 25% solvent content. The third cell is similar to the previous case but with a $p2$ unit cell rather than pm (Fig. 5.3(c)). Computation of $\Omega_c = 1/2p$ for these crystals indicates no unique solution can be expected if taken alone. However, using $\varepsilon = 0.8$ and equation (5.4) the

*ai*MR constraint ratio is calculated as (a) $\Omega_{aiMR} = 1.25$, (b) $\Omega_{aiMR} = 1.06$ and (c) $\Omega_{aiMR} = 1.6$ indicating a unique solution may be obtained using *ai*MR. Note that, for the third case, the second crystal is centric so that $\Omega_c = 1/p$.

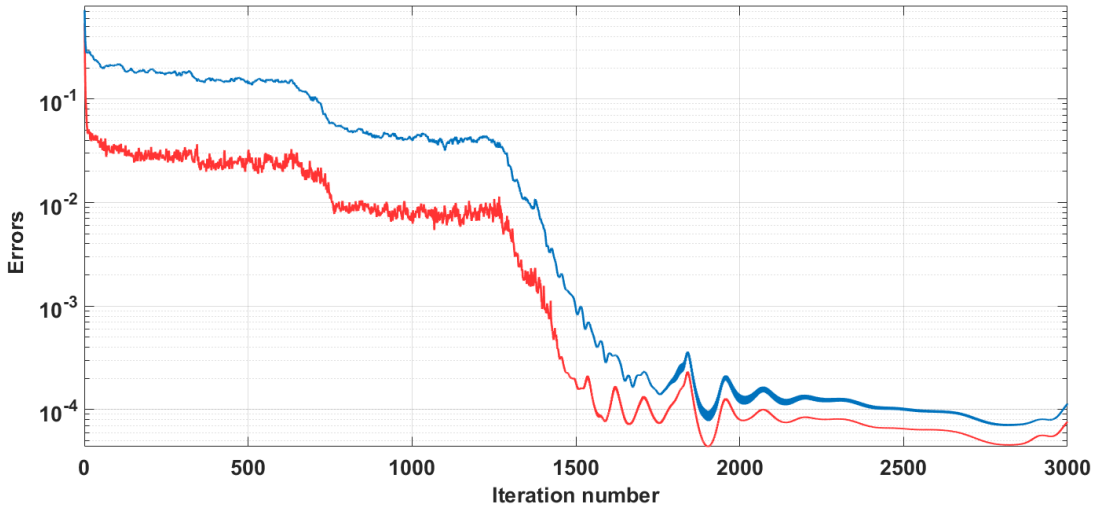


Figure 5.4 Real space (blue) and reciprocal space (red) errors e and E as a function of iteration for the unit cell and reconstruction shown in Fig. 5.3(a).

For each simulation, the reciprocal and real space errors were computed with the usual error metrics given in Section 1.4.5. Representative error plots are shown in Fig. 5.4, Fig. 5.5 and Fig. 5.6 for case (a), (b) and (c) respectively. Successful reconstructions were obtained in all cases, in accordance with the values of Ω_{aiMR} above. The ease of reconstruction is correlated to the values of Ω_{aiMR}. Case (b) is the hardest with $\Omega_{aiMR} = 1.06$ as can be seen in Fig. 5.5 with a lengthy search for an attractor, followed by a slow convergence to the solution. Case (c) is the easiest with $\Omega_{aiMR} = 1.6$, converging in less than a thousand iterations (Fig. 5.6), while case (a) with $\Omega_{aiMR} = 1.25$ converges in about 1500 iterations (Fig. 5.4). The average number of iterations required for convergence decrease, as Ω_{aiMR} increases, as expected. The smaller final error for case (c) may be due to the problem's centric nature which constrains the phase to just the two values, 0 and π. These results are consistent with the uniqueness theory described in Section 5.3.

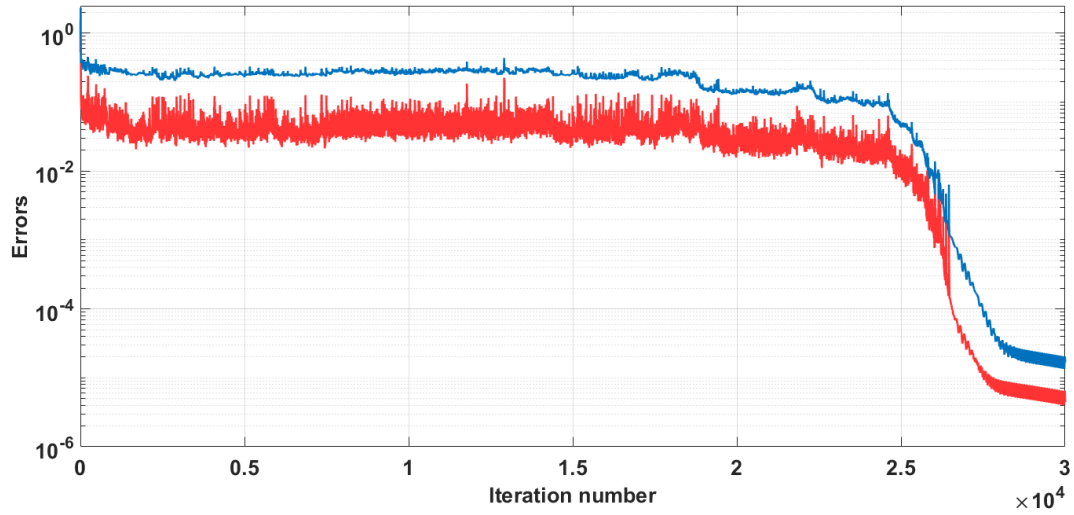


Figure 5.5 Real space (blue) and reciprocal space (red) errors e and E as a function of iteration for the unit cell and reconstruction shown in Fig. 5.3(b).

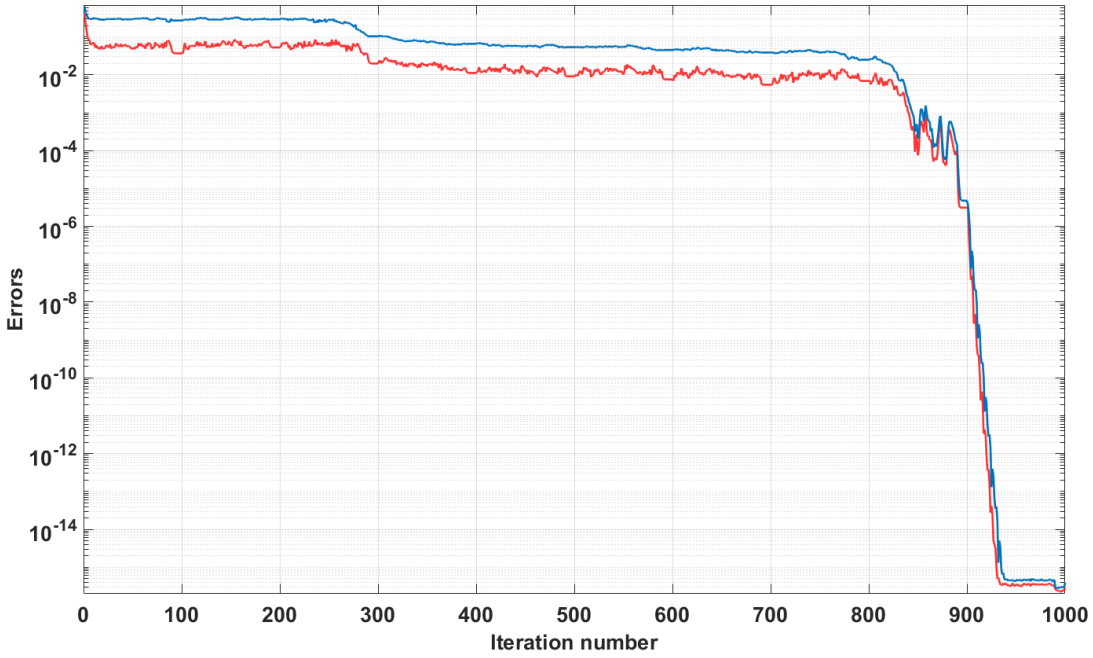


Figure 5.6 Real space (blue) and reciprocal space (red) errors e and E as a function of iteration for the unit cell and reconstruction shown in Fig. 5.3(c).

5.4.2 Implementation B

In implementation B, the algorithm operates on the set of densities of the individual molecules in their respective unit cell $g_{nk}(\mathbf{x})$.

5.4.2.1 Reciprocal space projection

Writing the complex number $G(R_{nk}^T\mathbf{h})\exp(i2\pi\mathbf{h}\cdot\mathbf{t}_{nk})$ in equation (5.2) as $a_{nk}+ib_{nk}$, equation (5.2) can be written as

$$I_n(\mathbf{h}) = \left|\sum_{k=1}^{K_n}(a_{nk}+ib_{nk})\right|^2 = \left(\sum_{k=1}^{K_n}a_{nk}\right)^2 + \left(\sum_{k=1}^{K_n}b_{nk}\right)^2, \tag{5.11}$$

where, for convenience, the $\mathbf{h}$ dependence has been dropped.

The diffracted intensity for a single crystal form can thus be equivalently expressed with the real numbers a_{nk} and b_{nk}. Different values of a_{nk} and b_{nk} lead to the same intensity, and equation (5.11) describes a $(2K_n-1)$–dimensional surface, denoted χ, in the $2K_n$–dimensional space Υ. Following [Chen et al., 2016], the diffracted intensity is normalised by K_n, giving

$$I'_n(\mathbf{h}) = I_n(\mathbf{h})/K_n. \tag{5.12}$$

The reciprocal space projection moves a point $\Phi_n = [a_{n1},...,a_{nK_n},b_{n1},...,b_{nK_n}]$ in Υ to the closest point $\Phi_n^\chi = [a_{n1}^\chi,...,a_{nK_n}^\chi,b_{n1}^\chi,...,b_{nK_n}^\chi]$ that lies on the surface χ. To find Φ_n^χ, the problem is formulated into an optimisation problem by defining two functions - a distance function and a constraint function:

- The constraint function $\zeta(\Phi_n)$ ensures that Φ_n lies on χ and is defined as

$$\zeta(\Phi_n) = \frac{1}{K_n}\left[\left(\sum_{k=1}^{K_n}a_{nk}\right)^2 + \left(\sum_{k=1}^{K_n}b_{nk}\right)^2\right] - I'_n(\mathbf{u}). \tag{5.13}$$

- The distance to be minimised is the distance between an arbitrary starting point, $\Phi_n^0 = [a_{n1}^0,...,a_{nK_n}^0,b_{n1}^0,...,b_{nK_n}^0]$ and Φ_n, is given by

$$\psi(\Phi_n) = \sum_{k=1}^{K_n}\left((a_{nk}-a_{nk}^0)^2 + (b_{nk}-b_{nk}^0)^2\right). \tag{5.14}$$

The method of Lagrange multipliers is used to solve this optimisation problem by defining the Lagrange function

$$\mathcal{L}(\Phi_n;\lambda) = \psi(\Phi_n) + \lambda\zeta(\Phi_n), \tag{5.15}$$

where λ is the Lagrange multiplier. The gradients of $\mathcal{L}(\Phi_n, \lambda)$ are calculated as

$$\nabla\mathcal{L}(\Phi_n, \lambda) = \left(\frac{\partial\mathcal{L}}{\partial a_{nk}}, \frac{\partial\mathcal{L}}{\partial b_{nk}}, \frac{\partial\mathcal{L}}{\partial\lambda}\right), \tag{5.16}$$

which at a local minima gives

$$\frac{\partial\mathcal{L}}{\partial a_{nk}} = 0 = 2(a^{\chi}_{nk} - a^{0}_{nk}) + \frac{2}{K_n}\lambda\sum_{m=1}^{K_n} a^{\chi}_{nm}, \tag{5.17}$$

$$\frac{\partial\mathcal{L}}{\partial b_{nk}} = 0 = 2(b^{\chi}_{nk} - b^{0}_{nk}) + \frac{2}{K_n}\lambda\sum_{m=1}^{K_n} b^{\chi}_{nm}, \tag{5.18}$$

and

$$\frac{\partial\mathcal{L}}{\partial\lambda} = 0 = \psi(\Phi^{\chi}_n). \tag{5.19}$$

Summing equations (5.17) and $i\times$(5.18) over all k gives

$$2\sum_{k=1}^{K_n}(a^{\chi}_{nk} - a^{0}_{nk}) + \frac{2}{K_n}\lambda K_n\sum_{m=1}^{K_n} a^{\chi}_{nm} + 2i\sum_{k=1}^{K_n}(b^{\chi}_{nk} - b^{0}_{nk}) + \frac{2i}{K_n}\lambda K_n\sum_{m=1}^{K_n} b^{\chi}_{nm} = 0, \tag{5.20}$$

and denoting $z^{0}_{nk} = a^{0}_{nk} + ib^{0}_{nk}$ and $z^{\chi}_{nk} = a^{\chi}_{nk} + ib^{\chi}_{nk}$, we obtain

$$\sum_{k=1}^{K_n} z^{\chi}_{nk} = \frac{1}{1+\lambda}\sum_{k=1}^{K_n} z^{0}_{nk}. \tag{5.21}$$

This expression can be substituted back into equations (5.17) and (5.18) to give

$$z^{\chi}_{nk} = z^{0}_{nk} - \frac{1}{K_n}\left(\frac{\lambda}{1+\lambda}\right)\sum_{m=1}^{K_n} z^{0}_{nm}, \tag{5.22}$$

and into equation (5.19) to give

$$I'_n = \frac{1}{K_n}\frac{|\sum_{k=1}^{K_n} z^{0}_{nk}|^2}{(1+\lambda)^2}. \tag{5.23}$$

Solving for λ in equation (5.23) gives

$$\lambda = \sqrt{\frac{I'^{0}_n}{I'_n}}, \tag{5.24}$$

where

$$I_n'^0 = \frac{1}{K_n}\left|\sum_{m=1}^{K_n} z_{nm}^0\right|^2. \tag{5.25}$$

Finally, substituting equation (5.24) into equation (5.22) gives

$$z_{nk}^{\chi} = z_{nk}^0 + \frac{1}{K_n}\left(\sqrt{\frac{I_n'}{I_n'^0}} - 1\right)\sum_{m=1}^{K_n} z_{nm}^0. \tag{5.26}$$

This results in the reciprocal space projection given by

$$P_M g_{nk}(\mathbf{x}) = \mathcal{F}^{-1}\left\{G_{nk}(\mathbf{h}) + \frac{1}{K_n}\left(\frac{\sqrt{I_n}}{|F(\mathbf{h})|} - 1\right)F(\mathbf{h})\right\}. \tag{5.27}$$

5.4.2.2 Real space projection

The real space projection starts by merging the object estimates of $g_{nk}(\mathbf{x})$ and applying other real space constraints such as the support constraint,

$$g'(\mathbf{x}) = \frac{s(\mathbf{x})}{P}\sum_{n=1}^{N}\sum_{k=1}^{K_n} g_{nk}(R_{nk}^T(\mathbf{x} - \mathbf{t}_{nk})), \tag{5.28}$$

and then repositioning to give

$$P_S g_{nk}(\mathbf{x}) = g'(R_{nk}\mathbf{x} + \mathbf{t}_{nk}). \tag{5.29}$$

Inspection of equation (5.28) shows the advantage of implementation B. The projection uses the isolated, positioned molecules $g_{nk}(\mathbf{x})$, which are not subject to overlap, even if they overlap in the unit cell. The projection given by equation (5.29) can therefore be applied in the presence of overlap. Note that equation (5.28) is identical to equations (5.8) and (5.10) of implementation A in the case of no overlap.

5.5 SIMULATION METHODS

Simulations illustrating the *ai*MR method for a protein molecule using implementation B are presented in the next section. For convenience, specific background material to the simulations in Section 5.6 is gathered in this section.

5.5.1 Simulated data

Simulations were conducted with two crystal forms of the Gallus gallus hen egg-white lysozyme protein (HEWL), the second protein structure to be solved by X-ray crystallography, and remaining today a popular model for protein X-ray crystallographic

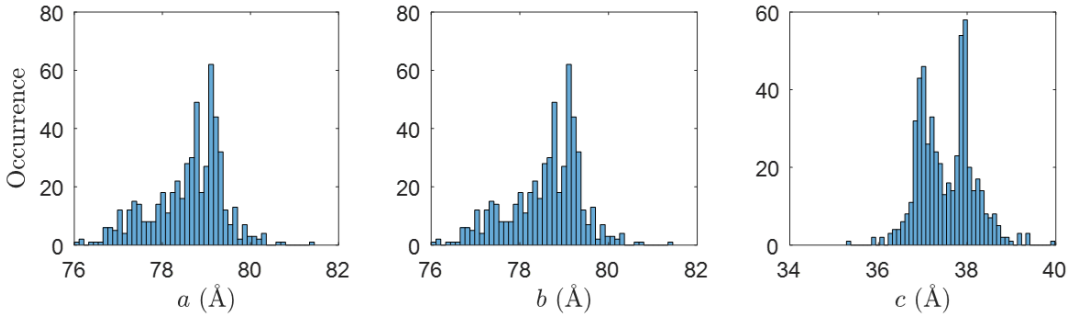


Figure 5.7 Histograms of the distribution of unit cell parameters for HEWL structures with space group $P4_32_12$ and $K_n = 8$ in the PDB.

studies. Due to the relatively small, globular shape of this protein, structures of HEWL in different crystal forms, including with significant space group changes, are easily found in the PDB.

A search through the PDB returned 658 structures of HEWL solved by X-ray crystallography. Amongst these HEWL structures, nine space groups were represented, with $P4_32_12$ constituting 84% of the entries. For simulation purposes, a number of possibilities can be considered involving the minimum number of crystal forms needed in *ai*MR:

- Using two or more crystal forms in the same space group.
- Using two or more crystals with different space groups.

Consider first the case with crystals forms in space group $P4_32_12$. The unit cell parameter ranges for structures with space group $P4_32_12$, with 8 copies of the protein in the unit cell, are shown in the Fig. 5.7. Most of the unit-cells dimensions are within about 3% of the average unit-cell size. This can be explained by the limited range of inter-protein distances available for globular shaped proteins before the crystal looses integrity. The largest unit cell volume is $2.5 \times 10^5 \text{Å}^3$ and the smallest is $2.1 \times 10^5 \text{Å}^3$, a change of about 15%. According to equation (5.4), and assuming two crystal forms of the median solvent content of 39%, and assuming $\varepsilon = 0.8$, the constraint ratio is $\Omega_{aiMR} = 1.28$, a value that might not be sufficient to obtain *ai*MR reconstructions.

Now, consider the case of crystal forms of HEWL in different space groups. Details of two particular structures, in space groups $P2_12_12_1$ and $P4_32_12$ are listed in Table 5.1 and shown in Fig. 5.8. The constraint ratio for these structures (for $\varepsilon = 0.9$) is $\Omega_{aiMR} = 1.55$ a value that is more likely to be sufficient for *ai*MR reconstruction than for the previous case. These two structures were thus used for the simulations.

Table 5.1 Details of the HEWL crystal forms used in the simulations.

Structure	132L [Rypniewski et al., 1993]	193L [Vaney et al., 1996]
Space group	$P2_12_12_1$	$P4_32_12$
a (Å)	30.6	78.5
b (Å)	56.3	78.5
c (Å)	73.2	37.8
α, β, γ	90°	90°
p_n	55.9%	60.5%
K_n	4	8

Figure 5.8 Crystals of the structures 132L (right) and 193L (left) of the HEWL protein used in simulations.

5.5.2 Structural homology

The structural homology between the two HEWL structures was characterised using the structure comparison tools in UCSF Chimera [Pettersen et al., 2004]. A C^α superposition of 193L to the reference structure 132L using the MatchMaker tool gave a C^α root-mean-square deviation (RMSD) of 0.83 Å, indicating quite similar structures.

The structural differences are illustrated by the differences between the C^α of each residue of 193L and 132L shown in Fig. 5.9. Overlaid in black is the reference structure 132L. As can be seen, the structural differences are concentrated near the edges of the proteins, while the core structure, is preserved in the crystal forms.

5.5.3 Molecular and crystal models

Although the structures of 132L and 193L could be used directly, in initial experiments a simplified structure was used as follows.

First, solvent molecules and ligands (sodium and cloride ions) were removed. The resulting structures were superimposed using *procrustes analysis* [Kroonenberg et al., 2003], giving the rotation matrix and translation vector between the asymmetric units $g_{11}(\mathbf{x})$ and $g_{21}(\mathbf{x})$. By choosing the reference structure $g(\mathbf{x})$ as $g_{11}(\mathbf{x})$, the rotation matrices R_{nk} and translation vectors $\mathbf{t}_{nk}$ are determined, i.e. $R_{11} = I$ and $\mathbf{t}_{11} = \mathbf{0}$.

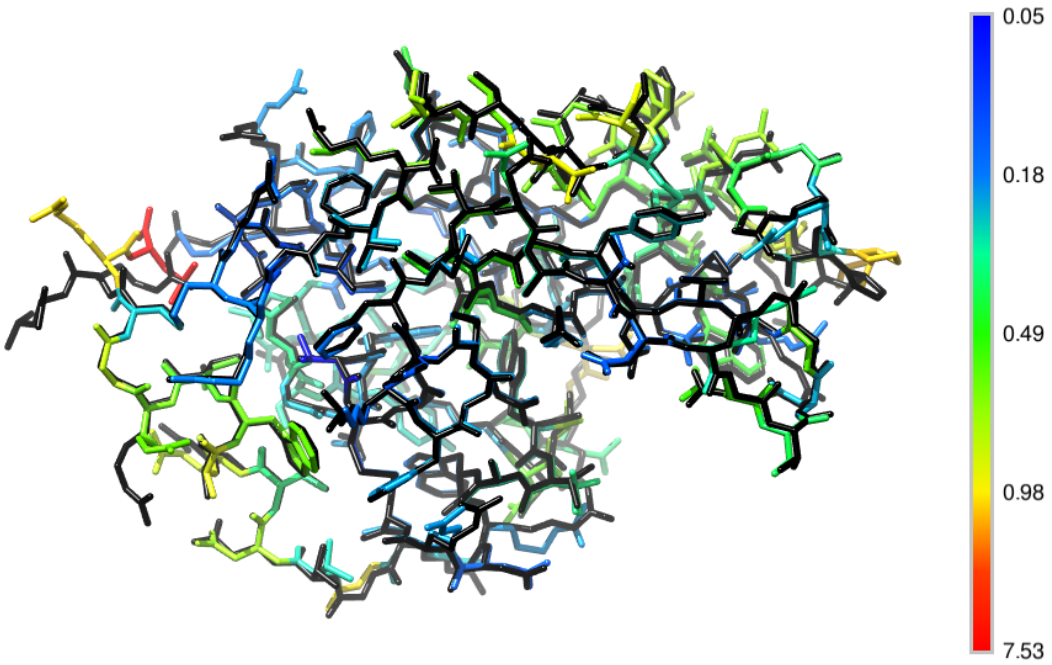


Figure 5.9 A C^{α} superposition of 193L to 132L. The color scale indicates the RMSD in Å of each alpha carbons of 193L taking 132L (black) as the reference structure.

Table 5.2 Electron density grids sizes and spacings in Å.

Crystal form	132L	193L
Number of samples along a (sample spacing Å)	32 (0.956)	80 (0.982)
Number of samples along b (sample spacing Å)	60 (0.938)	80 (0.982)
Number of samples along c (sample spacing Å)	80 (0.915)	40 (0.944)

Finally, the RMSD was reduced to zero between the two structures by replacing the structure of 193L by the translated and rotated version of the version of 132L. After these steps, two crystal forms of the same version of the molecule 193L were obtained.

The electron density of both crystals was then computed from their atomic coordinates. The command line tool phenix.fmodel was used to obtain the structure factors with the high-resolution parameter set to 2 Å and using an electron scattering table. No anisotropic scale matrix, flat bulk solvent model parameters or anisoScale was used in this case. The mtz file output was then converted into a ccp4 map with phenix.mtz2map with a grid resolution of 0.5 Å, to obtain a map sampled to at most $2 \times 0.5 = 1$Å spacing, which was then sigma-scaled. The electron density map details are given in Table 5.2.

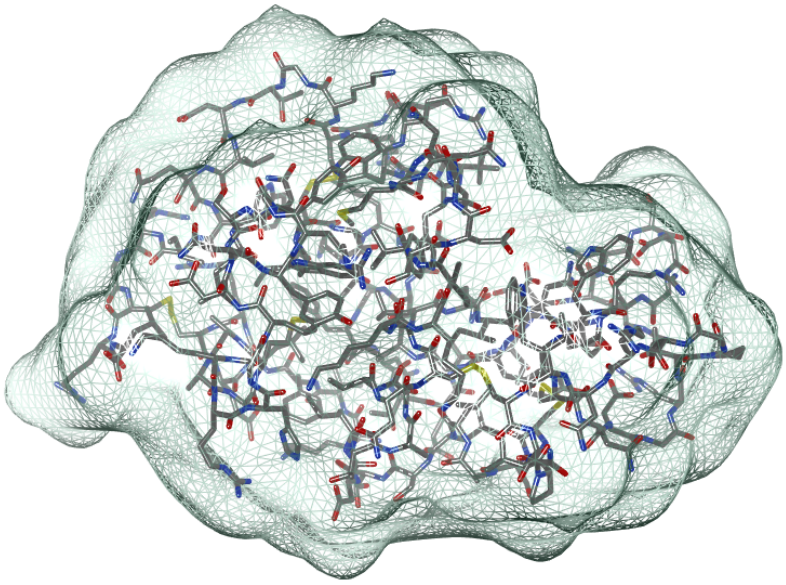


Figure 5.10 Envelope of 193L used in the simulations with the protein stick model of 193L fitted inside.

5.5.4 Determination of the envelope

The protein envelope (support) constitutes one of the *a priori* constraints used in *ab initio* MR. The envelope was computed from the atomic coordinates by first creating a logical mask of the protein atomic positions onto a grid and subsequently smoothing the logical mask with a Gaussian. The resulting mask was then thresholded to obtain an envelope containing no holes. The envelope obtained is determined by the standard deviation σ of the Gaussian while the threshold was chosen to obtain an envelope containing no holes. The envelope used in the simulation was calculated using $\sigma = 4$Å, and is shown in Fig. 5.10. Placing this envelope in the two unit cell gives solvent contents of about 0.24 for 193L and 0.33 for 132L. These solvent contents are smaller than those of the structure itself, so the envelope can by considered "generous".

5.5.5 Implementation of the real space merging

The merging step described in Section 5.4.2.2 is a crucial step in *ai*MR. It involves averaging the electron density from two electron densities sampled on different grids and thus requires potentially computationally intensive multivariate interpolation. For successful phasing, good interpolation accuracy is required. As this is an intricate procedure, and the bottleneck step in the *ai*MR, its implementation is described below.

The real space merging step can be broken down into three substeps that are shown as diagrams in Fig. 5.11:

1. Averaging within the same crystal form.
2. Averaging between the crystal forms.

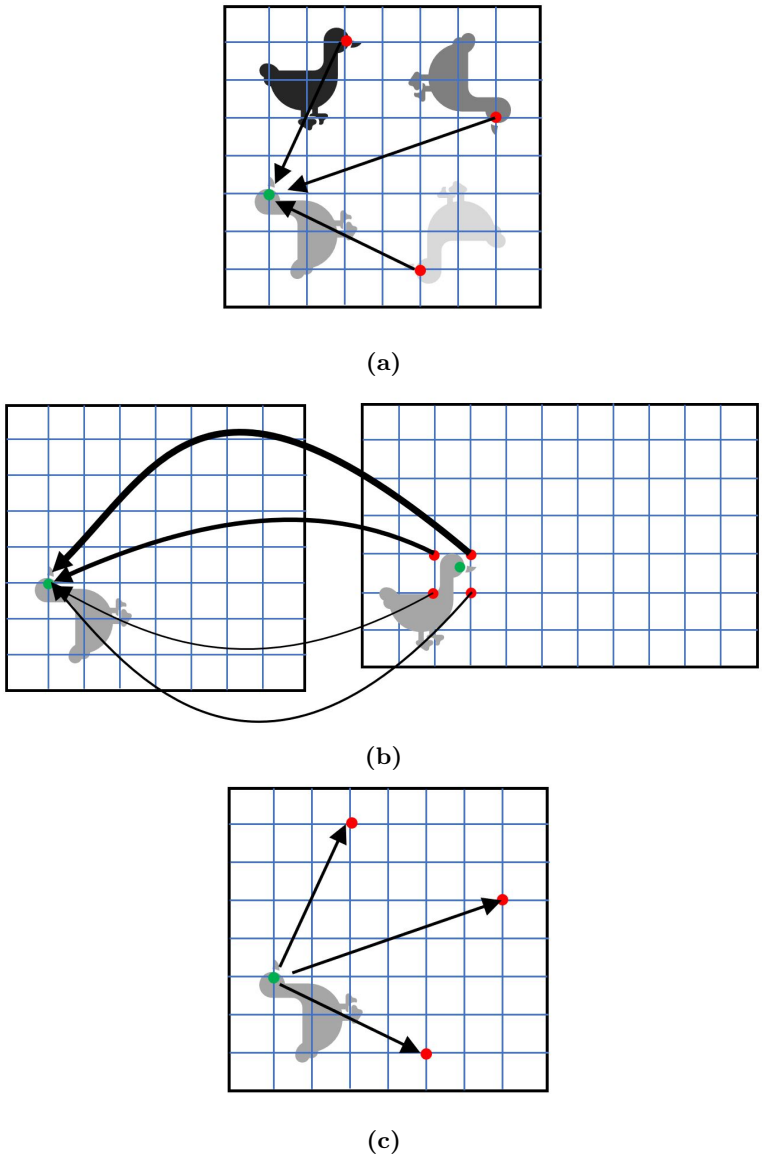


Figure 5.11 Diagrams of the three steps involved in the real space merging. (a) Averaging within the same crystal form. (b) Averaging between the crystal forms. (c) Redistributing the averaged electron density to each crystal form.

3. Redistributing the averaged electron density in each crystal form.

The first and last substeps are done on the same computational grids, interpolations and sampling issues can be avoided by working in the crystal basis of the respective crystal forms. Averaging within the same crystal form is shown in Fig. 5.11(a) and reduces to averaging the pixel values from all the copies (the sample locations for each copy are the same). Redistributing the averaged electron density in each crystal form is the inverse procedure and is shown in Fig. 5.11(c).

The second substep must be approached differently. Because the sampling grids are different between crystal forms, the sampling of the protein electron density is also different. This case is shown in Fig. 5.11(b). To average the electron density between crystal forms each pixel in the support of the protein in the first crystal form are mapped using the rotation and translation operators to a similar location in the second crystal form (for instance the green position in Fig. 5.11(b)). The closest pixels (red positions) are then used to linearly interpolate the value at the matching position.

5.6 SIMULATION RESULTS

Diffraction amplitudes for the two crystal forms calculated as described in Section 5.5.3, the molecular envelope, and the rotation R_{nk} and translation t_{nk} operators are assumed known. Implementation B described in Section 5.4.2 was used. The intensity data $I_n(\mathbf{h})$ were obtained from the true electron densities $f_n(\mathbf{x})$. Noise was not added to the intensity data as the objective here was to investigate the uniqueness and quality of the reconstructions under ideal conditions. The difference map algorithm and error reduction algorithms were used to iteratively apply the real space and reciprocal space projections described in Section 5.4.2. A positivity constraint was not applied. The difference map parameter $\beta = 0.9$ was used.

This algorithm was implemented using MATLAB® and run on an INTEL® i7-4600M quad-core processor. A generic run with 1800 DM iterations and 200 ER iteration takes approximately 4 hours.

Convergence was monitored by calculating the RMS difference between the amplitude data for crystal form n and that of the estimated amplitude of its reconstruction at iteration i. This error metric is defined by

$$E^i = \sqrt{\frac{\sum_n \sum_{\mathbf{h}} (\sqrt{I_n(\mathbf{h})} - \sqrt{I_n^i(\mathbf{h})})^2}{\sum_n \sum_{\mathbf{h}} I_n(\mathbf{h})}}, \tag{5.30}$$

where $I_n(\mathbf{h})$ is the Fourier intensity data for crystal form n and $I_n^i(\mathbf{h})$ is the estimated Fourier intensity of crystal form n at iteration i. A small value for E indicates a converging algorithm.

The quality of the reconstructed electron density was monitored by calculating the RMS difference between the correct electron density, and the estimated electron density

of the reconstruction at iteration i, given as,

$$e^i = \sqrt{\frac{\sum_k \sum_n \sum_{\mathbf{x}} (g(\mathbf{x}) - g^i_{nk}(\mathbf{x}))^2}{\sum_k \sum_n \sum_{\mathbf{x}} g^2(\mathbf{x})}}, \tag{5.31}$$

where $g^i_{nk}(\mathbf{x})$ is the electron density of the reconstruction of protein k in crystal form n at iteration i. A small value for e indicates a good reconstruction.

The final reconstruction was chosen as the iterate where the Fourier space error E^i is smallest. Small values for e and E indicate a successful and unique solution to the phase retrieval problem. A small value for E and high value for e indicates either an ambiguity or a non-unique problem. Inversion ambiguities were checked by computing e with $g(\mathbf{x})$ replaced by $g(-\mathbf{x})$.

The algorithm was started with a random electron density map and the difference map algorithm run for 250 iterations followed by 25 iterations of the error reduction algorithm, with this pattern repeated until convergence. Convergence was defined as when the Fourier space error E reached 8×10^{-2}, a value which, it was determined, results in an interpretable electron density map. The algorithm converged and uniquely

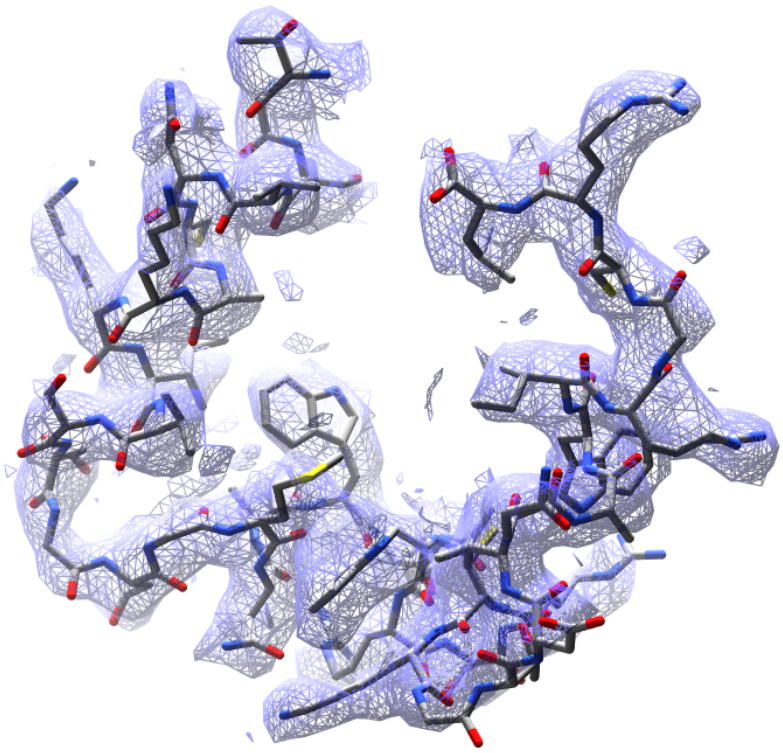


Figure 5.12 Reconstructions of residues 89 to 129 of the HEWL protein obtained with *ai*MR.

reconstructed the electron density. The reconstruction of the first crystal form is shown in Fig. 5.12. For this reconstruction, the final errors are $e = 0.20$ and $E = 7.6 \times 10^{-2}$.

As can be seen, the fit of the protein structure to the electron density is good and of sufficient quality for initial chain tracing (threading of the protein molecule into the electron density). Representative plots of the error metrics versus iteration are shown in Fig. 5.13. The random starting point of the simulation is evidently responsible for

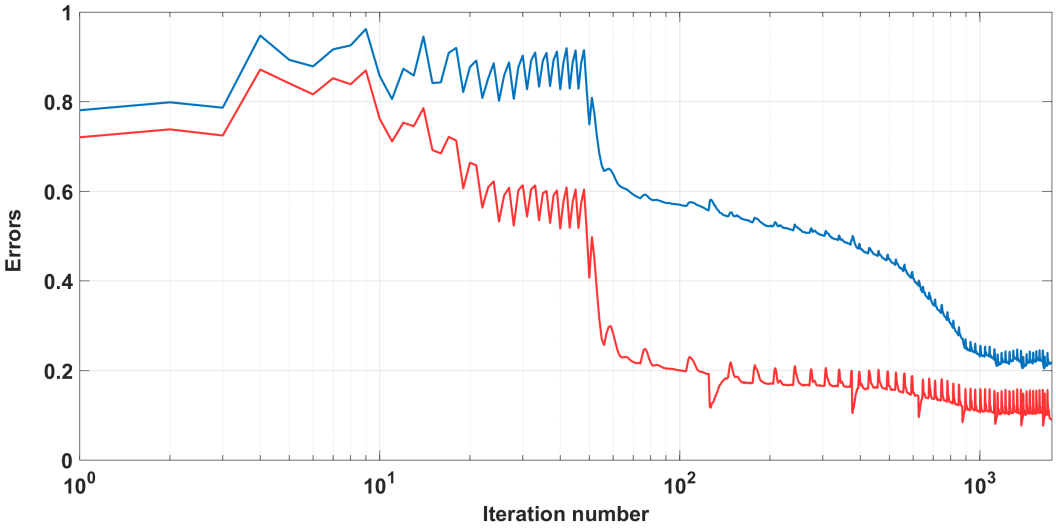


Figure 5.13 Real (blue) and reciprocal (red) space errors versus iteration for the simulations described in the text.

the high starting values for E and e (in the first 10 iterations). This is followed by a small decrease of the Fourier amplitude error while the real space stays high (the next 40 iterations). A sharp decrease in both errors is then observed, followed by a slow but steady decrease in both errors. This behaviour is typical of the convergence of IPAs.

5.7 DISCUSSION

The *ai*MR technique offers new phasing possibilities in protein crystallography. The technique's attractiveness lies in it being immune to model bias and its ability to find new folds. The need for knowledge of the molecular envelope $s(\mathbf{x})$ and transformations $(R_{nk},\ \mathbf{t}_{nk})$ between the crystal forms is a limitation of the current study, but it is likely that these restrictions can be relaxed. For example, if a low resolution electron density was available, from solution scattering or electron microscopy, for example, then the envelope may be able to be refined during the reconstruction using the shrink-wrap technique [Marchesini et al., 2003]. It may be possible to estimate the required rotation operators using Patterson techniques discussed in Section 1.3.2.4. It may then

be possible to search for the translations. Information may also be available from a failed MR attempt. For the case of similar unit cells, it is likely that the position and orientation changes between unit cells are correlated with the change in the unit cell dimensions, simplifying the search. Data from different crystal forms can sometimes be collected in a single experimental setup, and recent work with XFELs and a fixed target sample delivery has shown the collection of such data is possible using the natural variation in humidity across the target chip.

6 CONCLUSIONS AND FURTHER RESEARCH SUGGESTIONS

The advent of X-ray free electron lasers, with their extreme brightness, ultra-short pulses and megahertz repetition rate is offering new opportunities for the study of a variety of new samples (2D crystals, fibers, nanocrystals, and single particles) and the development of new techniques (SFX, SPI) in protein X-ray crystallography. The phase problem is not exempt from this revolution, and the problem can be potentially eased if continuous diffraction can be measured along one or more dimensions in reciprocal space. For the case of 2D crystals, the additional intensity data falls short of rendering the solution to the phase problem unique, so that additional real space data or knowledge on the sample is still needed. A parameter, denoted by Λ'_{2dc}, which depends on the shape of the molecular envelope and the resolution was defined and proved more useful than the usual constraint ratio Ω_{2dc} for determining the uniqueness of the solution.

Ab initio phasing algorithms such as IPAs are ideal candidates to take advantage of this source of additional information. In fact, because of the limited access to, and running costs of XFEL sources, phase retrieval techniques that do not require additional beam time are advantageous. The absence of model bias and ability to solve new structural folds are also welcomed. Unfortunately, *ab initio* phasing algorithms do not assure a unique solution to the phase problem. Calculation of the constraint ratio can help to determine if the solution to a phase problem is likely to be unique, and how constrained the problem is.

A new phasing approach, the *ai*MR phasing algorithm, described in Chapter 5, has been developed. This approach has shown to be feasible in noiseless simulations and has considerable potential. Here, again, a constraint ratio, Ω_{aiMR}, offers the best indicator for the possibility of *ab initio* phase retrieval. The approach is not affected by bias and can find new structural folds contrary to MR. It was shown that a difference of 5–10% between crystal forms is often sufficient for the diffraction data to be mostly independent.

6.1 FURTHER RESEARCH SUGGESTIONS

1. The volume constraint is shown in Chapter 2 to be difficult to use in practice with 3D crystal diffraction data. This is due to the large number of hyperplanes making up the solution manifold. Extrapolating to any realistic computation grids size, the search for the solution in the extremely non-convex space is bound to fail. For 2D crystals, the solution space is also extremely non-convex but the membrane is more structured reducing somewhat this complexity.

   Application of the volume constraint to the 2D crystal phase problem may thus have different outcomes to that of the 3D crystal phase problem. Research in this direction could lead to a useful *ab initio* algorithm using only the volume information that is easily obtainable in practice.

   The shrink-wrap technique could be used with the volume constraint as the protein extends from the membrane in a connected manner. Simulations of *ab initio* phase retrieval with the derived shrink-wrap volume projection could be used to investigate potential uses.

2. The change of unit cell size between crystal forms can be discrete as in a change of space group, but can also be continuous as in the swelling of a crystal with changes of humidity. Chapter 5 only considered a discrete and rather important change between the crystal forms but, in practice, collecting data from two different forms might require changing the experimental set-up and thus lead to longer collection times.

   In the case of a continuum of changes between two crystal forms, changes to the experimental set-up may be more easily introduced through varying the humidity. The small changes in unit cell dimensions might give direct access to the diffraction intensity gradients or be used to help determine the rotations and translations of the proteins within the crystal forms.

   A first order expression for the diffraction resulting from small unit cell changes could be derived for a few simplified cases in which only the orientation/ positions of the proteins or the unit cell parameters change. The continuous change in the diffraction intensities is a source of additional information that could potentially be used to help solve or ease the phase problem.

3. The simulations of *ai*MR in Chapter 5 were limited due to the absence of noise considerations. Addition of noise in the diffraction data will affect the *ai*MR approach, but uncertainties over the rotation and translations of the proteins in the crystal forms and the unknown structural changes of the proteins might be even bigger issues. Finally the *ai*MR approach is dependent on the knowledge of the molecular envelope and original positions and orientations of the proteins in

the crystal forms.

To make the *ai*MR method practical for *ab initio* phasing, all of the above considerations needs to be incorporated into the reconstruction algorithm. Testing on experimental data will then be required.

4. In Chapters 2, 3, and 4 *ab initio* phasing retrieval algorithms suffered from the highly non-convex solution space and the curse of dimensionality. Furthermore the weakness of some of the constraints is slowing down convergence.

   The simulations are computationally extensive and larger problems cannot in practice be solved on a single laptop as is the case with molecular replacement.

   The iterative projection algorithms should be implemented in tailored hardware GPU, FPGAs or supercomputers to tackle larger problems.

REFERENCES

[Abergel, 2013] Abergel, C. (2013). Molecular replacement: tricks and treats. *Acta Cryst. D*, 69:2167 – 2173.

[Allahgholi et al., 2015] Allahgholi, A., Becker, J., Bianco, L., Delfs, A., Dinapoli, R., et al. (2015). AGIPD, a high dynamic range fast detector for the European XFEL. *J. Instrum.*, 10:C01023.

[Aquila et al., 2015] Aquila, A., Barty, A., Bostedt, C., Boutet, S., Carini, G., et al. (2015). The linac coherent light source single particle imaging road map. *Structural Dynamics*, 2:041701.

[Barty et al., 2014] Barty, A., Kirian, R., Maia, F., Hantke, M., Yoon, C., White, T., and Chapman, H. (2014). Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *J. Appl. Crystallogr.*, 43:1118 – 1131.

[Berman et al., 2000] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., et al. (2000). The protein data bank. *Nucleic Acids Res.*, 28:235 – 242.

[Blundell and Patel, 2004] Blundell, T. and Patel, S. (2004). High-throughput X-ray crystallography for drug discovery. *Curr. Opin. Pharmacol.*, 4:490 – 496.

[Boutet et al., 2018] Boutet, S., Fromme, P., and Hunter, M. (2018). *X-ray Free Electron Lasers: A Revolution in Structural Biology.* Springer, Cham.

[Bricogne, 1974] Bricogne, G. (1974). Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Cryst. A*, 30:395 – 405.

[Brito and Archer, 2013] Brito, J. A. and Archer, M. (2013). *Practical Approaches to Biological Inorganic Chemistry.* Elsevier, Oxford.

[Caffrey, 2015] Caffrey, M. (2015). A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes. *Acta Cryst. F.*, 71:3 – 18.

[Carter et al., 1990] Carter, C., Crumley, K., Coleman, D., Hage, F., and Bricogne, G. (1990). Direct phase determination for the molecular envelope of tryptophanyl-trna synthetase from bacillus stearothermophilus by x-ray contrast variation. *Acta Cryst. A*, 46:57 – 68.

[Chapman, 2009] Chapman, H. (2009). X-ray imaging beyond the limits. *Nat. Mater.*, 8:299 – 301.

[Chapman et al., 2014] Chapman, H., Caleman, C., and Timneanu, N. (2014). Diffraction before destruction. *Philos. Trans. Royal Soc. B*, 369:20130313.

[Chapman et al., 2011] Chapman, H., Fromme, P., Barty, A., White, T., Kirian, R., Aquila, A., Hunter, M., Schulz, J., DePonte, D., Weierstall, U., and Doak, R. (2011). Femtosecond X-ray protein nanocrystallography. *Nature*, 470:73 – 77.

[Chayen, 2004] Chayen, N. (2004). Turning protein crystallisation from an art into a science. *Curr. Opin. Struct. Biol.*, 14:577 – 583.

[Chen et al., 2016] Chen, J., Arnal, R., Morgan, A., Bean, R., Beyerlein, K., Chapman, H., Bones, P., Millane, R. P., and Kirian, R. (2016). Reconstruction of an object from diffraction intensities averaged over multiple object clusters. *J. Opt.*, 18:114003.

[Crowther, 1969] Crowther, R. A. (1969). The use of non-crystallographic symmetry for phase determination. *Acta Cryst. B*, 25:2571 – 2580.

[DePonte et al., 2008] DePonte, D., Weierstall, U., Schmidt, K., Warner, J., Starodub, D., Spence, J., and Doak, R. (2008). Gas dynamic virtual nozzle for generation of microscopic droplet streams. *J. Phys. D*, 41:195505.

[Drenth, 2007] Drenth, J. (2007). *Principles of Protein X-ray Crystallography*. Springer Science & Business Media, New York.

[Elser, 2003a] Elser, V. (2003a). Phase retrieval by iterated projections. *J. Opt. Soc. Am. A*, 20:40 – 55.

[Elser, 2003b] Elser, V. (2003b). Random projections and the optimization of an algorithm for phase retrieval. *J. Phys. D*, 36:2995 – 3007.

[Elser, 2003c] Elser, V. (2003c). Solution of the crystallographic phase problem by iterated projections. *Acta Cryst. A*, 59:201 – 209.

[Elser and Millane, 2008] Elser, V. and Millane, R. P. (2008). Reconstruction of an object from its symmetry-averaged diffraction pattern. *Acta Cryst. A*, 64:273 – 279.

[Evans and McCoy, 2008] Evans, P. and McCoy, A. (2008). An introduction to molecular replacement. *Acta Cryst. D*, 64:1 – 10.

[Fienup, 1982] Fienup, J. (1982). Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21:2758 – 2769.

[Friedrich et al., 1912] Friedrich, W., Knipping, P., and von Laue, M. (1912). Interferenz-Erscheinungen bei Röntgenstrahlen. *Sitzungsberichte der Mathematisch-Physikalischen Classe der Kniglich-Bayerischen Akademie der Wissenschaften zu München*, 92:303 – 322.

[Fromme, 2015] Fromme, P. (2015). XFELs open a new era in structural chemical biology. *Nat. Chem. Biol.*, 11:895 – 899.

[Giacovazzo, 1999] Giacovazzo, C. (1999). *Direct Phasing in Crystallography: Fundamentals and Applications*. Oxford University Press, Oxford.

[Hao, 2006] Hao, Q. (2006). Macromolecular envelope determination and envelope-based phasing. *Acta Cryst. D*, 62:909 – 914.

[Harker and Kasper, 1948] Harker, D. and Kasper, J. (1948). Phases of Fourier coefficients directly from crystal diffraction data. *Acta Crystallographica*, 1:70 – 75.

[Hasnain, 2015] Hasnain, S. S. (2015). Crystallography in the 21st century. *IUCrJ*, 2:602 – 604.

[He and Su, 2015] He, H. and Su, W. (2015). Direct phasing of protein crystals with high solvent content. *Acta Cryst. A*, 71:92 – 98.

[Henderson, 1990] Henderson, R. (1990). Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. *Proc. Royal Soc. Lond.*, 241:6 – 8.

[Hendrickson et al., 1990] Hendrickson, W., Horton, J., and LeMaster, D. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of threedimensional structure. *The EMBO Journal*, 9:1665 – 1672.

[Henrich et al., 2011] Henrich, B., Becker, J., Dinapoli, R., Goettlicher, P., Graafsma, H., Hirsemann, H., Klanner, R., Krueger, H., Mazzocco, R., Mozzanica, A., and Perrey, H. (2011). The adaptive gain integrating pixel detector AGIPD a detector for the European XFEL. *Nucl. Instrum. Methods Phys. Res.*, 633:S11 – S14.

[Hilbert and Cohn-Vossen, 1999] Hilbert, D. and Cohn-Vossen, S. (1999). *Geometry and the Imagination (No. 87)*. American Mathematical Soc, Providence, Rhode Island.

[Holton and Frankel, 2010] Holton, J. and Frankel, K. (2010). The minimum crystal size needed for a complete diffraction data set. *Acta Cryst. D*, 64:393 – 408.

[Hunter et al., 2014] Hunter, M., Segelke, B., Messerschmidt, M., Williams, G., Zatsepin, N., Barty, A., Benner, W., Carlson, D., Coleman, M., Graf, A., and Hau-Riege, S. (2014). Fixed-target protein serial microcrystallography with an x-ray free electron laser. *Sci. Rep.*, 4(6026).

[Ishchenko et al., 2018] Ishchenko, A., Gati, C., and Cherezov, V. (2018). Structural biology of G protein-coupled receptors: New opportunities from XFELs and cryoEM. *Curr. Opin. Struct. Biol.*, 51:44 – 52.

[Jaskolski et al., 2014] Jaskolski, M., Dauter, Z., and Wlodawer, A. (2014). A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *The FEBS Journal*, 281:3985 – 4009.

[Johansson et al., 2017] Johansson, L., Stauch, B., Ishchenko, A., and Cherezov, V. (2017). A bright future for serial femtosecond crystallography with XFELs. *Trends Biochem. Sci.*, 42:749 – 762.

[Kendrew et al., 1958] Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., and Phillips, D. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662 – 666.

[Khatter et al., 2015] Khatter, H., Myasnikov, A., Natchiar, S., and Klaholz, B. (2015). Structure of the human 80S ribosome. *Nature*, 520:640 – 645.

[Kirian et al., 2010] Kirian, R.A.and Wang, X., Weierstall, U., Schmidt, K., Spence, J., Hunter, M., Fromme, P., et al. (2010). Femtosecond protein nanocrystallographydata analysis methods. *Opt. Express*, 18:5713 – 5723.

[Kroonenberg et al., 2003] Kroonenberg, P., Dunn, W., and Commandeur, J. (2003). Consensus molecular alignment based on generalized procrustes analysis. *J. Chem. Inf. Comp. Sci.*, 43:2025 – 2032.

[Landau and Rosenbusch, 1996] Landau, E. and Rosenbusch, J. (1996). Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. *Proc. Natl. Acad. Sci.*, 93:14532 – 14535.

[Liu et al., 2012] Liu, Z., Xu, R., and Dong, Y. (2012). Phase retrieval in protein crystallography. *Acta Cryst. A*, 68:256 – 265.

[Lo et al., 2015] Lo, V., Kingston, R., and Millane, R. P. (2015). Iterative projection algorithms in protein crystallography. II. Application. *Acta Cryst. A*, 71:451 – 459.

[Lo et al., 2009] Lo, V., Kingston, R. L., and Millane, R. P. (2009). Determination of molecular envelopes from solvent contrast variation data. *Acta Cryst. A*, 65:312 – 318.

[Lomb et al., 2011] Lomb, L., Barends, T. R., Kassemeyer, S., Aquila, A., Epp, S. W., Erk, B., and Schlichting, I. (2011). Radiation damage in protein serial femtosecond crystallography using an x-ray free-electron laser. *Phys. Rev. B*, 84:214111.

[Marchesini et al., 2003] Marchesini, S., He, H., Chapman, H. N., Hau-Riege, S. P., Noy, A., Howells, M. R., Weierstall, U., and Spence, J. C. H. (2003). X-ray image reconstruction from a diffraction pattern alone. *Phys. Rev. B*, 68:140101.

[Millane, 1990] Millane, R. P. (1990). Phase retrieval in crystallography and optics. *J. Opt. Soc. Am. A*, 7:394 – 411.

[Millane, 1993] Millane, R. P. (1993). Phase problems for periodic images: effects of support and symmetry. *J. Opt. Soc. Am. A*, 10:1037 – 1045.

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

[Millane and Lo, 2013] Millane, R. P. and Lo, V. (2013). Iterative projection algorithms in protein crystallography. I. Theory. *Acta Cryst. A*, 69:517 – 527.

[Millane and Stroud, 1997] Millane, R. P. and Stroud, W. (1997). Reconstructing symmetric images from their undersampled Fourier intensities. *J. Opt. Soc. Am. A*, 14:568 – 579.

[Neutze et al., 2000] Neutze, R., Wouts, R., Van der Spoel, D., Weckert, E., and Hajdu, J. (2000). Potential for biomolecular imaging with femtosecond x-ray pulses. *Nature*, 406:752 – 757.

[Nickolls et al., 2008] Nickolls, J., Buck, I., and Garland, M. (2008). Scalable parallel programming. *IEEE Hot Chips 20 Symposium (HSC)*, pages 40 – 53.

[Oberthür, 2018] Oberthür, D. (2018). Biological single-particle imaging using XFELs - towards the next resolution revolution. *IUCrJ*, 5:663 – 666.

[Oberthür et al., 2017] Oberthür, D., Knoska, J., Wiedorn, M., Beyerlein, K., Bushnell, D., Kovaleva, E., Heymann, M., and Gumprecht, L. (2017). Double-flow focused liquid injector for efficient serial femtosecond crystallography. *Sci. Rep.*, 7:44628.

[Pettersen et al., 2004] Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., and Ferrin, T. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25:1605 – 1612.

[Prasad et al., 1999] Prasad, B., Hardy, M., Dokland, T., Bella, J., Rossmann, M., and Estes, M. (1999). X-ray crystallographic structure of the Norwalk virus capsid. *Science*, 286:287 – 290.

[Prince, 2006] Prince, E. (2006). *International Tables for Crystallography, Volume C: Mathematical, physical and chemical tables*. International Tables for Crystallography. Kluwer Academic Publishers, Dordrecht.

[Robin et al., 2016] Robin, L. O., Jordi, J., and Martin, F. (2016). Current advances in synchrotron radiation instrumentation for mx experiments. *Arch. Biochem. Biophys.*, 602:21 – 31.

[Roedig et al., 2015] Roedig, P., Vartiainen, I., Duman, R., Panneerselvam, S., et al. (2015). A micro-patterned silicon chip as sample holder for macromolecular crystallography experiments with minimal background scattering. *Sci. Rep.*, 5:10451.

[Rossmann, 1972] Rossmann, M. G. (1972). *The Molecular Replacement Method*. Gordon and Breach, New York.

[Rypniewski et al., 1993] Rypniewski, W., Holden, H., and Rayment, I. (1993). Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: An x-ray analysis at 1.8 Å resolution. *Biochemistry*, 32:9851 – 9858.

[Spence, 2017] Spence, J. (2017). XFELs for structure and dynamics in biology. *IUCrJ*, 4:322 – 339.

[Spence and Doak, 2004] Spence, J. and Doak, R. (2004). Single molecule diffraction. *Phys. Rev. Lett.*, 92:198102.

[Tong and Rossmann, 1997] Tong, L. and Rossmann, M. G. (1997). Rotation function calculations with glrf program. *Meth. Enzymol.*, 276:594 – 611.

[Usón and Sheldrick, 1999] Usón, I. and Sheldrick, G. (1999). Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.*, 9:643 – 648.

[Vaney et al., 1996] Vaney, M., Maignan, S., Ries-Kautt, M., and Ducruix, A. (1996). High-resolution structure (1.33 Å) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. *Acta Cryst. D*, 52:505 – 517.

[Verschueren et al., 1993] Verschueren, K., Selje, F., Rozeboom, H., Kalk, K., and Dijkstra, B. (1993). Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase. *Nature*, 363:693 – 698.

[Wang, 1985] Wang, B. C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Meth. Enzym.*, 115:90 – 112.

[Weichenberger and Rupp, 2014] Weichenberger, C. X. and Rupp, B. (2014). Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. *Acta Cryst. D*, 70:1579 – 1588.

[Weierstall et al., 2014] Weierstall, U., James, D., Wang, C., White, T., Wang, D., W., L., Spence, J., Doak, R., Nelson, G., Fromme, P., and Fromme, R. (2014). Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat. Commun.*, 5:3309.

[Woolfson, 1987] Woolfson, M. (1987). Direct methods from birth to maturity. *Acta Cryst. A*, 43:593 – 612.

[Zwick et al., 1996] Zwick, M., Lovell, B., and Marsh, J. (1996). Global optimization studies on the 1-D phase problem. *Int. J. Gen. Syst.*, 25:47 – 59.