An exploration of some aspects of molecular replacement in macromolecular crystallography

Richard Mifsud

Christ's College

Submitted: June 2018

This dissertation is submitted for the degree of Doctor of Philosophy

Declarations

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Acknowledgements

I would like to thank all those who helped fund my research that greatly increased my understanding of this field, allowed me to contribute to the wealth of knowledge, and enabled me to generate this thesis. I have been on the Cambridge Institute for Medical Research 4 year PhD Programme, funded by the Wellcome Trust. I would like to thank INSTRUCT for granting a travel fellowship to allow me to attend the Gordon Research Conference on structural biology in 2014.

I would like to give thanks to Professor Randy Read for allowing me to do a Doctor of Philosophy project in his group. He has provided invaluable advice throughout, and has helped guide and develop my project. He has also supported me throughout the project, providing funding to attend the informative *phenix* conferences held every six months.

I would like to thank all the scientists in the Read Group who have helped increase my understanding in several areas. In particular, Gabor Bunkozi for his assistance in explaining the various Python libraries in *phenix*, and the intricacies of some of the existing code. I would also like to thank him for his advice with the concepts of the *NaCelleS* project.

The initiative shown by Cavan Bennett and his group leader Cedric Ghevaert in introducing me to the issues that had arisen with crystallising the CRLF3 protein is gratefully acknowledged, as is their subsequent support and collaboration in determining the structure of major elements of this protein, as reported in this thesis.

The CRLF3 structural determination was collaborative, as it required a range of skills. I designed the constructs that were taken forward for experimentation. The cloning and expression of the protein was undertaken by Cavan, who also purified and prepared the crystals for data collection, working with Yahui Yan. The later preparation of heavy atom derivatives was also done by Yahui. The remote data collection was led by Yahui, with my direct involvement. I led the immediate and subsequent analysis of the collected data, and the solving of the phase problem. Model building and refinement was my work; and the data will shortly be deposited.

Airlie McCoy is another person I would like to thank. She has always cheerfully fixed any *Phaser* bugs I identified and has helped in a number of other ways. Rob Oeffner has been helpful in providing test data, helping to run programs in Windows and on the cluster. Tristan Croll has been a great support in the laboratory and the discussions with WeeLee Chan early in the project were helpful.

More broadly within the CIMR, I am very grateful to Jim Huntington and Andreas Floto for the laboratory rotations that I undertook. The CIMR programme has many two hour core seminars, which have helped me to gain a broader understanding, as well as various Wellcome Trust PhD days, which have allowed a broad overview into what other PhD students have been doing.

I would like to thank the following people for proof-reading my thesis. Beth Humphreys, John Mifsud, John Aldridge and Lucy Oswald, along with Gordon Frazer for showing me how to bind the thesis.

Finally, I would like to thank the CIMR Level 7 Canteen for coping with my innumerable allergies, Christ’s College for all the support they have shown throughout the years, and my long suffering girlfriend Lucy Oswald, for astronomical levels of patience.

Abstract

This thesis reports work in three areas of X-ray crystallography. An initial chapter describes the structure of a protein, the methods based on the use of X-rays and computer analysis of diffraction patterns to determine crystal structure, and the subsequent derivation of the structure of part or all of a protein molecule. Work to determine the structure of the protein cytokine receptor-like factor 3 (CRLF3) leading to the successful generation of a structural model of a significant part of this molecule is then described in Chapter 2. A variety of techniques had to be deployed to complete this work, and the steps undertaken are described. Analysis was performed principally using *phaser,* using maximum likelihood methods. Areas for improvement in generating non-crystallographic symmetry (NCS) operators in existing programmes were identified and new and modified algorithms implemented and tested. Searches based on improved single sphere algorithms, and a new two-sphere approach, are reported. These methods showed improvements in many cases and are available for future use. In Chapter 4, work on determining the relative importance of low resolution and high intensity data in molecular replacement solutions is described. This work has shown that high intensity data are more important than the low resolution data, dispelling a common perception and helping in experimental design.

Abbreviations

2D	two-Dimensional
3D	three-Dimensional
ASA	Accessible Surface Area
ATP	Adenosine TriPhosphate
BCA	BiCinchoninic Acid
BIS-TRIS	Bis-tris methane
BRIDGE	The BRIDGE Consortium is an organisation umbrella for Next Generation Sequencing at Cambridge
CASP	Critical Assessment of protein Structure Prediction
CCD	Charge-Coupled Device
CCP-EM	Collaborative Computational Project for Electron cryo-Microscopy
CCP4	Collaborative Computational Project Number 4 in protein crystallography
cDNA	Complementary DNA
CREME9	Cis-Regulatory Module Explorer for the human genome 9 (aka CRLF3)
CRLF3	Cytokine Receptor-Like Factor 3
CRLM9	Cytokine Receptor-Like Molecule 9 (aka CRLF3)
CYTOR4	Cytoskeleton Regulator 4 (aka CRLF3)
DNA	DeoxyriboNucleic Acid
DSSP	Database of Secondary Structures Program
EDTA	EthyleneDiamineTetraacetic Acid
EG	Ethylene Glycol
EGA	European Genome-phenome archive at the European Bioinformatics Institute
eLLG	Expected Log Likelihood Gain
EM	Electron Microscopy
EMPIAR	Electron Microscopy Public Image Archive
ExAC	Exome Aggregation Consortium
FN3	Fibronectin type III domain
FN3con	Fibronectin type III domain consensus protein
GCSF	Granulocyte colony-stimulating factor
GST	Glutathione S-Transferase

GTP	Guanosine-5’-TriPhosphate
Hg-SAD	Mercury atom single-wavelength anomalous diffraction
HMM	Hidden Markov model
HySS	Hybrid Substructure Search
IhhN	Indian hedgehog gene
Ihog	Interference hedgehog
IKMC	International Knock-out Mouse Consortium
IPTG	IsoPropyl β-D-1-ThioGalactopyranoside
IQ	Intelligence Quotient
iRNA	interfering RNA
IUCr	International Union of Crystallography
JAK-STAT	JAK (Janus kinases)-STAT (Signal Transducer and Activator of Transcription proteins) signalling pathway
LLG	Log-Likelihood Gain
MAD	Multiple-wavelength Anomalous Dispersion
MES	2-(N-Morpholino)EthaneSulfonic acid
MIR	Multiple Isomorphous Replacement
MIRAS	Multiple Isomorphous Replacement with Anomalous Scattering
mmCIF	Macromolecular Crystallographic Information File format
MR	Molecular Replacement
NCAM1	Neural Cell Adhesion Molecule 1
NCS	Non-Crystallographic Symmetry
NF1	NeuroFibromatosis type 1
NGS	Next Generation Sequencing
NHS	National Health Service
NIHR	National Institute for Health Research
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NRF	No Reported Factors
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	PolyEthylene Glycol
PFPE	PerFluoroPolyEther oil

phenix	Python-based Hierarchical ENvironment for Integrated Xtallography
PPI	Poly-Proline type I helix
PPII	Poly-Proline type II helix
PSF	Point Spread Function
REFMAC	REFinement of MACromolecular structures program
RF	Random Forest (algorithm)
RIP	Radiation-damage-Induced Phasing
RIPAS	Radiation-damage-Induced Phasing with Anomalous Scattering
RMS	Root-Mean-Square
RMSD	Root-Mean-Square Deviation
RNA	RiboNucleic Acid
ROBO1	ROundaBOut guidance receptor 1
SAD	Single-wavelength Anomalous Dispersion
S-SAD	Sulphur SAD
siRNA	Small interfering RNA
SIR	Single Isomorphous Replacement
SIRAS	Single Isomorphous Replacement with Anomalous Scattering
SPRY	domain named from SPla and the RYanodine Receptor
STRIDE	STRuctural IDEntification algorithm
SVN	SubVersioN (version control software)
TAE	40mM Tris, 20mM Acetic acid, 1mM EDTA
TFZ	Translation Function Z-score
TFZ0	Initial Z-score for the translation search
Tris	Tris(hydroxymethyl)aminomethane
XDS	X-ray Detector Software
XFEL	X-ray Free Electron Laser

Table of Contents

Table of Figures

Table of Tables

Chapter 1 Introduction

1.1 Summary of thesis

The manner in which X-ray crystallography can be used to obtain protein structures is presented in this introductory chapter. Protein structures and their use are discussed briefly, and challenges in using X-rays to obtain the necessary information to determine these structures are described. The following chapters in this thesis present work in three areas of X-ray crystallography. First, the successful derivation of part of the structure of CRLF3 is described, with an initial interpretation of medically significant mutations. Second, generally applicable software was developed to assist in the identification of non-crystallographic symmetry using new algorithms. Third, the effect of removing different parts of data upon the ease of obtaining a molecular replacement solution was investigated to guide experimental data gathering priorities.

1.1.1 Introduction to macromolecular crystallography and associated techniques

This introductory chapter includes a guide to the structure of a protein. The use of X-ray diffraction to derive protein structures is then discussed, including the difficulties associated with this technique. The stages in deriving a structure from the information provided in diffraction patterns is then elucidated, from obtaining an initial estimate of the phases, through to improving this estimate, and finally refinement to obtain the best possible structural model. Various metrics to analyse the quality of fit to the data are presented. Finally other techniques which give similar, often complementary, structural information are introduced.

1.1.2 Cytokine receptor-like factor 3 (CRLF3) structure determination

Chapter 2 of this thesis presents work to determine the structure of the murine protein cytokine receptor-like factor 3 (CRLF3). Experimental results from the International Knockout Mouse Consortium (IKMC)[1] had suggested that this was an important protein in the blood clotting pathway.

A molecular structure was desired for this protein in order to facilitate a better understanding of its function, and to enable interpretation of the effects of possible mutations on the observed phenotypes. The original crystallisation of the full length protein had failed to form crystals, and so new constructs were designed that could be used to generate crystals. In order to achieve this, the sequence was analysed in various ways, from predicting the secondary structure, to exploring homologous structures. This allowed as much understanding as possible of the reasons for the failure of the first crystallisation attempts, along with a sensible rationale for the selection of better constructs for crystallisation.

However, even with the growth of acceptable crystals, obtaining the structure was not trivial. It involved multiple attempts using the following techniques: molecular replacement and SAD. Eventually, with a second data collection, the structure of the construct of part of murine CRLF3 could be solved.

From this structure some useful conclusions were drawn. Taking known mutations associated with blood clotting disease in this protein together with the close similarity between murine and human CRLF3 (93% identical), it could be determined whether the structure predicted for a particular mutation would have a significant effect on the pathology. If it did, then this potentially suggested that the selected mutation in this protein could be associated with the blood clotting disease.

Part of this work was done in collaboration with Dr. Cavan Bennett in Dr. Cedric Ghevaert’s group from Department of Haematology in the NHS Blood and Transplant Centre at the University of Cambridge, who undertook the production of the constructs designed as part of this work, and Dr. Yahui Yan from Professor Randy Read’s group in the Cambridge Institute for Medical Research, University of Cambridge, who led the crystallisation trials and remote data collection.

1.1.3 Elucidation of non-crystallographic symmetry

Chapter 3 of this thesis describes a project to improve *phenix.find_ncs_from_density*[2]. Areas for improvement were identified when using this program in the work on CRLF3, and novel algorithms were generated. This program takes in an electron density map and tries to identify which regions of the map are linked by non-crystallographic symmetry (NCS). *phenix.find_ncs_from_density*[2] does a phased translation search of every single possible rotation on a 20° net of a sphere of electron density with a 10 Å radius, in order to see whether it is similar to any other volume of electron density. It also limits the resolution to 4 Å to allow for identification of the NCS operators within a reasonable computational time.

This project resulted in the creation of a new program called *NaCelleS*, which was an attempt to develop a better way of finding NCS, improving upon the above method. *NaCelleS* uses all the data, but, to ensure a quick solution, a rotation search is done initially. This means that only a selected list of rotations is required with a phased translation search. Thus it was hoped that this would provide a more sensitive way of identifying NCS.

A further idea explored in this project was called the *NaCelleS* two-sphere approach. This method allowed the deduction of whether anomalous scatterer sites are linked to each other by NCS, by comparing the electron density only around these sites.

Further, the idea of using an iterative approach towards finding NCS was explored. In *phenix.find_ncs_from_density*[2], once a subset of the operators have been found there is no way of improving or extending the NCS operators using the NCS information now known. Therefore the question of whether finding a few operators would allow for more operators to be located was considered. This approach exhibited some success and some issues, as will be discussed later.

Overall, this project aimed to improve a program that is in current use, by trying to make it more intelligent, and to attempt to make better use of the data that are provided. The improvement in the method should provide some benefit to other scientists attempting to gain protein structures in their own experiments.

1.1.4 Effects of data pathology on ease of molecular replacement solution

In Chapter 4, work on determining the importance of low resolution and high intensity data in molecular replacement solutions is described. There has been a long held hypothesis that low resolution data are uniquely important to molecular replacement and that they must be recorded with great accuracy. This, however, seemed to be a misunderstanding of the work of Davies[3], who actually said that the highest intensity reflections were the most important, not the low resolution ones. However, frequently the highest intensity reflections on a diffraction pattern are in the low resolution region, leading to a potential misinterpretation of his work. Therefore the question has been asked as to whether it was the most intense reflections or the low resolution reflections that were the more important.

The question of whether there was a difference between maximum likelihood methods and Patterson-based methods regarding the unique importance of low resolution data was explored, and whether the two methods behaved in different ways. These questions were explored through using the implementations in *phaser*[4], using a test data set curated by Dr. Rob Oeffner in the Cambridge Institute for Medical Research, University of Cambridge.

1.2 Structure of a protein

Proteins are biological molecules that perform many of the complex functions in a living cell. Each protein is highly specialised for its function, for example to provide channels through membranes, break down substances or provide structural support to the cells. Comprehending the structure of such molecules helps us to understand how the basic functions of life are carried out, as it is this structural shape that largely determines the function of the protein.

One very relevant and beneficial application of this knowledge, obtained from these structural studies, is in medicine. Firstly, such studies can help us to understand disease mechanisms. For example, sickle cell anaemia occurs because of a change in a single residue (Section 1.2.1) in haemoglobin, which leads to a change in the structure of this protein that allows red blood cells to aggregate in the deoxygenated state. This structural change makes the red blood cells inflexible, leading to difficulty in entering small blood vessels, and reduces their average life time. Thus the ability of the haemoglobin to carry oxygen is diminished. However, such structural changes can also be beneficial, as it has been found that this mutant haemoglobin potentially allows greater resistance to malaria[5].

Another important application of these protein structures is to facilitate the design of drugs. Before protein structures became available, some drugs were developed by taking natural compounds, which were known to have beneficial properties, and chemists would then create similar compounds to try to improve this biological activity in order to make a more effective drug[6]. Mass screening of non-natural (man-made) compounds, chemically similar to compounds with known beneficial properties, was also explored[7,8] in order to find effective medicines. Now, with protein structures becoming available through the development of a number of experimental techniques, the opportunity to design drugs to fit specific active sites or to act as pharmacological chaperones[9] based on this knowledge is growing, offering new medicines to tackle disease. Determining protein structures and developing techniques to permit such determination is thus very important to continue this beneficial work.

1.2.1 Primary structure of the protein

The primary structure of the protein is determined by the order of the amino acids it contains. Historically this structure referred to all the covalent linkages, including the disulphide bridges. More recently, only the sequence of amino acids is considered, while the other covalent bonds are considered as part of the tertiary or the quaternary structure[10]. This primary structure is a linear sequence of amino acids, written sequentially, beginning at the amino terminus (N-terminus) to the carboxyl terminus (C-terminus) (Figure 1-1), using a single letter code formalised in the 1960s[11]. This sequence of amino acids determines the structure of the protein[12] and, as discussed later, the structure of the protein can theoretically be predicted. This concept was suggested originally from the results of the Anfinsen experiment in which ribonuclease was unfolded and refolded, and found to have the same activity[13].

Amino acids differ in their side chains. These side chains can be hydrophilic (interact well with water), or hydrophobic (interact poorly with water), leading to the primary structure taking up a shape in the presence of water to reflect these interactions, e.g. alpha-helix or beta pleated sheet structures. There are twenty possible different amino acids (residues) in humans; typical examples are alanine, serine and aspartic acid.

1.2.2 Secondary structure of the protein

The secondary structure of the protein is the three-dimensional folded form of local segments of the protein backbone. This is determined largely by the sequence of residues, influenced by long-range interaction between residues that are in close proximity. To understand how a protein folds, there are three main-chain dihedral angles to consider in every residue of the protein (Figure 1-1). An understanding of restraints for these angles, and the most probable values, is very helpful in determining protein structures.

Firstly, we need to consider the ω angle. Investigations have determined that the central bond determining this angle has a partial double bond character, so it predominantly is most stable in a flat cis or trans conformation[14] (implying that ω is approximately 0 ° or 180 ° [15]). Furthermore, the trans conformation has been found to be far more common than the cis, although the cis conformation is still found. In non-proline residues, presence of a cis peptide can indicate that this residue is functionally active[16,17]. However there have been a number of cases where ω deviates from planarity[18], so this option must be borne in mind when reviewing the ω angle in structure determination.

Considering the other two angles, φ and ψ, there are fewer restraints. These angles can adopt a greater number of values, as there is no double bond characteristic as can be seen for the ω angle. However, there are still only certain angles that are usually available. These are best summed up in a Ramachandran plot[19] (Figure 1-2), which shows the possible φ and ψ angles. The plot was initially based on energetics and conformations that are sterically possible. It is based on an exploration of all the different combinations, showing the angles that occur in other proteins. There are many unfavourable angles (these angles lead to steric clash of the backbone with itself), which are rarely formed. The angles derived from the model are plotted on this chart and can be compared with the integrated knowledge from the protein data base, shown as coloured areas in Figure 1-2.

Finally, the side chain groups (R groups) on the protein chain also take certain conformations[20,21]. For glycine there is only a single option, as it is just a hydrogen, but for longer R groups, there can be a number of different conformations (for example, Figure 1-3).

The different side chains vary widely in chemical properties e.g. polar, hydrophilicity, charge and their physical configuration e.g. size, shape, β-branched. Cysteine side chains can also come together to form disulphide bonds.

The parameters described above are sufficient to determine the different areas of flexibility in each residue of the protein. These considerations result in a number of possible secondary structures, which are now discussed.

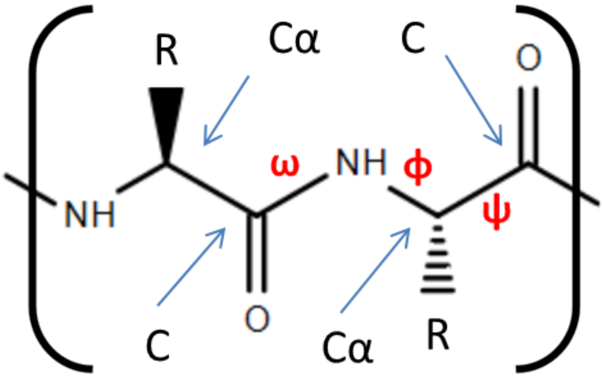


Figure 1-1 Diagram of angles in a protein backbone

This shows two repeat groups in a protein, with R being any of the side groups found in amino acids. The different angles that define the backbone conformation have been defined from the central bond around which the other rotations occur. φ is the right-handed angle between the Cα-C bond and C-N bond, ψ between C-N and N-Cα, and ω between Cα-C and N-Cα. (Picture drawn using *ChemSpider*[22] interface and *Microsoft Publisher 2017*).

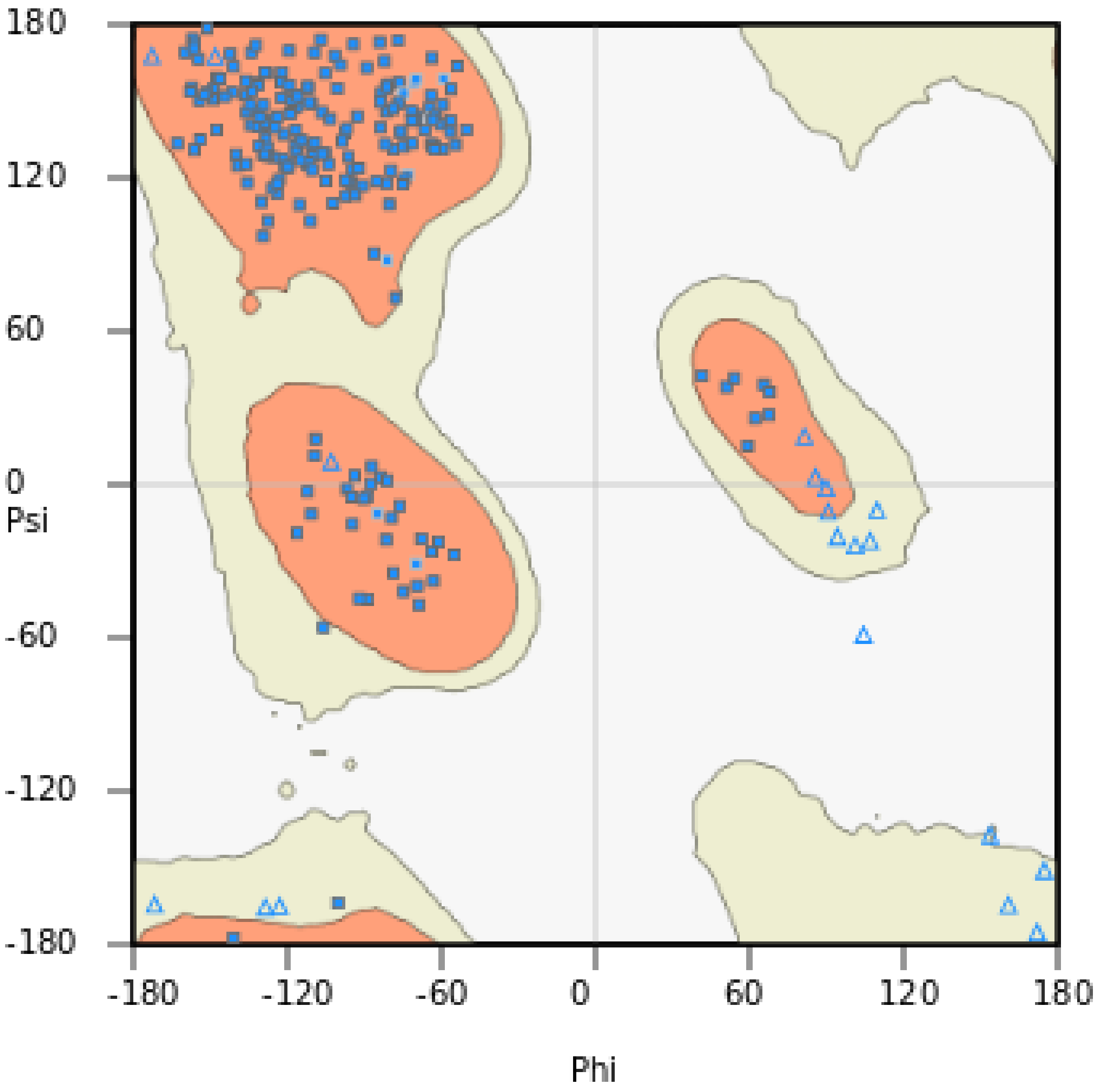


Figure 1-2 A Ramachandran plot for the CRLF3 protein model 2

This figure shows the φ and ψ angles (see Figure 1-1) for all modelled residues in the protein as dark blue triangles (showing glycine residues) or squares (all other residues). The orange regions show the favourable region for a glutamine residue, and yellow the allowed region. As can be seen, many areas are unfavourable for glutamine. The α-helices region (at φ = -60 ° and ψ = -40 °) and β-sheet region (at φ = -120 ° and ψ = 120 °) can also clearly be seen (plot generated using *Coot*[23]).

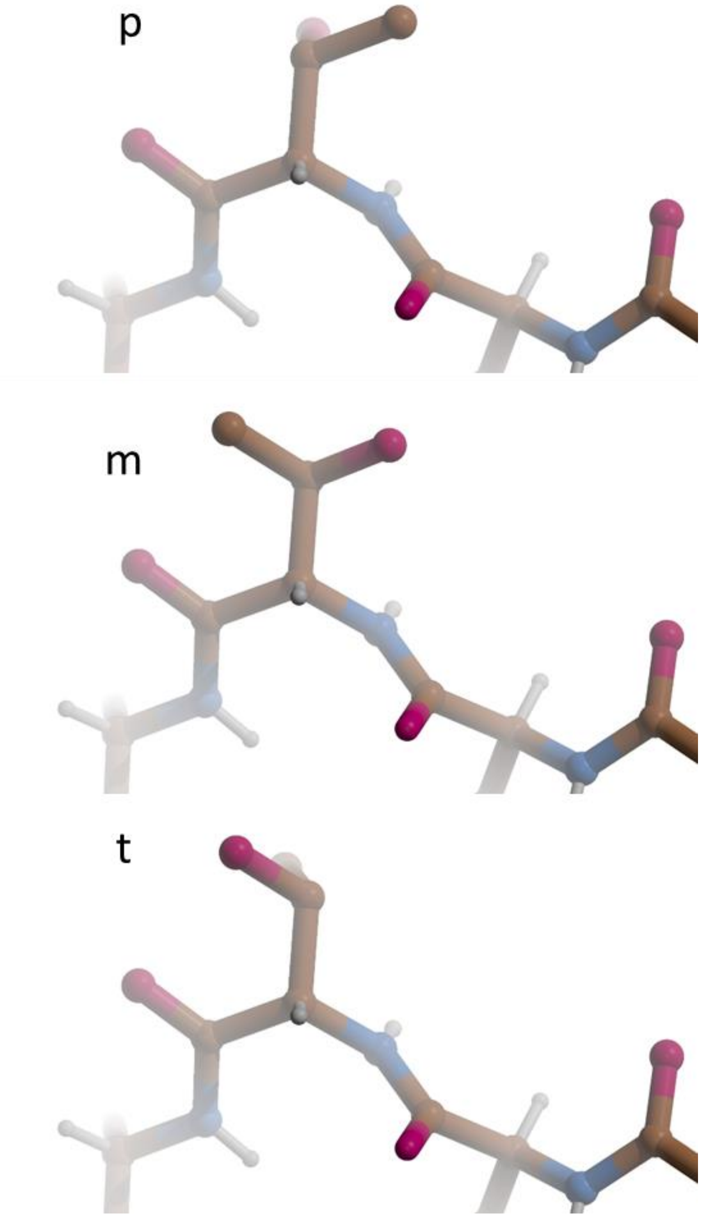


Figure 1-3: Three common rotamers of threonine, as an example of these orientations

For the residue threonine, there are three common rotamers. The para (p), meta (m) and the trans (t), also called gauche-, gauche+ and trans, are found in 49 %, 43 % and 7 % of structures, respectively. This is example of one residue's common rotamers – other residues have common rotamers as well, and can be described in similar ways (plot generated using *Coot*[23] and *RASTER3D*[24]).

1.2.3 Types of secondary structure

1.2.3.1 α-helices

The α-helix is named from Astbury's statements in the 1930s about keratin, which, whilst Astbury had the diffraction pattern for the α-helix, did not lead to a derivation of the structure[25]. The structure was finally postulated by Pauling[14,26], although it was later discovered that he had predicted the wrong enantiomer[27]. It was further demonstrated by the high resolution crystal structure of myoglobin[28] showing accurately the true α-helical structure.

For the α-helix structure, one complete turn of a helix requires 3.6 residues, so each residue contributes a 100° rotation. In the direction of the helix, a residue adds a 1.5 Å increase in length. The side chains all point outwards and downwards towards the C-terminus and not towards the centre of the helix. This occurs because the carboxyl oxygen (residue *i*) on the backbone hydrogen bonds with the N-H on the amine part of a residue further along the helix (residue *i*+4). Therefore all of the main-chain hydrogen bonding is within the helix, and all of the side chains are outside the central core of the helix. As all the N-H residues point towards the N-terminus and all the C=O amides point towards the C-terminus, a dipole exists over the entire helix, being positive at the N-terminus and negative at the C-terminus. This affects the structures of proteins[29]. The usual backbone angles in α-helices are φ = -60 ° and ψ = -40 °.

1.2.3.2 β-sheets

The β-sheet, first postulated by Pauling in 1951[26,30], was first experimentally shown to exist with the structure of egg white lysozyme in 1965[31]. The first structure proposed was broadly correct, but the actual β-sheets were later found to be twisted and to have a significant amount of shear. This has been found consistently in most β-sheets[32,33].

The β-sheet is a long sheet-like structure, with the outward facing sidechains alternating from side to side of the sheet. The dihedral angles φ and ψ have a larger distribution in β-sheets than the α-helices, but this distribution is clustered approximately around φ = -120 ° and ψ = 120 °.

One of the largest differences between α-helices and β-sheets is in the backbone hydrogen bonding. In the latter, the carboxyl C=O and the amide N-H have none of their possible hydrogen bonding interactions satisfied within a single strand (as opposed to α-helices where all the backbone hydrogen bonding is satisfied). Instead the β-sheets need to form interactions with other strands, as will be discussed shortly.

1.2.3.3 Other secondary structures

Other secondary structures can exist and have been described in the literature. An example is the β-turn, also known as the Venkatachalam-turn after the person who first postulated it[34]. He calculated the φ and ψ required to allow a turn where there is a hydrogen bond between residues *i* and *i + 3*, leading to three different types of turn, an example of which is the 3_{10} helix discussed below. There are many more types that have been first calculated[35] and then discovered experimentally[36]; these have been reviewed comprehensively[37].

The 3_{10} helix is a more tightly wound, thinner and more stretched out helix than the α-helix. Most 3_{10} helices tend to be short, with an average, not of 3 residues per turn as predicted by Pauling[14], but of 3.2–3.3 residues per turn[38], with a rotation of approximately 110 ° per residue. Each residue adds 1.93–2.00 Å to the helix length, and an helical pitch of 5.8–6.0 Å[39]. More recent studies have shown longer 3_{10} helices[40,41], including some acting as a transmembrane region important for voltage sensing[39].

There are also poly-proline helices. The two types predominately found are poly-proline type I helix (PPI) with φ = − 75 °, ψ = 160 ° and ω = 0 ° and poly-proline type II helix (PPII) with φ = −75 °, ψ = 145 ° and ω = 180 °[42,43]. PPI helices are rare in nature and have not yet been assigned to any biological function. However, PPII helices exist in a wide range of proteins[44–46], and have been found to be important in a wide range of functions from immune responses to cell motility[47,48].

1.2.3.4 Definition and prediction of secondary structures

The presence of certain residues, or combinations of residues, provides information on likely secondary structures. Particular residues have a greater likelihood of being found in an α-helix[49–52]. Some important examples are considered below.

Alanine is usually seen as the optimum residue, as it packs well between the residues above and below. Residues that are β-branched (e.g. valine) tend to be unfavourable due to steric clashes. Similarly, large bulky residues, such as tryptophan, can be unfavourable in the middle of a helix due to steric clashing. Residues such as lysine and arginine, which have a long linear hydrophobic region, tend to be energetically favourable for forming α-helices.

Proline is known as the "helix-breaker" and, because of steric clashes within the side chain when trying to form the φ and ψ required for the helix, it tends not to be found in helices. It also lacks an N-H, so the hydrogen bonding found inside the α-helix is not possible, and there is a carbon atom in that position which sterically clashes.

Glycine, with its very large range of conformational freedom in the denatured state, tends not to be found in helices since the large entropy loss in incorporating it, with no gain in favourable interactions, tends to be energetically unfavourable. Some literature has suggested that glycines and prolines that are found in helices can make the helices more flexible[53].

In contrast, at the terminal ends of the helices different residues tend to be found[54–56]. Residues are seen which can stabilise the unsatisfied H bonding of the backbone and the dipole e.g. for the N-terminus glutamic acid and aspartic acid, and for the C-terminus arginine and lysine. Furthermore, glycine is commonly found, as not only does it allow for a large range of φ and ψ angles and so allow the chain to twist into a different conformation, but it also can allow enough space for water to enter and hydrogen bond with the backbone.

1.2.3.5 Definition and prediction of secondary structures

Secondary structure, as has been seen, concerns folding on the local scale within the protein. The question arises whether such folding can be predicted based on the local sequence. There has been extensive work on this throughout the years[57,58]. The first programs were based on statistical methods, outputting the residues which were more likely to be involved in each secondary structure element[59]. More recent programs have used machine learning to provide more accurate predictions, and these will be discussed later[60].

Once the complete three-dimensional atomic protein structure is determined, there are a number of methods that can define the secondary structure that is present. The basis of these programs is to scan through the protein residues and match against the pattern of hydrogen bonding seen in other structures. There are two common programs that do this, called STRIDE[61,62] and DSSP[63,64].

1.2.4 Supersecondary structure

The supersecondary structure describes the local folding of the secondary elements to form small compact sections of the overall structure of the protein. It indicates the grouping of a small number of secondary structures in the final model, and common motifs that are seen. Examples include β-hairpins, βαβ motifs, β-barrels, Greek keys among many others.

1.2.5 Tertiary structure

The tertiary structure of a polypeptide chain within a protein is the three-dimensional structure of the molecular chain, i.e. the way the entire polypeptide protein folds in space, allowing for the hydrophobic residues to cluster into the centre, whilst the hydrophilic residues are on the outside[65]. The tertiary structure describes how all the secondary structure elements come together to form the three-dimensional structure of a polypeptide chain within a protein.

1.2.5.1 Tertiary structure motifs

There are certain motifs which are common, especially when associated with particular protein functions. These motifs fall broadly into four categories, namely antiparallel α-helices, parallel α-helices and β-sheets, antiparallel sheets, and a category containing small SS-rich or metal-rich motifs[37].

1.2.6 Quaternary structure

The quaternary structure of the protein is the way the tertiary structures (there can be several different polypeptide chains) come together to form the complete protein. For example, haemoglobin (Figure 1-4) is a combination of four polypeptide chains around iron ions. The combination of these allow it to carry out its function of carrying oxygen[5]. There can also be very large complexes, like the ferritin which forms a 480 kDa 24-mer[66].

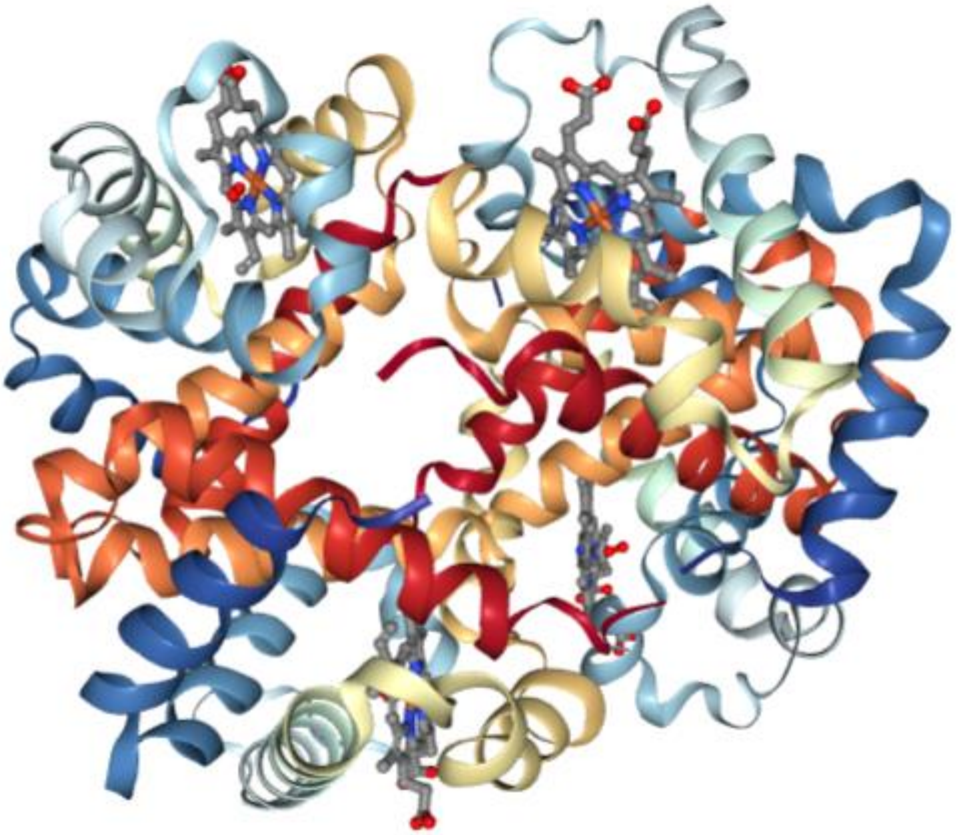


Figure 1-4: Image of haemoglobin (PDB code 1GZX)

This shows the quaternary structure well, with the four chains (two α-chains and two β-chains) and the iron ions in the haem rings. A rainbow colour scheme for each of the chains, ranging from blue at the N-terminus, to red at the C-terminus has been used. (Image created using the *NGL viewer*[67]).

1.3 Protein Data Bank

In order to make structures of proteins freely available for scientists to use and to publicise what proteins have had their structures determined, the Protein Data Bank (PDB) was set up in 1971[68,69]. The deposition of protein structures in this database was initially voluntary, but in the mid-1980s strong arguments for mandatory deposition were made, and in 1989 the International Union of Crystallography (IUCr) created a set of guidelines[70] for this purpose. Given the importance of protein structures to other scientists and to medicine, it was argued that holding them in a central place, freely available, was of extreme importance. These guidelines allowed for coordinates of a protein structure to be held confidentially for a period of time, but they would eventually make all the protein structures freely accessible.

Therefore there is now a central database where coordinates of atomic models can be deposited. For crystal structures, deposition of the experimental data in the PDB, whilst originally encouraged, is now a requirement. Other requirements include details about the experiment, composition of the structure and details of data handling and refinement, along with contact details of the relevant scientists.

For electron microscopy, there is the Electron Microscopy Public Image Archive (EMPIAR)[71], which allows researchers to deposit the raw data. In neutron diffraction and nuclear magnetic resonance (NMR)[72], there are separate places to deposit the data, but structures are still in the PDB.

1.4 Calculating the protein structure through X-ray diffraction

One of the most successful and widely used methods for the solving of protein structures is protein X-ray crystallography. On June 22nd 2018, 141,415 structures have been made publicly available and solved by X-ray crystallography. Furthermore, there are significantly more structures which are solved within pharmaceutical companies and within labs that have not yet been released. The following sections describe this method, which was used in the research described in this thesis.

In order to take a picture of a protein and be able to resolve the atoms in the structure, a wavelength which is of the same order as, or smaller than, the distance between the atoms is required. Visible light would be unsuitable, as it has a wavelength of 10^{-7} m whereas inter-atomic distances are of order 10^{-10} m. Therefore X-rays are used instead, as they have the appropriate wavelength[73,74].

1.4.1 Difficulties with X-rays

There are, however, several distinct problems with using X-rays in this application, including practical resolution limits, signal-to-noise issues and potential radiation damage to the proteins during measurement. These will now be considered in turn.

Firstly proteins interact very weakly with X-rays[75]. This interaction is with the electrons in the molecule, and it reflects electron densities in space. Electromagnetic radiation interacts with matter through its fluctuating electric field, which accelerates charged particles. The interaction is essentially with electrons, which are grouped around the nucleus, as the intensity of scattered radiation is proportional to the charge/mass ratio squared. This can be considered as the electrons emitting electromagnetic radiation after excitation. This weak interaction problem is reduced by crystallising a protein[76–79] so that there are many repeats of the protein in the path of the X-ray beam, interacting and enhancing the signal (Figure 1-5). Crystallisation also fixes the protein in a certain conformation, ensuring that it has minimal mobility[80]. This allows the diffracted beams to add up in phase and so results in a much stronger signal, reducing this initial problem to an acceptable level so that structures can be derived.

The second problem is that X-rays are energetic and can damage the protein. This implies that the derived protein model may be of a damaged protein and not of the protein as found in nature. There are many commonly observed examples of radiation damage to proteins, from disulphide bridges breaking to the decarboxylation of aspartates and glutamates[81,82].

Primary damage comes through mainly the photoelectric effect, along with a smaller component from direct inelastic scattering of photons from the protein, leading to energy being absorbed and electrons being excited. Secondary damage then occurs through radiolytic damage after the primary process. This causes chemical bonds to break, redox processes to occur, and free radicals to be generated[83].

This damage to the crystal can be a significant problem, as it degrades the high resolution data that can be obtained. Moreover, X-ray radiation can lead to different damage throughout the crystal, meaning that different protein conformations are averaged together producing a crystal with data diffracting to a lower resolution only.

In order to reduce damage to the crystal, the amount of time that the crystal is exposed to the beam is minimised[84]. Furthermore, if the crystal is large enough, a different part of the crystal can be used to complete the data set, thus spreading the absorption of energy and damage. Alternatively, if more than one crystal is present datasets from multiple crystals could be combined[84,85].

It has also been discovered that using gaseous nitrogen to cool the crystals to 100 K has significantly reduced radiation damage[86], although it remains a problem. Attempts to cool the crystal further using helium to 15 K did not yield significant reduction in radiation damage compared to 100 K[87,88]. Issues also arise in low temperature macromolecular crystallography. Ice can form, which will greatly increase the background in the diffraction patterns measured. Furthermore, ice can increase the mosaicity of the crystal, as the water now turned to ice can disturb the crystalline order. Therefore cryoprotectant is used to avoid as many of these issues as possible. The choice of cryoprotectant for the crystals has been somewhat difficult to rationalise but is becoming better understood[89].

There is another technique that can eliminate the problem of radiation damage in a new and unique way. An X-ray Free Electron Laser (XFEL)[90–95] shoots high intensity X-rays at a steady stream of crystals. This method obtains a diffraction pattern in such a short time frame (e.g. 70 fs) that the crystal does not have a chance to break down. The beam is also pulsatile. By the time the atoms start moving away from each other, the protein is no longer in the X-ray beam. Thus it was originally hoped that radiation damage is not a problem, though there is now growing understanding that this might not be true. Studies indicate that the damage inducing process is complex and depends both on the intensity and duration of the X-ray pulse. Simulation experiments suggested that differences in specific bond lengths measured by XFEL and traditional X-ray crystallography were not caused directly by radiation damage, and that pulses shorter than 15 fs should be used to avoid significant radiation damage[96,97].

Thirdly, direct images of molecules cannot be obtained with X-rays. No lens is currently available for X-rays which can provide a high enough resolution to see the atomic details in proteins. A resolution of ~ 2 Å is required, whereas the cutting edge for X-ray lenses is ~ 4 nm[98]. Therefore once the X-rays have passed through the sample, only the diffraction pattern can be determined, rather than a picture of the protein itself. Computer post-processing of the diffraction patterns is then used to reassemble the image of the molecule, as phase information cannot be collected experimentally.

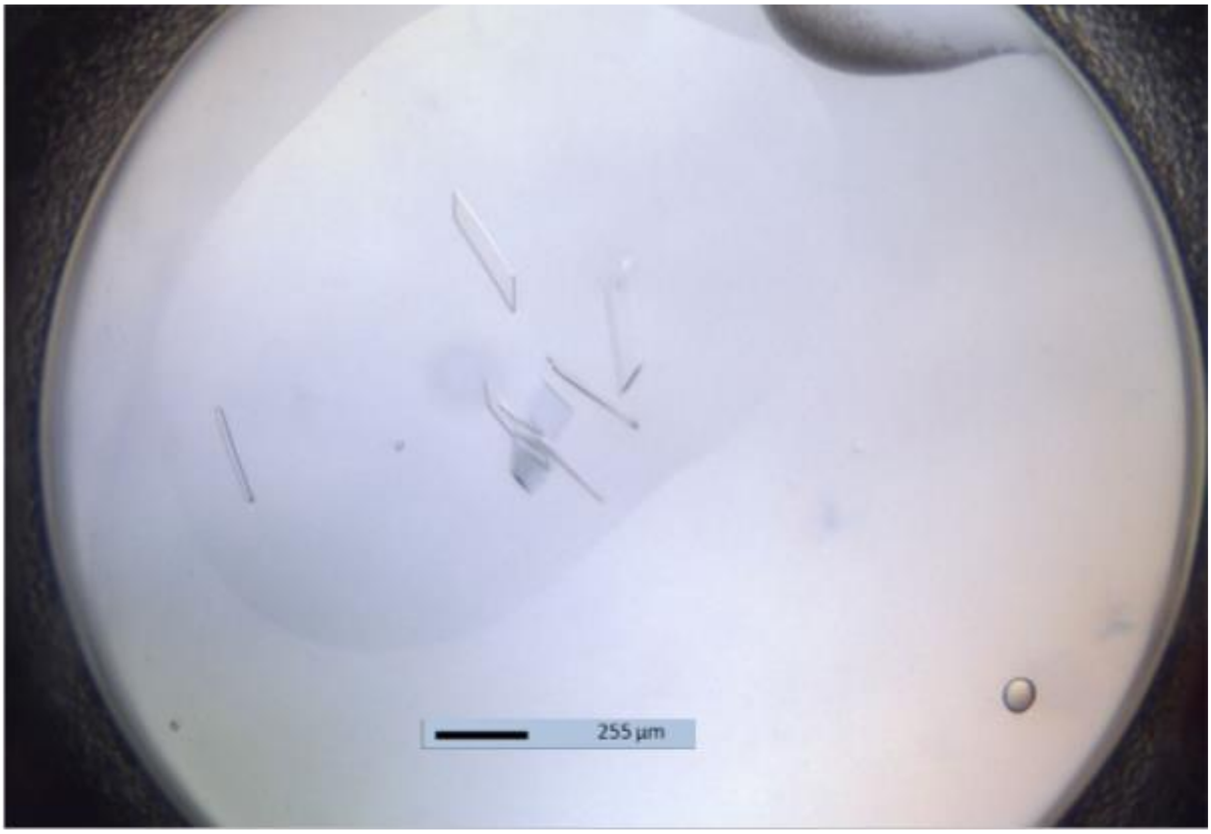


Figure 1-5: Example crystals from the CRLF3 protein that can be used in an X-ray diffraction experiment

This image shows the CRLF3 crystals before they are placed into cryoprotectant and liquid nitrogen. This was a droplet on a 96 well plate, in a 400 nl volume with a 1:1 mixture of CRLF3 (in 10 mM Tris pH 7.4, 150 mM NaCl with a solution of 12 % PEG 3,350, 0.1 M Sodium Acetate) under vapour diffusion with a 70 µl solution in the same sealed environment (Image taken using Rock Imager by Formulatrix).

1.4.2 Diffraction from a crystal

The diffraction pattern obtained from a crystal is a regular series of reflections, the positions of which are determined by the repeating lattice of a crystal.

The regular pattern of molecules in a crystal, with each atom elastically scattering the incoming X-rays, permits the creation of a diffraction pattern. Each electron can be considered as an emitter of electromagnetic waves at the frequency of the incoming X-rays, and these waves will add up and interfere with each other. Whether this interference is constructive or destructive will depend on the direction of the incoming and outgoing waves, and the relative position of the electron sources. It is easiest to determine the conditions for scattering in phase if the waves are considered as being reflected by planes passing through the atoms in the crystal. Bragg's Law then describes the relationship between the scattering angle and the interplanar spacing of the atoms.

$$2d\,sin\,\theta = n\lambda$$

Equation 1-1

where d is the interplanar distance, θ is the incident (half the scattering) angle relative to the direct beam as the reference zero angle, n is a positive integer and λ is the wavelength of the incident wave.

A number of deductions can be made from this law. Firstly, the longer the incoming wavelength, the lower the sensitivity to the spacing of the planes or changes in angle. Secondly, as the spacing between the atomic layers is decreased, the angle for the first peak in diffracted intensity becomes larger. This reciprocal relationship between interplanar spacing and the angle of diffraction leads to the naming of the diffraction space as 'reciprocal space', a term which is routinely used in the references and in this thesis.

The structure factor represents the wave that results from the diffraction. For a single atom, it can be expressed as follows

$$\boldsymbol{F}(\boldsymbol{S}) = f_n exp(2\pi i\boldsymbol{r}.\boldsymbol{S})$$

Equation 1-2

where f_n is the atomic scattering factor, which depends on scattering angle and type of atom (i.e. number of electrons around the atom) if we assume that the atom is spherically symmetrical, $\boldsymbol{r}$ is a real space vector denoting the position of the atom, and $F(\boldsymbol{S})$ is the structure factor. In this equation, the diffraction vector, $\boldsymbol{S}$, is given by

$$\boldsymbol{S} = (1/\lambda)(\boldsymbol{s} - \boldsymbol{s_0})$$

Equation 1-3

where $\boldsymbol{s}$ is the vector of length $(1/\lambda)$ parallel to the incident beam and $\boldsymbol{s_0}$ is the vector of length $(1/\lambda)$ parallel to the diffracted beams (Figure 1-6).

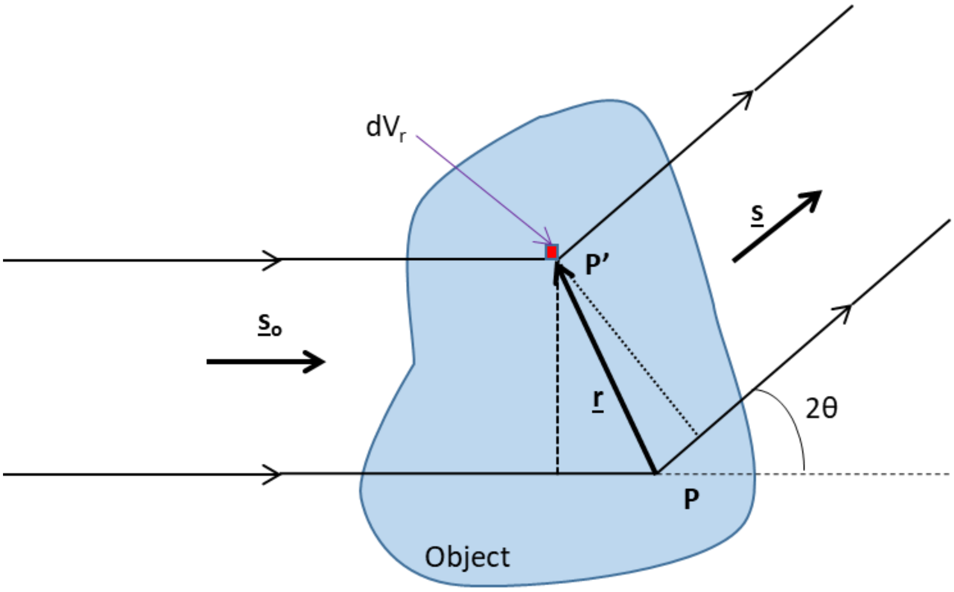


Figure 1-6: Illustration of path length difference between different paths

where P is a point in the object, and P’ is another random point in the same object. Other variables are as defined for Equation *1-2* and Equation *1-4*.

The scattered vector from a single electron in an atom can then be taken and summed up appropriately for all electrons diffracting in the space that the beam is passing through, as the resultant pattern from each will be a contribution. As this summation is over many electrons that exist in orbitals of many different shapes, this can then be written as

$$F(S) = \int \rho(\boldsymbol{r}) \exp(2\pi i \boldsymbol{r}.\boldsymbol{S})\,.\,dV_{\boldsymbol{r}}$$

Equation 1-4

where $\rho(\boldsymbol{r})$ is the electron density function of a unit cell, and $dV_{\boldsymbol{r}}$ is the small volume element over which the integral over all space is performed. The above equation is in the form of a Fourier Transform.

The condition that this is to be considered over a crystal, with a repeating unit cell, can now be added. This condition implies that defined reflections are only seen where there is constructive interference from all unit cells (a very large number). The beams of diffracted X-rays generating the resulting diffraction pattern are thus limited to discrete directions, resulting in the spot nature of the pattern. These beams are also dependent only on the distribution of the electron density within the unit cell, and not on the crystal, and so the following constants can now be refined and defined more precisely: $\boldsymbol{r}$ is a real space vector, with fractional coordinates in each of the crystal repeat lattice vectors ($\boldsymbol{a}$**,** $\boldsymbol{b}$ and $\boldsymbol{c}$), which in a unit cell is

$$\boldsymbol{r} = x_n\boldsymbol{a} + y_n\boldsymbol{b} + z_n\boldsymbol{c}$$

Equation 1-5

where (x_n, y_n, z_n) are the coordinates of an atom in the crystal and $\boldsymbol{S}$ is the diffraction vector, defined by the reciprocal lattice cell lengths ($\boldsymbol{a}^*$, $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$) and can be given by the following equation

$$\boldsymbol{S} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$$

Equation 1-6

where $h, k, l,$ are the Miller Indices labelling the allowed values of $\boldsymbol{S}$ in the crystal. This allows for the simple calculation

$$\boldsymbol{r}.\boldsymbol{S} = hx_n + ky_n + lz_n$$

Equation 1-7

The Fourier transform can then be written

$$F(hkl) = \iiint \rho(xyz)\exp\big(2\pi i(hx_n + ky_n + \ lz_n)\big).dxdydz$$

Equation 1-8

At this point it is useful to introduce the independent atom model as a simplifying assumption. The electron density is assumed to contain only individual atoms at fixed sites, ignoring anything that is not in a fixed (repeating) position inside a cell. Anything that is not fixed, when summed up over all unit cells, will average out and yield no contribution to a beam other than the F(0,0,0) structure factor. A second assumption that all electrons are located on an atom is also applied, neglecting the presence of any electrons in chemical bonds (or in particular any delocalised π systems). This assumption is chemically inaccurate, but permits the solving of structures.

This now allows the electron density for a unit cell as a sum of independent atoms to be written in the following form

$$p(xyz) = \sum_n \rho_n(xyz) * \delta(\boldsymbol{r} - x_n\boldsymbol{a} - y_n\boldsymbol{b} - \ z_n\boldsymbol{c})$$

Equation 1-9

where $p(xyz)$ is the total electron density of the molecule, $\rho_n(xyz)$ is the electron density for a single atom, $\boldsymbol{r}$ is the position in space in question, $*$ is the convolution operator, $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are the unit cell lattice vectors, and x_n, y_n and z_n are the fractional coordinates of each atom.

Taking the Fourier transform of this equation defines the structure factors as follows

$$F(hkl) = \sum_n f_n exp(2\pi i(hx_n + ky_n + \ lz_n))$$

Equation 1-10

This equation permits the structure factors from a known model to be determined. These structure factors correspond to the reflections seen in a diffraction pattern of the model.

Starting from a measured diffraction pattern, once the particular (hkl) beam is identified (indexing), the inverse Fourier transform can be taken from the sum of all the beams. If the phases were known, the electron density could be obtained. However, as explained previously, the phases cannot be determined experimentally from the diffraction pattern.

1.4.3 Experimental difficulties in measuring a diffraction pattern

There are a series of problems that might occur in data collection, which impede the acquisition of an ideal data set. Firstly, the vast majority of the incident X-rays pass through the crystal without diffracting at all. If recorded on detectors, in particular old photographic film, the resulting high intensity peak would overwhelm the signal from all the other spots and could damage the detector/film. Therefore a beam-stop is included in experiments, which prevents the vast majority of undiffracted X-rays from being recorded.

This low level of interaction makes it impossible to measure the *F(0,0,0)* spot in the diffraction pattern, as it is recorded in the same place as the beam that does not diffract. Furthermore, depending on the size of the beam-stop, various other very low resolution data can be lost. A larger beam-stop makes it easier to align the beam in the X-ray diffraction experiment, but leads to more data being lost.

Another potential problem can be the inaccurate recording of intensities of the most intense spots. Some X-ray detectors cannot record very intense spots, as the count rate becomes too high and the detector is overloaded. This leads to all the various different greatest intensity reflections having the same incorrect intensity recorded, which is determined by the maximum possible reading of the detector. In some detectors, very high signals can leak across individual pixels, further changing the spot intensity pattern.

These points are considered in more detail in Chapter 4.

1.5 Indexing

Once the images have been recorded, the different spots on the diffraction pattern need to be indexed and the space group determined as the first step towards the solution of the protein structure. These images can be complex and determination of the correct indices is not trivial. Attempts to average diffraction spots from different images before knowing their diffraction indices (indexing) have been largely unsuccessful[99]. Instead an indexing is calculated, usually for two images with the crystal rotated 90 ° between the two images, and applied to all other images before the intensities are measured from all the images[100]. Originally autoindexing, which is the automatic indexing of the images using computer algorithms, required knowledge of unit-cell parameters, but methods where this was not necessary have superseded them[101]. Indexing methods rely on diffraction images being recorded using the oscillation method, allowing for spots on the diffraction pattern to be mapped to reciprocal space vectors.

The oscillation method uses a smooth rotation, typically of around half a degree, in one axis whilst the image is being recorded. This can be sequential and without interruption between images, leading to the alternative name of rotation method. It allows collection of spot profile densities to permit accurate determination of the intensity of each spot, and helps with indexing.

There are different approaches to solving this indexing problem, as implemented in *XDS*[102] and *iMOSFLM*[103]. Once the diffraction pattern of the crystal has been indexed and measured, the next step towards the solution of the structure can be undertaken.

1.6 The Phase Problem

If the amplitudes and phases of these spots were known, then a Fourier transform would allow for the calculation of the electron density map, and a protein model could be built of the map. However, the intensity of the spot, which is the only measurement that can be recorded, is determined by the square of the Fourier transform of the electron density in a unit cell. Thus only intensities can be determined from the measurements, and the phases, which give most information of the structure of the protein, are not known (Figure 1-7). This is known as “The Phase Problem”. In order to progress beyond this problem, an initial estimate of the phases is needed.

There are a variety of different ways that this initial estimate of the phases can be achieved. These broadly fall into two categories. There are the experimental phasing methods, which aim to infer the phases from differences in the diffraction pattern, either by changing the crystal or by changing the wavelength of the beam. Another category is molecular replacement, which aims to use solved protein structures which are similar to the protein of interest as a guide for the initial estimates of phase. For very high resolution data there is also the option of direct methods, which uses the idea that any combination of phases which leads to negative electron density must be incorrect, and the idea of atomicity in the model, which can be seen from the electron density map. This can allow for a solution *ab initio* from the native dataset without requiring other information[104].

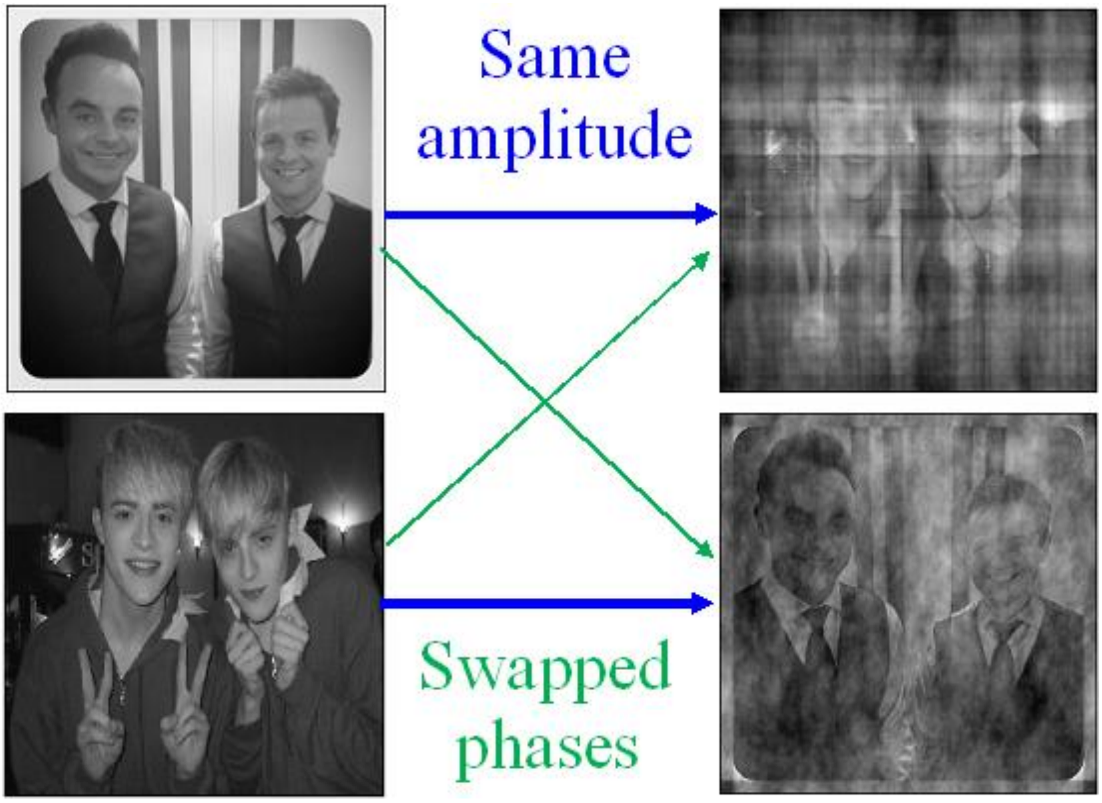


Figure 1-7: A demonstration of the importance of phase information

In the diagram above, diffraction patterns were calculated from the pictures on the left. The phase information was then exchanged between the two diffraction patterns and the patterns re-transformed to create the images on the right. As can be seen, the phases seem to carry most of the information. (Pictures created using *FTL-SE* program[105]).

1.6.1 Experimental Phasing Methods

One of the oldest experimental phasing methods is multiple isomorphous replacement (MIR)[106]. Here the addition of heavy atoms into the crystal leads to a change in the diffraction pattern. The introduction of such an atom will change the scattered intensity significantly, as heavy atoms contribute disproportionately to the overall intensity, and these atoms will scatter the incoming X-rays approximately in phase with each other. The heavy atoms need to bind in the crystal in the same places in each unit cell. If there were no errors in measurement and a perfect crystal were to be available, this approach would reduce the phase choice down to two possibilities for each structure factor (*F(hkl)*), and a second heavy atom substitution in a different crystal of the protein would resolve the ambiguity completely. The very first example of using this idea was in small molecular crystallography, with the structure determination of $CuSO_4$ and $CuSeO_4$[107] and later in macromolecular crystallography with the structure of myoglobin being solved[108].

The challenge with MIR is that native crystals and at least two other crystals, each having heavy atoms added to them, are required. The additional crystals can be made by soaking native crystals in solutions containing heavy atoms, thus introducing these atoms in different places in the crystals. If in the two substitutions the heavy atoms occupy the same location in the crystal, then MIR does not lead to resolution of the phase ambiguity. It is assumed that the crystal does not change with the addition of these heavy atoms. However, the act of adding the heavy atoms may have caused other dimensional changes in the unit cell, which make it no longer isomorphous to the original native crystal. Additionally, the act of heavy atom binding may change the structure of the protein, affecting the diffraction pattern unpredictably, and so not resolving the original phase ambiguity.

Single isomorphous replacement (SIR)[109] is similar to MIR, but restricts the experiment to a single substitution, leading to a choice of two phase angles. This is actually a simplification as it does not take into account errors with the measurements of the structure factors or the errors in the heavy atom position and its occupancy. These errors result in a lack of closure error, from which phase probabilities can be calculated. This phase probability forms a bimodal distribution, which can be described by a set of four Hendrickson–Lattman coefficients[110]. The integral of the structure factor over all phases then gives the best possible structure factor, from which an initial electron density map and an initial model can be calculated. SIR has similar experimental problems as MIR but for one heavy-atom derivative rather than two, usually making it experimentally simpler, but less phase information is obtained, leading to poorer electron density maps and requiring improvement in the initial phase estimates (Section 1.7).

There are other techniques in experimental phasing that do not require more than one crystal type. Multiple-wavelength anomalous dispersion (MAD)[111–114] uses the fact that certain elements have absorption edges in the X-ray band. Around these absorption edges there is an anomalous component to the scattering factor, which is shifted π/2 radians away from the original phase of the scattering reflection from these atoms. This leads to the breakdown of Friedel's Law, which states that ***F****(hkl)* equals ***F****(-h-k-l)* for the data collected. Thus if a number of datasets around these absorption edges are recorded at different X-ray wavelengths, then the phases of the different spots can be deduced from the change in intensities, as the anomalous signal will change with wavelength. However, there are drawbacks to this method, as indicated below.

One of the key requirements for MAD is a crystal containing elements that can scatter anomalously with a different strong signal at two wavelengths in the useful range. Suitable anomalous scattering elements can be found natively, such as potassium, calcium and magnesium. Heavy atoms can also be added to allow MAD to occur[115]. One of the most common ways of performing MAD experiments is to replace the S in the protein with Se by growing it in media that has methionines containing Se instead of S[116]. This can be expensive to do, although it has been successful in a very large number of cases.

Another problem is that multiple datasets from one crystal usually need to be taken, and radiation damage can become a problem, leading to a degradation in the data obtained. Alternatively, several crystals prepared under the same conditions could be used, but again there is the issue of whether the crystals are truly isomorphous.

In contrast, single-wavelength anomalous dispersion (SAD) [117], whilst requiring the same anomalous scatterers as MAD, only requires one dataset to be taken. Accordingly, in addition to the elements suitable for MAD, it can usefully exploit elements for which only one dataset (containing anomalous signal) can be recorded, such as chlorine, sulphur or phosphorus. Sulphur is commonly present in the native proteins but it only scatters very weakly. The one dataset can allow for the determination of a single correct phase solution. This is a simplification, as after taking into account all the errors, the output is a Gaussian distribution through the different phases. The best structure factor can then be calculated by integrating the structure factors over all phases, and an initial estimate of the target model is thus achieved.

Radiation-damage-induced phasing (RIP)[118] is a technique whereby two datasets are taken from the same crystal but between the recording of the first dataset and the second there has been some deliberate damage to the crystal. This damage can be introduced in various ways, such as through the exposure to X-rays or ultraviolet light-emitting diodes[119]. If the damage can be made specific to a particular part of the unit cell and can be calculated from prior knowledge[120] or empirically[121] then the change in the diffraction pattern can lead to the determination of the phases and so the structure of the protein.

Finally, in experimental phasing, there is a variety of ways to combine the above methods to improve signal-to-noise in difficult cases. For example, isomorphous replacement can be combined with anomalous scattering to form single isomorphous replacement with anomalous scattering (SIRAS)[122] or multiple isomorphous replacement with anomalous scattering (MIRAS)[123]. Alternatively, radiation damage can be combined with anomalous scattering in radiation-damage-induced phasing with anomalous scattering (RIPAS)[124–126].

Overall, the purpose of all of these different methods is to determine an initial estimate of phases before refinement of these phase estimates. Refinement is the improvement of the initial phase estimates to create a better model of the protein. The experimental phasing methods all usually require several crystals and diffraction measurements, so molecular replacement is typically attempted first, as computer analysis is less time consuming. Therefore, the next section details how molecular replacement allows the calculation of an initial estimate of the phases.

1.6.2 Molecular replacement

Molecular replacement involves the use of a known structure to form an initial estimate of the phases of the diffraction pattern of the unknown protein. The known structure needs to be similar to the unknown protein. The process involves taking the known structure and rotating and translating it in an asymmetric unit to find the best possible match with the observed data.

The success of molecular replacement[127] is highly dependent on the quality of the input model, and its similarity to the protein of interest. The input model must be as similar as possible to the structure being solved, with between 25–35 % sequence identity to the target structure being considered a good minimum. The input model is then orientated and positioned in the cell to cover all cases within acceptable computational time limits. From each distinct placement of the model, an initial estimate of the quality of fit can be obtained. Once the optimum location and orientation are determined, the most likely phases can be calculated, which then allows for an initial calculated electron density to be generated. Once the optimum placement is determined, model building and refinement can be attempted.

In molecular replacement, there are two broad methods that most software use, namely Patterson-based methods[128–130] and maximum likelihood methods [131–135]. Both methods usually split up the six dimensional searches into two separate three-dimensional searches. First, a rotation search of the input model is undertaken and then, taking the output from this, a translation search is carried out. The two three-dimensional searches require far less computational power than the six dimensional search and so are more efficient in finding the most likely solution.

1.6.3 Patterson-based methods

These methods are based on Patterson maps[136], which are vector maps containing peaks that represent the interatomic distance vectors between different atoms. The intensities of the peaks are proportional to the product of the atomic numbers of the atoms at the two ends of the vector. At a given displacement from the origin, **r**, the Patterson function, *P(**r**)*, can be mathematically described through the use of an inverse Fourier Transform, FT^{-1}, of the intensity of a diffraction vector, *I(**S**)*, in the following way.

$$P(\boldsymbol{r}) = FT^{-1}[I(\boldsymbol{S})]$$

Equation 1-11

The above equation is useful because it is dependent only on intensities, and hence avoids the phase problem described before. Thus Patterson maps can be calculated immediately from the data obtained from the X-ray diffraction experiment and need no phase information. The Patterson map produced is the self-convolution of the electron density $\rho(\mathbf{r})$, i.e. the convolution of the electron density with its inverse

$$P(\boldsymbol{r}) = \rho(\mathbf{r}) * \rho(-\boldsymbol{r})$$

Equation 1-12

Once a Patterson map has been generated from the input model to the molecular replacement process, it can then be compared to the Patterson map calculated from the experimental data. Usefully, following the divide-and-conquer strategy[112,127,137,138], the intramolecular vectors only depend on the rotation of the input model (and so the input model's Patterson map) compared to the Patterson map of the diffraction data, allowing the rotation search to be undertaken first. The intermolecular vectors depend on both the rotation and translation of the input model, so this now allows a translation search to be undertaken using the results from the rotation search. Once the model is placed from these two three-dimensional searches an initial phase estimate can be given to the data of interest.

1.6.4 Maximum likelihood methods

The maximum likelihood methods are based on the idea that the best model will fit the data the best. The best model can be chosen by picking the model that gives the highest probability of observing the data[139]. This modelling is achieved by using a likelihood function, which is the probability of observed structure factors, given the calculated structure factors. It should be noted that care needs to be taken to consider Bayesian statistics in using this method. In particular, we can measure the probability of

$$p(F_O; F_C)$$

Equation 1-13

where $p()$ is the probability, F_O is the observed structure factor, ";" indicates a probability on the left hand side, given that the observables on the right hand side have occurred, and F_C is the calculated structure factor. However, the result required is

$$p(F_C; F_O)$$

Equation 1-14

This is given by the equation below (Bayes' theorem):

$$p(F_C; F_O) = \frac{p(F_O; F_C)}{p(F_O)} \times p(F_C)$$

Equation 1-15

The probability of the data will not vary as the model is changed, so the equation can be simplified to

$$p(F_C; F_O) \propto p(F_O; F_C) \times p(F_C)$$

Equation 1-16

Choosing the best model is performed by scoring the results (identifying the most likely) of placing a search model in different locations and orientations in the asymmetric unit. These individual scores are often small numbers, and as they must be multiplied to determine overall probabilities, computational complications arise, and it is better to work in logarithmic values. The multiplications are then converted to simple additions, and a much higher dynamic range can be accommodated in the computer programming. As the functions in the implementations often involve minimisation rather than maximisation of results, minus log likelihoods are often used.

Both rotation and translation searches are needed to find the most likely location. One option is to perform a six-dimensional search, but this is computationally very demanding.

However, rotation and translation searches can be undertaken separately, as in the contribution from a single model, translation only affects the phase, whereas rotation affects both the phase and the amplitude of the reflections (rigid model). As in Patterson-based methods the rotation search is done before the translation search, reducing the task to two three-dimensional searches rather than a full six-dimensional search. For likelihood to work well, errors need to be accounted for and incorporated in the likelihood functions.

1.6.4.1 Rotation search function

In the rotation search function, the translation of the search model for a particular orientation is not yet known, so the relative phases of the symmetry related copies are not known. Therefore, for a trial orientation, the structure factor amplitude for every copy is known, but not its phase, or its phase relative to other copies of the search model. Accordingly, addition in the complex plane to generate a resultant structure factor can only result in a probability distribution function. This distribution can be calculated for an observed structure factor, given the set of structure factors in the trial rotation.

This distribution can be approximated by a Wilson distribution[140]. The Wilson distribution is the probability distribution of structure factors when the composition of the unit cell is fully known, but none of the position of the atoms has been determined. It has a mean of zero and a variance of Σ_N. In this application, the random walk used in deriving the Wilson distribution is considered for the molecules rather than for a distribution of atoms in an asymmetric unit.

Model errors need to be included in considering the structure factors. This reduces the contribution of each factor by *D*, the fraction of the structure factor that is correct, and increases the variance. Equations that reflect these considerations are shown below. The first equation shows the probability of an observed structure factor, given the set of structure factors in the trial orientation of the search model.

$$p\left(F_O;\{\boldsymbol{F}_{jk}\}\right) = \frac{2F_O}{\epsilon\Sigma_W}\exp(-\frac{F_O^2}{\epsilon\Sigma_W})$$

Equation 1-17

where F_O is the observed structure factor, $\{\boldsymbol{F}_{jk}\}$ is the set of contributions of calculated structure factors from the symmetry copies k of molecules j, ϵ is the expected intensity factor, D_j is the Luzzati weighting factor[141], Σ_W is

$$\Sigma_W = \left[\Sigma_N - \sum_j \sum_k D_j^2 \Sigma_N\right] + \sum_j \sum_k D_j^2\, |\mathbf{F}_{jk}|^2$$

Equation 1-18

and $\Sigma_N$is defined as

$$\Sigma_N = \langle\frac{F_O^2}{\epsilon}\rangle$$

Equation 1-19

The contribution to the variance indicated within the square brackets reflects the error in the model. If the model is unknown, this D_j factor is zero, and the resultant distribution reduces to a Wilson distribution with variance Σ_N. If the model is perfect, D_j is one and the term in the square brackets goes to zero.

The probability distribution of the observed structure factor, given the set of structure factors in the oriented search models in a unit cell can be better modelled by the Sim distribution. The Sim distribution[142] extracts the largest structure factor and ascribes it an arbitrary fixed phase. A distribution of the remaining structure factors is then created around this defined offset. The equivalent equations are:

$$p(F_O; \{\boldsymbol{F}_{jk}\}) = \frac{2F_O}{\epsilon\Sigma_S} \exp(-\frac{F_O^2 + F_{big}^2}{\epsilon\Sigma_S}) I_0(\frac{2F_O\ F_{big}}{\epsilon\Sigma_S})$$

Equation 1-20

where F_O is the observed structure factor, $\{\boldsymbol{F}_{jk}\}$ is the set of contributions of calculated structure factors from the symmetry copies k of molecules j. F_{big} is the calculated structure factor from the biggest contribution, ϵ is the expected intensity factor, and Σ_S is defined as

$$\Sigma_S = \left[\Sigma_N - \sum_j \sum_k D_j^2 \Sigma_N\right] + \sum_j \sum_k D_j^2 \, |\mathbf{F}_{jk}|^2 - F_{big}^2$$

Equation 1-21

where D_j is the Luzzati weighting factor, and $\Sigma_N$is defined as in Equation 1-19.

However, this distribution is not readily integrated and it is often not easy to determine which will be the largest structure factor.

1.6.4.2 Translation search function

Having completed a rotation function search, a list of possible rotations of the search model is available. All the possible relative phases between the structure factors of the symmetry related copies are now known for each possible rotation, so these can be summed to give a calculated structure factor with known amplitude but unknown phases, The scoring within the translation function can now be calculated by integrating out the nuisance variable (the phase of F_O). The translation function itself is indicated below:

$$p(F_O; F_C) = \frac{2F_O}{\sigma_\Delta^2 + \sigma_\mathrm{F}^2} \exp(-\frac{F_O^2 + D^2 F_C^2}{\sigma_\Delta^2 + \sigma_\mathrm{F}^2}) I_0(\frac{2F_O \; DF_C}{\sigma_\Delta^2 + \sigma_\mathrm{F}^2})$$

Equation 1-22

where F_O is the observed structure factor, F_C is the calculated structure factor, D is the Luzzati weighting factor, ϵ is the expected intensity factor, σ_F^2 is the inflation of the variance due to experimental error and

$$\sigma_\Delta^2 = \langle \frac{F_O^2}{\epsilon} \rangle - D^2 \langle \frac{F_C^2}{\epsilon} \rangle$$

Equation 1-23

This search only compares the intensity of the reflections and does not include any known phase information from an input electron density map.

1.6.4.3 Phased translation search function

A different way of scoring the translation search is to use a phased translation search function[134], which is a way of correlating the electron density predicted by the trial orientation with the electron density in the input map of the unit cell. The implementations so far do not include model errors, as these are not easily included in the scores. The search is performed in reciprocal space to make it less computationally expensive. A finite value can be assigned to the correlation between the two electron densities, ranging from -1 to +1, although values below zero are rarely seen. This correlation can be stated as

$$\rho = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 + \sum_i (Y_i - \bar{Y})^2}}$$

Equation 1-24

where X_i and Y_i are two variables, and $\bar{X}$is the mean, and ρ is the cross-corelation. However, when looking at electron densities, the real space correlation coefficient can be defined as:

$$\rho = corr(\rho_P, \rho_M) = \frac{cov(\rho_P, \rho_M)}{[var(\rho_P)\ var(\rho_M)]^{0.5}}$$

Equation 1-25

where var is the sample variance and cov is the sample covariance, ρ_P is the observed electron density given by the prior information i.e. the input map, and ρ_M the calculated electron density i.e. the search model after rotation and translation. This can then be written more succinctly in the following way

$$C(t) = \frac{\int_V [\rho_P(\boldsymbol{x}) - \bar{\rho}_P][\rho_M(\boldsymbol{x}-\boldsymbol{t}) - \bar{\rho}_M]d\boldsymbol{x}}{\{\int_V [\rho_P(\boldsymbol{x}) - \bar{\rho}_P]^2 d\boldsymbol{x} \int_V [\rho_M(\boldsymbol{x}-\boldsymbol{t}) - \bar{\rho}_M]^2 d\boldsymbol{x}\}^{0.5}}$$

Equation 1-26

where $C(t)$ is the correlation coefficient as a function of translation. The integrals are performed over all space. Now, if we assume that the average electron density is zero (so that in effect we do not have the *F(0,0,0)* term), we can simplify it to the following

$$C(t) = \frac{\int_V \rho_P(\boldsymbol{x})\rho_M(\boldsymbol{x}-\boldsymbol{t})d\boldsymbol{x}}{\{\int_V \rho_P(\boldsymbol{x})^2 d\boldsymbol{x} \int_V \rho_M(\boldsymbol{x}-\boldsymbol{t})^2 d\boldsymbol{x}\}^{0.5}}$$

Equation 1-27

If we write out all the electron density using the Fourier Transforms and the cross-correlation theorem, we find the following.

$$C(t) = \frac{V^{-1}\sum_{\boldsymbol{h}} m_p|F_O(\boldsymbol{h})|e^{i\alpha_P}F_M^*(\boldsymbol{h})e^{-2\pi i\boldsymbol{h}.\boldsymbol{t}}}{(\frac{1}{V})[\sum_{\boldsymbol{h}}\left(m_p|F_O(\boldsymbol{h})|\right)^2 \sum_{\boldsymbol{h}}(|F_M(\boldsymbol{h})|)^2]^{1/2}}$$

Equation 1-28

where m_p is the figure of merit, and V is the volume of the unit cell. This is the equation that is implemented in *phaser*[4].

Whenever scores are being calculated, the contribution from the entire unit cell must be taken into consideration, making use of the known crystallographic symmetry operators. This allows for the correct scoring of the rotation and translation functions of a model in the crystal. However, when searching, the search only needs to be undertaken over the asymmetric unit, as this will sample all possible positions of the model without repetition of positioning.

1.6.4.4 Packing Function and Refinement within molecular replacement

After obtaining a list of likely translations from the translation functions above, there are some ways of improving the results, particularly when placing more than one copy of the molecule in the asymmetric unit. A packing test is done to ensure that two molecular replacement solutions do not overlap. One way of implementing this is by checking that no C^{α} are within 2 Å of each other from either solution. Any two solutions that disobey the above rule will lead to a rejection of one of the structures.

Secondly, the solutions can be refined. In the rotation and translation search, there has been a grid through which the search model has been rotated and translated. By allowing a rigid-body refinement, an improved solution can be obtained.

1.7 Improving the initial phases

From the molecular replacement analysis, a candidate electron density map can be calculated. Ideally, model building could be undertaken to improve the model of the protein based on this map. Improved phase information can be calculated from this model, leading to a better electron density map, and so a better model (Section 1.9) to feed into the analysis. Thus iterative cycles can be done to obtain a final model of the protein that matches the experimental results as closely as possible

However, there are many occasions where the initial electron density map is too poor and noisy to build a good original model, and so the new estimate of phases is no better than the initial estimate of phases. In this case, iterative model improvement cannot be started and so the structure cannot truly be solved. Therefore the initial estimate of phases needs to be improved upon and the wide range of methods that undertake to achieve this are collectively known as density modification[143].

To improve the initial phases, heuristic constraints are placed upon either the electron density map or the structure factors. There are a variety of methods to do this, including solvent flattening[144–146], iterative skeletonisation[147], histogram matching[148], Sayre's equation[149] and molecular averaging[150,151].

The technique of solvent flattening[144–146] first scans through the electron density map and draws a mask around all regions in the asymmetric unit that are identified as part of the protein that is being crystallised. Given that the protein is highly ordered, the electron density is expected to vary greatly, so this characteristic is used to identify these areas. Solvent, on the whole, is disordered and has little to no long-range order. Hence the electron density in this bulk solvent should average to a uniform electron density. If there are any peaks in the bulk solvent region this must be as a result of phase errors, so these are flattened as the second step. This should reduce the phase error and lead to improved phases. The most difficult part of solvent flattening is defining a good mask around the protein region, which significantly affects the success of this method. Once this is defined, everything outside of it is considered to be bulk solvent and can be flattened i.e. reduced to a constant value (often an average mean value). Considering that the bulk solvent accounts for 30–70 % of the unit cell in a crystal, the scope for improvement can be large.

Iterative skeletonisation[147] uses the idea that protein is a polymer, which is continuous throughout the length of the protein. Therefore electron density should be continuous and connected for it to be a protein. If the choice of phases lead to increased compliance with this criterion then it should result in improvement in the initial estimate of phases. In reality, this technique is rarely used, as building an atomic model to the density is far more powerful than iterative skeletonisation.

Histogram matching[148] takes advantage of the fact that all proteins are made up of only twenty amino acids and so all have similar chemical properties. Therefore the electron density distribution should be primarily dependent only on resolution and overall temperature-factor values (Section 1.9). A considerably different electron distribution is likely to rise from phase errors. Therefore changing the electron density so that the distribution matches the expected distribution tends to improve the phases and make the electron density easier to interpret.

Sayre’s equation[149] uses the concept (which is one of the direct methods approaches that are more commonly used in small molecule crystal studies) that electron density must always be positive. It also requires atomicity, meaning that all the electrons are situated around the location of the nucleus, regardless of the element. The equation F_h for the structure factor at reciprocal vector h is given by

$$F_h = q \sum_{h'} F_{h'} F_{h-h'}$$

Equation 1-29

The sum is calculated over the whole of reciprocal space, and q is a constant dependent on $\sin(\theta) / \lambda$, where θ is the angle of incidence of the X-ray beam, and λ is the wavelength of the X-ray beam. This equation requires a structure at atomic resolution and uses the idea that if such a very high resolution electron density map (which is made up of individual peaks corresponding to the individual atoms) is squared, then a map which has a similar shape with separate peaks is obtained, leading to unchanged phases. The convolution theorem then implies that there is a relationship between the phases of different structure factors.

Molecular averaging[150,151] is very commonly used when trying to improve the initial estimate of phases. It makes the assumption that two molecules of the same protein that are in different crystallographic environments are going to be identical. These different crystallographic environments can be two physically different crystals of the same protein (usually crystallised in different conditions). Alternatively, it can be two or more copies of the protein within the same crystal not linked to each other by any crystallographic symmetry. The vectors between these copies are described by non-crystallographic symmetry (NCS) operators. Any differences in the calculated electron density maps between the copies when the NCS operators are applied must be due to phase errors. Examining the difference between the various structures, the phases of both copies can be improved. This molecular averaging can be used even when only part of the protein is similar (partial NCS) or when different domains of the protein obey different NCS relationships (multiple-domain NCS). All of these NCS relationships can be used to help improve the initial estimate of phases and, given that they are found in over a third of Protein Data Base (PDB) [152,153] structures, this technique can be quite useful.

There are a number of alternative ways of determining NCS relationships/operators. The simplest way is to solve the structure using molecular replacement. If the input model has already been placed at different points in the asymmetric unit then this will usually provide the NCS operators in the asymmetric unit.

A different approach can be taken when the sites of the heavy atoms or anomalous scatterers are known. If there are more than two of these sites per copy of the protein, then the relationship between the monomers can be derived by simple geometry and the NCS relationship calculated[154,155]. This technique does not work if there are only one or two sites per monomer. It also relies upon the heavy atom or anomalous scatterers binding at the same places in each monomer of the protein, which is not always true for crystals soaked in solutions containing heavy atoms.

GETAX[156] uses a poor electron density map to calculate proper NCS i.e. NCS operators that are linked by point group symmetry. *GETAX* only uses low resolution data up to 6 Å and by a brute force search determines whether any points are linked by a proper rotation axis through comparison of the electron density.

Lastly, there is the most generic approach. Here the electron map is looked at directly, and the similarity in the electron density between one part of the asymmetric unit and another part is determined[157]. This does not require any other information, such as heavy atom sites or anomalous scatterers, and still gives the NCS operators. This approach has been implemented in programs such as *phenix.find_ncs_from_density*[2], and is described more fully in chapter 3.

Through application of one, some or all of the above methods, the intention is to generate enough information to have a sufficiently good initial estimate of the phase to continue with the iterative solution of the structure of the protein.

1.8 **Model building**

Given the initial or improved estimate of phases and with the initial or improved electron density map calculated, a model can be built of the structure. Originally, model building of the electron density was undertaken using a stack of Perspex sheets, and they were painstakingly assembled by the scientists.

A later development is known as Richards' box[158]. This was a large optical comparator device which allowed for the creation of complex models, and was highly popular in the 1970s. Eventually, computer based systems, such as Frodo[159], were developed to allow scientists to rapidly build models.

Later model building was done manually on a computer, using visualising software such as *O*[160]. Nowadays, model building is done mostly automatically through programs such as *phenix.autobuild*[161], *Buccaneer*[162] or *ARP/wARP*[163,164]. There are different parts to this model building, including main chain tracing, secondary structure recognition, building of loops, placing of side chains, ligand building and solvent building. All of these parts are done in an iterative fashion. The intention is to build the best possible initial model to start the refinement process.

It should be noted that, for many of the later refinement and structural validation steps, hydrogen atoms need to be added to the appropriate bond positions to create the entire model. *REDUCE*[165] can be used to ensure that this is done appropriately.

1.9 Refinement and Structural validation

Once the structure has been built it should be refined to improve its match with the diffraction data (refinement). However, in order to know how best to improve the structure, structure validation is helpful. This provides some level of confidence that the model is the best possible solution for the data observed and also gives an idea of the areas that require improvement. It gives confidence that the model is a correct representation of the protein in the crystal, and the structure can then be applied to understand how the protein works and why certain mutations might be harmful.

There are two different ways in which chemical knowledge from previous structures can be introduced into a refinement. Firstly, constraints, which imply a mathematical equality, can be defined. For example, one common constraint is reducing the three anisotropic B-factors that an atom can have in three-dimensional space into one isotropic B-factor. This has the effect of reducing the number of parameters that need to be found in a refinement. Secondly, restraints, which imply equality within some experimental error, can be applied. Bond length is a good example of a restraint. A typical bond will have a characteristic length but there will be a distribution around this length in different environments. Therefore this bond length has an ideal value, but has a spread around this figure. In the refinement this has the effect of increasing the available data, as it can provide an extra energy term describing the likelihood of this bond length.

The various refinement programs set different targets to best improve a structure, and sequence through different algorithms to achieve a solution. These algorithms need to balance fitting the observed data with ensuring that the model makes chemical sense e.g. correct bond lengths, angles, no steric clashes. If no restraints of this type are set then the protein becomes over-fitted to the data. Refinement methods balance the match to the data with the match to previous chemical knowledge.

There are a number of refinement programs that use various statistical measures in their calculations, leading to the generation of models that are closest to the protein structure itself. Two examples of these are *REFMAC5*[166] and *phenix.refine*[167].

Another useful part of refinement is manual fitting. Despite the recent advances in automated model building and codifying human pattern-recognition processes, an experienced scientist can still improve on the output of current automated algorithms. Automated programs are very strong in placing backbone chains in the electron density map. They are also good at making sure that the side chains are pointing in the correct direction. However, they are poor at deciding on protein masks in maps of lower quality, and in determining secondary structure in these situations.

This manual fitting is often done through a program such as *Coot*[23], which allows manual manipulation of the model. It generates a variety of metrics to allow the user to improve the structure and identify whether the structure present is supported by the electron density maps. This program will be discussed in more detail after the various measures in model quality are considered.

Programs which assess the protein model quality are available. One of the most commonly used programs is *MolProbity*[168,169]. The PDB itself also publishes statistics about any of the proteins that it stores.

1.9.1 Global measures

There are two types of measure of the model quality. First there are global measures, which measure how well the model fits the data overall. The R-value (sometimes called the residual factor or the reliability factor) measures the fit of a simulated diffraction pattern and the experimentally-observed equivalent. It is defined by the following equation:

$$R = \frac{\sum_h \sum_k \sum_l ||F(hkl)|_{obs} - |F(hkl)|_{calc}|}{\sum_h \sum_k \sum_l ||F(hkl)|_{obs}|}$$

Equation 1-30

where $|F(hkl)|$ is the amplitude of the structure factor, with the subscripts "obs" indicating that it is taken from the observed data set and the subscript "calc" from the calculated dataset.

A totally random set of atoms will result in an R-value of about 63%, and a perfect fit a value of 0%. Typical values of models in the PDB data base are around 20%, dependent largely on resolution. However, the refinement process introduces bias into the matching of the calculated electron density with the observed data, as the atomic model and the measured diffraction pattern are both used to calculate the electron density. Setting aside a fraction of the experimental observations and restricting the refinement to using the remainder permits the generation of an R-free value[170] by calculating how well the model predicts the set aside data. Ideally, the R-value and the R-free values will be very similar. If these differ significantly, then over-fitting is likely to have occurred. The value of this difference that is significant is hard to define, and is resolution dependent, but a value of 3 %-5 % is typically a reasonable cut off[171]. The R-value does not become arbitrarily low as knowledge of chemical bonds is included in the creation of a model. This helps ensure that the model is not over-fitted to noise. The presence of bulk solvent is also not well modelled in the crystal structure, which also sets a lower limit on the R-value, as $\boldsymbol{F}(hkl)_{calc}$ is incorrectly calculated.

The temperature factor (also known as the B-factor) is a measure of the disorder at the site of a given atom in the structure. It describes the shape of the electron distribution of the atom, averaged over many unit cells. There is global inherent thermal motion of the atom about its mean position within the protein, and a second, larger effect due to alternative conformations of sidechains being possible. Higher B-factors indicate a broader spread of this distribution, signalling that the atom can adopt many more positions.

1.9.2 Local measures

There are a variety of measures of local structure that can be used to analyse the protein. One of the simplest measures is looking at the angles along the backbone. Any models that lead to unfavourable angles that are not seen in other structures are likely to be incorrect. Furthermore, there are only certain angles that φ and ψ can take to avoid steric clashing along the backbone[19]. *MolProbity*[168,169] has a reference of 100,000 residues from 500 protein models, each of which has been assessed to ensure that the relevant angular values are of high quality (manually curated). At higher resolution, more data are available to support the placing of Ramachandran outliers[172]. Residues with higher B-factors (temperature factors)[173] have higher uncertainty, allowing such outliers to be accommodated more easily. Both are taken into account by *MolProbity* when calculating an overall score.

Another examination that can be undertaken is an all-atom contact analysis. The idea behind this is to ensure that atoms are not too close to each other in space. Whilst atoms experience Van de Waals forces and hydrogen bonds which keep the structure together, if the atoms get too close there is a very large repulsive interaction, which prevents them from getting any closer. Thus these clashes are far more likely to be due to incorrect modelling than to a genuine rare clash. *MolProbity* undertakes this all-atom contact analysis by conceptually rolling a 0.5 Å sphere over the entire Van de Waals radius of the atom and looking at any possible clashes that might occur with neighbouring atoms.

There are two primary outputs for these data. Firstly, in a molecular graphics programs, such as *Coot*, there is a visible representation of the results. In such an output, green dots indicate atom surfaces exhibiting 0.25 Å of overlap, which has no energy penalty, along with pale green for hydrogen bonding interactions[165,174]. As they approach and get closer and more unfavourable, the colour turns from green to yellow to bright red spikes to indicate increasingly unfavourable interactions. The second output from this investigation of possible unfavourable atom overlaps is a clash score, which can be reported back giving an overall statistic of the quality of the structure. There are many different quantitative scores that could be used in defining this clash score, but the simplest version (number of overlaps > 0.4 Å per thousand atoms) is the one used throughout *MolProbity*.

Another measure is the rotamer analysis[21]. This checks all of the side chains and ensures that they adopt a conformation that has been seen in other proteins. It gives a probability of seeing the different conformations and makes sure that the side chains are not just being fitted randomly or to noise in the map.

Peptide ω analysis concerns the backbone of the protein. The ω value should always be around 0 ° or 180 °, as this allows for the nitrogen lone pair to donate into the C=O π* orbital and so allows a favourable conjugation to occur. Furthermore, predominantly trans-amino rather than cis-amino acids should be present in the protein (unless the residue in question on the amide side is proline). Therefore the addition of too many cis-amino acids, which can be seen by the ω analysis, should be avoided, as they are unlikely to be true. Indeed it has been found in lower resolution data that cis-residues seem to act as if they fit the electron density map better, even though they are less likely to actually be present in the protein.[175]

Geometry analysis is the analysis of all the dihedral angles in the protein. This is achieved by reviewing the C-α sites along the backbone, and ensuring that the secondary structure is maintained. Once such implementation is *CaBLAM*[176].

Density-fit analysis looks at residues on a local scale to determine whether the protein fits the electron density map around it. This ensures that the placement of residues is supported by the data.

The puckering of sugars in RNA are particularly hard to determine in electron density. Restraints can be added in refinement to maintain the correct angles, and deviations can be reported.

1.9.3 Coot

Coot is a commonly-used program for visualising and improving model structures of proteins. It allows comparison of the model to electron density map data by creating a mesh at a certain contour level, which can be easily adjusted, and displaying the difference maps as green and red, showing clearly the places where more or less electron density should be modelled respectively. These contour levels are defined by mapping out a level representing a given value of standard deviation from the mean electron density in the map.

Coot itself is able to show a number of the measures that have already been discussed, including the Ramachandran plot[173], geometry analysis (using the mmCIF REFMAC dictionary[166]), peptide ω analysis, rotamer analysis[21], density-fit analysis and probe clashes[165,174]. These are either shown on the map or in pop-up windows, with most measures updating automatically as the model is corrected.

Moreover, *Coot* is fully integrated into other programs, such as *CCP4*[177] and *phenix*[4]. This allows it to be loaded as soon as model building is complete, with the full information from *MolProbity* also included, thus permitting the rapid improvement of the model and allowing for the best possible structure to be created.

Coot also allows the protein structure to be explored in detail interactively, and the sites of mutations (i.e. local changes in the residues) to be rapidly analysed. It allows residues to be manipulated easily, and steric or chemical effects arising from residue change to be assessed. It also allows residues to be added or deleted easily and sidechains to be mutated or stubbed. Thus effects such as steric clashes, removal of disulphide bonds or electrostatic interactions can be seen, and the effect of mutations on the stability of the protein, or effects on its activity, can be hypothesised.

1.9.4 Other factors considered in refinement

There are other factors considered in refinement, which are less relevant to structure validation, that should be briefly discussed. Firstly, refinement programs can take into account any NCS relationships that have already been discovered, as this helps to the creation of a better model. Secondly, occupancy refinement[165] can be used. This can be applied to the residues, if they can form one of two possible alternate conformations, or to the ligands, which only partially occupy some of the sites. Accounting for this allows for better structures to be obtained.

A bulk solvent model is also required, as this makes up anywhere between 30-70 % of most unit cells. There are a number of ways in which the bulk solvent can be modelled. For example, in REFMAC5 there are two different models available. These are the Babinet's bulk-solvent model[178,179], which uses an assumption that the only difference between solvent and protein at low resolution is a scale factor, and the model proposed by Jiang & Brünger[180], which uses the assumption that there is a uniform distribution of solvent over the region where there is not a protein.

Finally problems in the data need to be managed. A good example is twinning, which can severely hamper the efforts to solve a protein structure, and leads to significant changes in the targets that are used. Mosaicity can also be an issue and can result in difficulties in indexing and integration in particular. Incorrect indexing of points on the diffraction pattern can also complicate the refinement of the structure solution. Furthermore, incorrect space group determination can lead to significant problems.

1.10 Other methods for finding protein structure

Evaluation and refinement of the structure of proteins using X-ray diffraction has been presented, and the resultant structure will be deposited into the PDB when it is satisfactory. X-ray diffraction is currently the main source of protein structures for scientists, but there are other techniques that allow the structures of proteins to be elucidated.

1.10.1 NMR structure determination

Nuclear magnetic resonance (NMR) is one such technique that can be used to determine a model of the protein. The technique relies on resonances being observed in magnetically active spin nuclei. These resonant frequencies depend very much on their chemical environment of the atom, and their precise location gives structural information. There have been a variety of high resolution structures obtained from this technique, including the murine prion proteins[181] (central backbone to a RMSD between the models of 1.4 Å, with all heavy atoms with RMSD 2 Å).

It is a highly complementary technique to protein X-ray crystallography for solving a protein structure. Unlike macromolecular X-ray crystallography, the protein is in solution and so no crystals are needed for the structure to be obtained. NMR also reveals the ionisation state of the side chains of the protein far more easily than a protein X-ray crystallographic model and allows for transient weak complexes formed in solution to be explored. Furthermore, a number of models are obtained, so it is easy to see where the flexible regions of the protein are when in solution. However, the protein has to be relatively small (under 50 kDa) in order to be easily solvable. NMR also tends to provide less detail than an X-ray crystallographic protein structure and it requires ^{13}C and ^{15}N to be incorporated to allow the structure to be solved[182].

1.10.2 Protein Neutron Crystallography

A complementary technique is neutron diffraction[183,184]. This technique uses neutrons which diffract from the nuclei of the atoms, rather than from the electrons, allowing very good placement of the positions of the hydrogen atoms in the structure, which are usually poorly determined in X-ray diffraction methods. This can be highly important, such as in the case of aldolase reductase[185] where, despite a 0.66 Å X-ray structure, it was the neutron structure that allowed determination of the position of the catalytically active hydrogen atoms.

However, there are significant difficulties in using protein neutron diffraction crystallography. Firstly, very large crystals of the protein are needed, which for most proteins are impossible to produce. Secondly, it takes a long time to collect the data (hours to days), as neutron sources tend to have low beam fluxes. This therefore means the crystal must also be stable over a long period. Both of these are severe limitations in practice and so prevent neutron diffraction being able to assist in solving many structures.

It has, however, provided illuminating insight in some cases and is a useful technique to use if possible. The software developed for protein X-ray crystallography can be used to help refine its structures, as seen in *phenix*[4]. Effective comparisons and useful conclusions can be drawn by considering X-ray and neutron data together[186].

1.10.3 Electron Microscopy

A technique that has developed significantly in recent years is cryo-electron microscopy[187–189]. The resolving power for structures has become higher and higher, and indeed in September 2016 a 1.8 Å structure was solved[190]. The main benefit of single particle cryo-EM is that a density map can be reconstructed directly from the experiment. From that point on, many of the refinement steps used in protein X-ray crystallography can be used with minor adjustments and so software already developed can be rapidly adapted. Indeed, much software is now being developed specifically for this technique, such as CCP-EM[191] and *phenix.real_space_refine*[192]. The other advantage of this technique is that crystals are not needed (it uses a thin layer of frozen protein solution), thus it works well for proteins that are usually hard to crystallise, such as large proteins. However, cryo-electron microscopy is still developing and does not yet have the same high throughput as protein X-ray crystallography.

Chapter 2 Cytokine receptor-like factor 3 (CRLF3) protein

2.1 Introduction to CRLF3

CRLF3 is a protein that has been ascribed many functions in the human cell cycle in earlier work. To date there has been little clarity on specific functions (Section 2.2), but recent work has indicated its importance in blood clotting in a mouse model. No structure had been proposed for this protein.

The group led by Dr Cedric Ghevaert within the Department of Haematology at the University of Cambridge had a particular research interest in Cytokine receptor-like factor 3 (CRLF3) protein, due to observations of the phenotype of CRLF3 knockout mice in the International Knockout Mouse Consortium (IKMC)[1]. These mice exhibit a drop in the number of platelets to 60 % of the baseline level as they are inefficient in creating platelets (thrombopoiesis). This is believed to be due to an increased stability of micro-tubules leading to a failure of maturation of proplatelet/preplatelets, and so an increased clearance of the immature forms, leading to a drop in the number of platelets. A molecular structure was wanted for this protein to facilitate a better understanding of the function of the protein, and so an interpretation of how mutations could lead to the observed phenotypes.

Crlf3 is a 2.4kb gene and has been shown to express in the haematopoietic system. In *Ensembl*[193], the Crlf3 gene has 6 transcripts, along with 69 orthologues. CRLF3 has 442 amino acids, with a molecular weight of 49.6 kDa. Murine CRLF3 is 93% sequence identical to human CRLF3. All protein work was done with murine CRLF3. The isoelectric point is predicted to be 4.93 (Section 2.5). Neither transmembrane helices nor signal peptides are predicted. Sequence analysis indicated that a Fibronectin type-III domain is located between residues 181-274, along with coiled coils ascribed to residues 10-57 in *UniProt*[194]. Other names for this protein include p48.2, CYTOR4, CREME9, and CRLM9.

2.2 Previous work on CRLF3

Previous work on CRLF3 had not been linked to blood clotting, so the work of the group led by Dr Ghevaert was a new line of investigation. This work has shown CRLF3 knockout mice to have a 25-40% decrease in platelet count, but the platelets that are there function normally in *in vitro* assays. Through bone marrow transplantations in mice, it can be shown that the decrease is caused by abnormalities in haematopoietic cells. In the CRLF3 knockout mice, upregulation of megakaryopoiesis and normal looking proplatelet morphology is seen, suggesting that there is a down-regulation of platelet clearance with the loss of CRLF3 protein. As a result, platelets are being removed from the body quicker. When a spleen from a wild type mouse is placed in a Crlf3 knockout mouse, platelet levels return to normal within a week. Abnormally large platelet structures are seen in the circulatory system in older Crlf3 knockout mice, suggesting increased and aberrant tubulin expression in the platelets. This evidence led to the hypothesis that Crlf3 controls platelet creation by destabilising the tubulin. This hypothesis proposes that Crlf3 knockout mice have delayed maturation of the pre-platelets in the peripheral circulation, potentially due to increased structural stability caused by the aberrant tubulin. Therefore these immature forms are then removed, so there is a decrease in platelet count. The removal of platelets occurs in the liver and spleen.

Other work has tried to ascribe functions to CRLF3. Cytokines are molecules which mediate intercellular signalling. These cytokines interact with receptors, which then allow for a signalling pathway to be activated within a cell. CRLF3 has been previously classified as a type I cytokine receptor of group 1 due to sequence similarities. It has been grouped together with other receptors such as the thrombopoietin receptor, prolactin receptor, erythropoietin receptor and the growth hormone receptor[195]. These proteins tend to have a single transmembrane helix, and sit in the cell surface membrane. The cytokine binds to the receptor on the outside part of the protein, causing it to dimerise, and then allowing transduction through a JAK (Janus kinases)-STAT (Signal Transducer and Activator of Transcription proteins) signalling pathway[196]. CRLF3 has been described as an orphan receptor, as to date its ligand has not been identified.

Evidence of the importance of this protein in the cell cycle was identified in human embryo kidney cells. It has been linked with cutaneous squamous cell carcinoma, but the mechanism has not been established. This indicates involvement in the cell cycle. The deletion of a group of genes including Crlf3 led to neurological problems, suggesting that this protein has an important role in the development of these cells and the communication between neurones. In experiments with rat neuronal cells, the synaptic cells appeared abnormal if CRLF3 levels were reduced. Finally, CRLF3 has been a useful target protein in tests of protein production knockdown systems.

The function of p48.2, a protein which is identical to CRLF3, but lacking the first four residues, was analysed in human embryo kidney 293T cells, and evidence suggested that it was important in the cell cycle. In particular, in these cells it was found that an over-expression of p48.2 would keep cells in G0/G1, and not allow entry of the cells into S phase in cell division. p48.2 also appears to have a role in up-regulating cyclin-D1 and cyclin-D3, though how this affects the cell cycle is unclear[197].

In 2006, there was a suggestion that CRLF3 was upregulated in cutaneous squamous cell carcinoma[198]. This was one of many genes that was deregulated which would lead to the most common form of skin cancer amongst Europeans. The investigation was looking at the accumulation of mutations through the use of quantitative real-time reverse transcription PCR. Previous literature had already suggested that CRLF3 was involved with cell communication, and in particular had ATP/GTP binding. The role CRLF3 may have in cell communication might be important in this cancer, but there was little further evidence presented in this paper.

In 2009, there was a deep analysis of p48.2. This protein was found at human chromosome 17q11.2, and was part of the microdeletion found in some Neurofibromatosis type I (NF1). NF1, a multisystem genetic disorder characterised by skin disorders, which results in unusual cell differentiation and growth. This can lead to a broad spectrum of problems, from learning disabilities to benign neurofibromas and increased likelihood of cancer[199]. This microdeletion also led to lower IQ, cutaneous neurofibromas and dysmorphic facial features[200].

In a rat neuronal cell line (PC12 cells), CRLF3 was mutated using a gene trap method[201]. This led to a phenotype of abnormal morphology and atypical distribution of synaptic microvesicles. In order to distinguish between a lack of CRLF3 causing the phenotype, or a truncated protein with a dominant negative function, two experiments were carried out. Firstly, in the non-mutated cell line, siRNA was used to reduce the amount of CRLF3 produced, and the same phenotype occurred. In a different experiment, cDNA was used to express the truncated protein. This led to no phenotype, and so showed that the lack of CRLF3 protein was causing the phenotype. Therefore, from these results, it was suggested that CRLF3 has a role regulating neuronal morphology, and is potentially involved in the transport or biogenesis of synaptic vesicles. However, there does not seem to have been any follow up work to these results from November 2012[201].

In 2005, CRLF3 was used in the development of a tetracycline-controlled inducible iRNA System, and its visualisation in Western Blots was sufficiently good to make it useful as a protein to develop a system against chronic myelogenous leukaemia cells[202]. The expression of CRLF3 could thus be controlled, but it did not lead to further functional understanding.

There has been a study into an insect orthologue of CRLF3[203]. This has shown that in insects the orthologue was an erythropoietin receptor. A primary neuronal cell culture for the *Tribolium castaneum* beetle would only survive in the presence of erythropoietin. On eliminating Crlf3 protein expression via various knockdown techniques, these cells stopped surviving even with the presence of erythropoietin. This allowed the authors to suggest the receptor role, and further to speculate that the CRLF3 in vertebrates might have a tissue protection-mediating receptor role.

2.3 Starting Point for this work

Dr. Cavan Bennett, a PhD student in in Dr. Cedric Ghevaert’s group in the Department of Haematology at the University of Cambridge at the time, had been attempting to crystallise the full length murine CRLF3 protein to achieve a full understanding of its structure and so of its mechanism. However, the initial crystallisation trials had been unsuccessful. There had been difficulty in expressing the protein and in achieving the required concentration for crystallisation. The protein had to be expressed with a Glutathione S-Transferase (GST) tag attached to it in order to aid purification. One litre of bacterial culture produced only 0.6 mg of CRLF3 after GST cleavage and purification. A concentrated protein sample (5 mg ml^{-1}) was obtained from 4 litres of bacterial (*Escherichia Coli)* culture. This was just sufficient to allow crystallisation screens to be set up (using a vapour diffusion method). However, no crystals were formed (crystallisation work performed by Giles Lewis, Cambridge Institute for Medical Research, University of Cambridge, Cambridge UK). An alternative approach would be required to make progress in determining the structure of this protein.

Accordingly, a collaboration was set up with Dr. Cavan Bennett to progress the determination of the structure of this protein, and to examine the use of some of the underlying computational tools. Given the difficulties observed in crystallising the protein, this work was initially aimed at designing new constructs to achieve better crystallisation of the protein, then to solve the crystal structure of these constructs and finally to aid with the refinement and deposition of the structural models in the PDB database. The author was responsible for the design of the constructs and all analysis of diffraction patterns and structure determination, with the collaborators manufacturing the constructs. Dr Yahui Yan performed the crystallisation and led the remote data collection.

2.4 Flow of experimentation and analysis

Flow chart (Figure 2-1) indicates the sequence of the work described in this chapter. Starting from the position described in Section 2.3, experiments were conducted to create crystals suitable for X-ray diffraction experiments to determine the structure of the CRLF3 protein or suitable constructs. This work was carried out in stages, with the results of each stage determining the next steps within this overall aim.

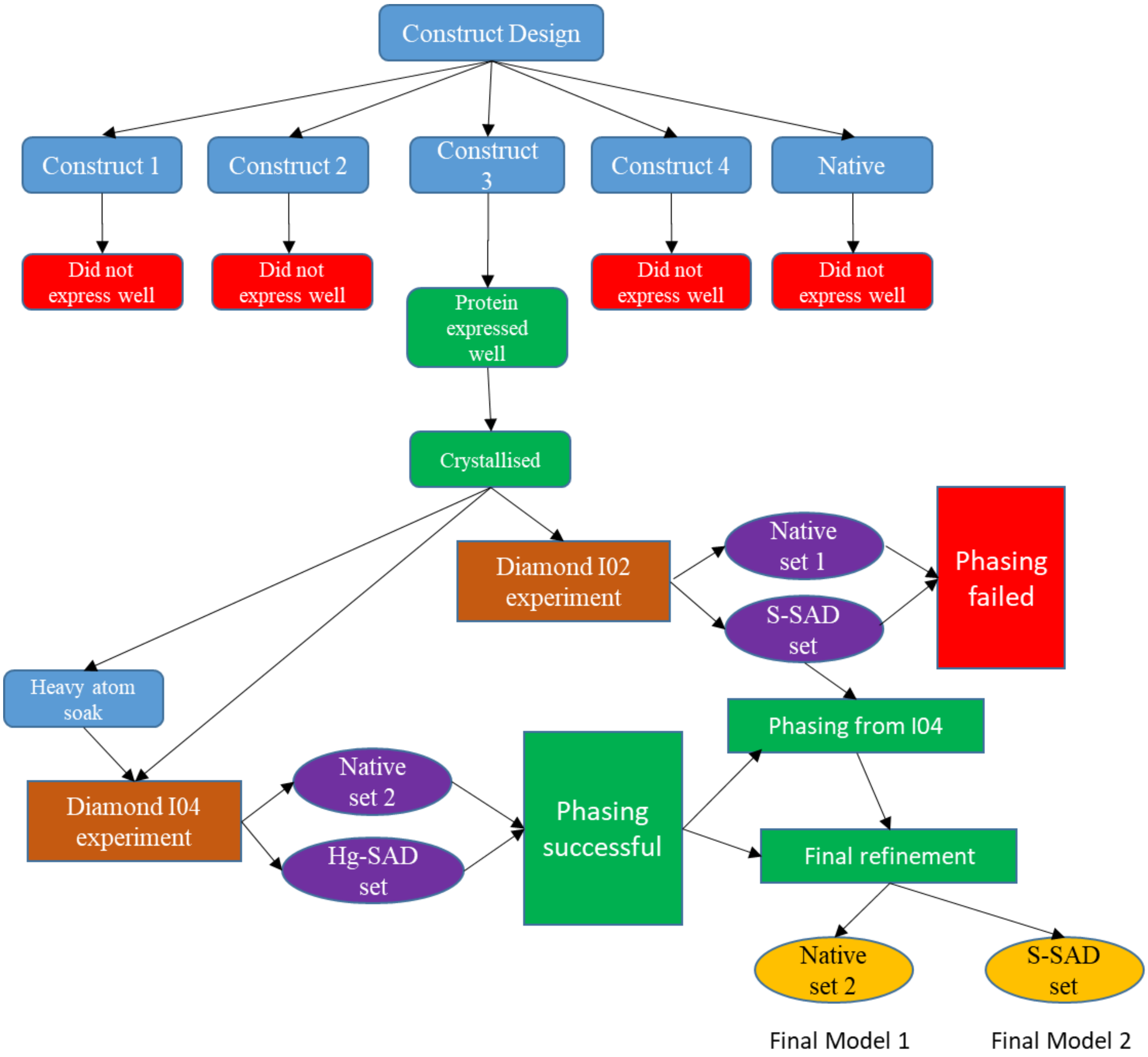


Figure 2-1: Stages in determining the structure of the CRLF3 construct 3

Flowchart generated in *Microsoft PowerPoint.*

2.5 Stage one: Designing the constructs

Firstly, the question of designing a construct was considered. There was concern that the full length protein might not crystallise (Section 2.3) and so suitable constructs were required. Therefore, to obtain an initial understanding of the protein, the online webserver *HHPRED*[204] was used.

HHPRED uses the hidden Markov model (HMM)[205]. A Markov model compares sequences; a hidden version implies that the rules used in the model are not visible to the user. These rules will be applied to the unknown protein in turn to determine a set of scores for the amino acids in a given position within the sequence when compared with a target sequence e.g. a protein in the data bank. These scores are then compared to the proteins in the reference database of proteins used and a series of best matches found.

In order to do this, the program first uses *PSI-BLAST*[206] to identify library sequences that are suitable homologs. From this multiple alignment a profile HMM is created for this sequence, which is a concise statistical description of the likelihood of each possible residue at a particular sequence position, allowing for secondary structure, insertions and deletions. This profile is then compared to pre-calculated database HMMs, and returns a list of matches from the database that might be useful for the sequence in question, i.e. providing models that might be suitable for molecular replacement, along with two scores (measures of success) for each. The first is probability (in percentage) that the solution given is a true match to the target model. The second is an E-value, defined as the expected number of false positives per database search that return a score at least as high as this particular sequence match.

This was done with the CRLF3 protein. The results (Figure 2-2) seemed to show quite clearly that the protein can be characterised as comprising of two regions.

The first region, containing the first 160 amino acids, seemed to have very few matches in the database. The matches were regions containing many coiled coils, which are known to be difficult to crystallise. The matches seemed to be proteins involved in ubiquitination. However, given there were only two matches this was an uncertain result (e.g. 4NQJ_C[207] – probability 97.99 %, E-value 7.5E-6, 13 % sequence identical).

The second region had plenty of matches (e.g. 3T1W[208] – probability 98.32 %, E-value 3.5E-7, 17 % identical). This region matched structures with a fibronectin type domain, and there were a variety of functions, from cell adhesion to being a hydrolase. All of these proteins were primarily made up of β-sheets.

These two different regions started to inform the choice of constructs. It suggested that cutting the protein at the juncture between these two regions might lead to some useful construct designs.

Next, the secondary structure of the protein was predicted using *PSIPRED*[60,209,210] (Figure 2-3). *PSIPRED* predicts the secondary structure of the protein based solely on its sequence. The first step is to use *PSI-BLAST* to create a position-specific scoring matrix (profile) of the given sequence. This is then used to predict the secondary structure using a neural network, in overlapping blocks of 15 amino acids, combined with information on the position of this sub-sequence within the chain. A second neural network filters the output from the first network to indicate if each residue is in a helix, sheet or coil.

The results from *PSIPRED* (Figure 2-3) suggested that the first part of the protein contains a number of helices, and there is high confidence in this prediction. It is in agreement with the *HHPRED* results that matched up that part of the protein to coiled coils. The second part of the protein (after approximately residue 160) contains primarily β-sheets, with the helical parts of the protein given less confidence in the prediction. This is useful for construct design, as coiled coils tend to be highly dynamic and tend to crystallise poorly[211]. Thus the first part of the protein might be hindering the crystallisation.

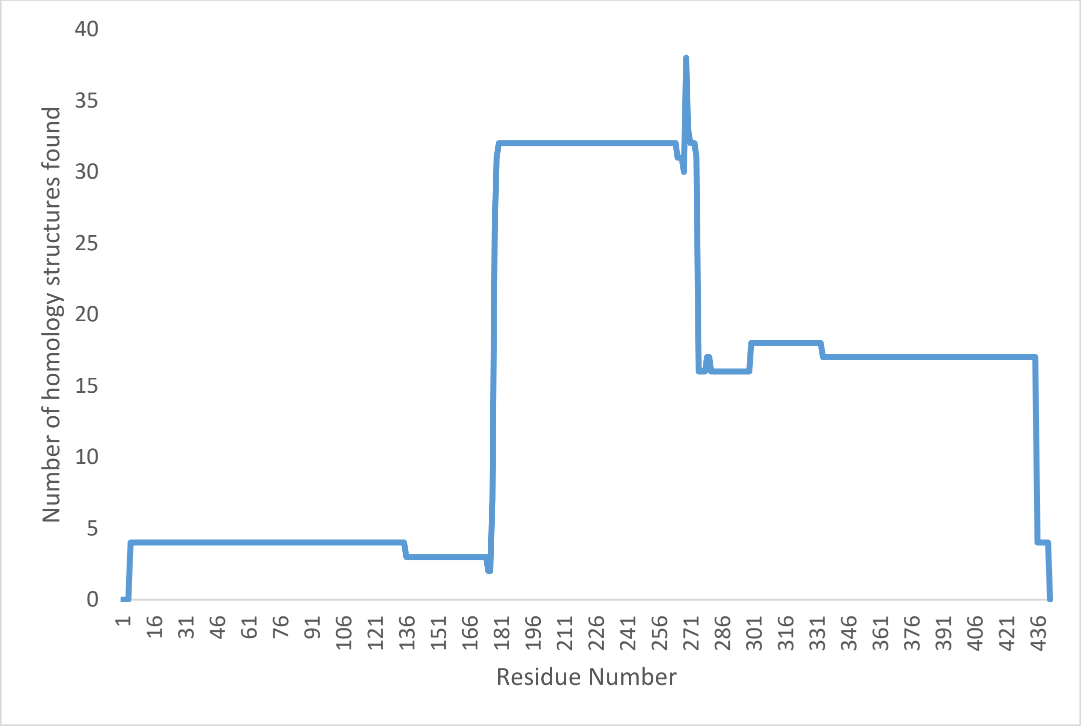


Figure 2-2 HHPRED results for CRLF3, showing the number of homology matches for the different section of the sequence

HHPRED[204,205] does a homology search on the sequence and here graphically shows where the matches are with the sequence. As can be seen, there are very few matches for the first 171 residues, a large number of matches for the middle section of the protein, and a reasonable number of homologous structures in the end section (output re-plotted using *Microsoft Excel*).

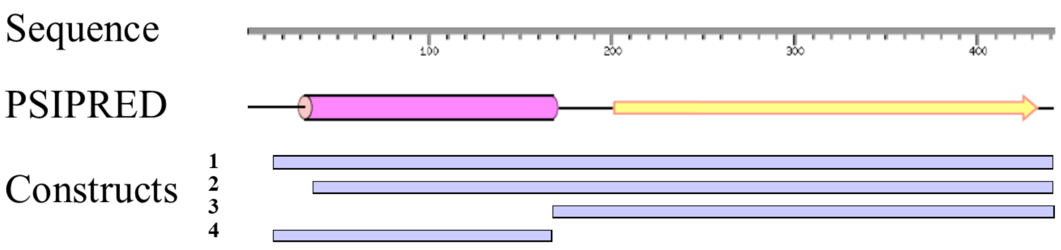


Figure 2-3 A summary of various different measures used to inform the choice of constructs

PSIPRED[60,209,210] predicts the secondary structure based on the sequence provided (simplified plot shown in this diagram). The purple cylinder indicates the likely presence of helices, with the yellow arrow indicating the likely presence of sheets. Four constructs (shown by the blue bars) were created with the following residues from the 442 amino acid protein used respectively, 16-442, 38-442, 174-442, 16-173.

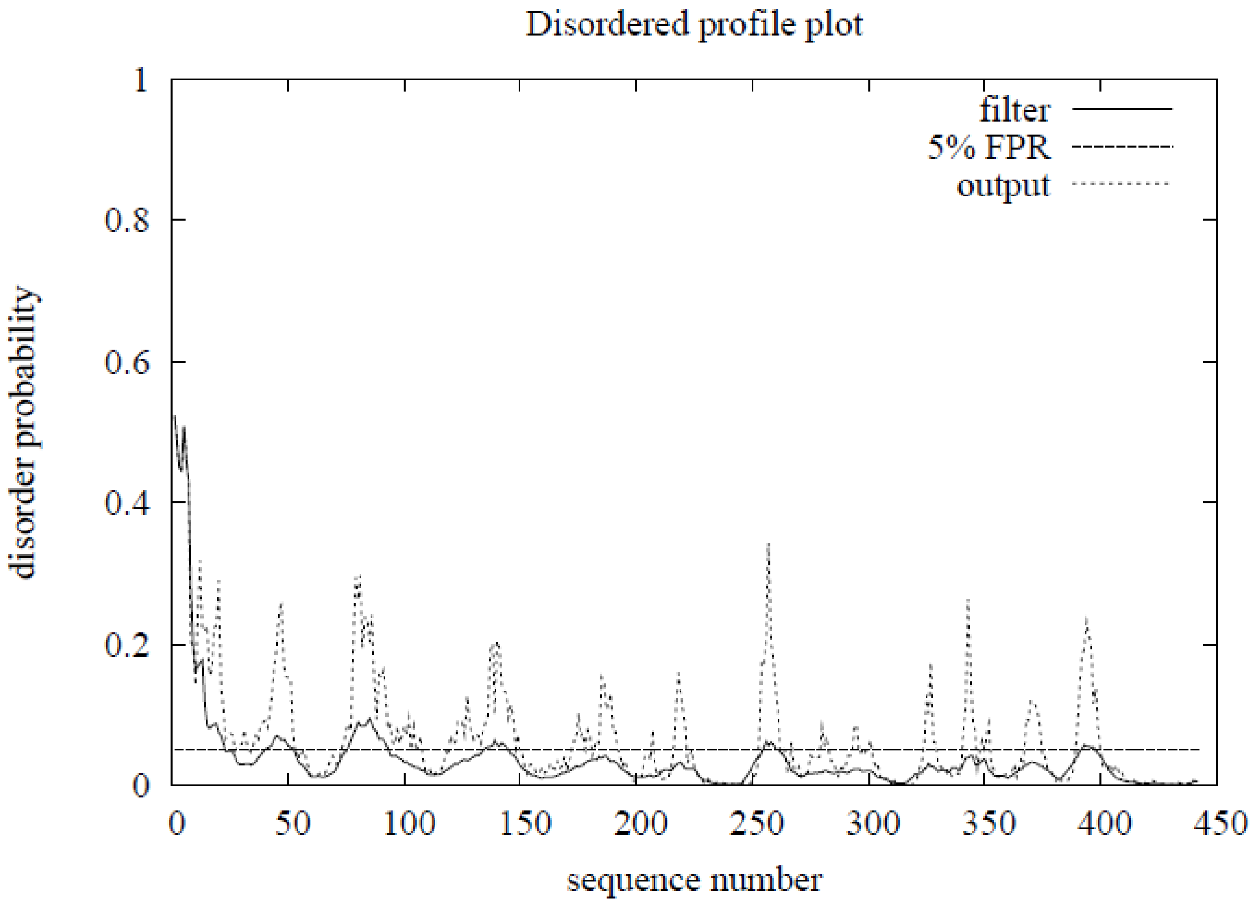


Figure 2-4 DISOPRED2 output for CRLF3

DISOPRED2[212] has been used to create the disordered profile plot. This indicates the likelihood that the given residue within a sequence is in a disordered region of a protein, which would indicate difficulty in crystallisation. 5% FPR stands for false positive rate being set to 5 %.

Predictions from the *DISOPRED2* server[212] were reviewed. This program has recently been updated, and the equivalent result is included (Figure 2-4). These identify parts of the protein which are likely to be natively disordered in the structure. This particular server uses a type of machine learning and had been trained on a test set of 715 protein chains where the resolution was better than 2 Å. In this test set a residue was deemed to be disordered if it was not modelled in the map, i.e. if residues exist in the sequence but not in the coordinate list. The server examines the protein sequence to identify regions that are similar to disordered regions in the test set.

This result showed that the first fifteen residues are likely to be disordered. This is helpful as disordered parts of the protein tend to hinder crystallisation[212]. Therefore, in the design of the construct, removing the disordered parts of the protein will make it more likely that the construct will crystallise.

The next measure examined was how likely a given protein sequence (CRLF3 or construct) is to crystallise. To determine this, the *XtalPred* server[213–216] was used. This looks at a variety of predictors created by a range of developers, which it combines into a single score indicating how likely the protein is to crystallise (Table 2-1). These predictors include *PSIPRED*[60,209,210] and the *DISOPRED2* server[212], in addition to other measures such as surface ruggedness[217] (using the relative surface accessibility calculations from the NetSurfP server[218]), gravy index[219], instability index[220] and predicted transmembrane helices[215]. The *XtalPred*[213–216] server combines these measures to create a prediction of crystallisation success. *XtalPred*[213–216] returns two crystallisation likelihood values, one based on an older method (the expert pool method, which uses eight protein features), and a second from a more modern method (the random forest (RF) algorithm, which uses twelve features and a Random Forest classifier). Both are still considered to have value, though the latter claims an almost two-fold improvement in prediction success; the former gives a score out of 5, and the newer out of 11. Higher number classes indicate a poorer success rate in crystallisation.

Predictor	Summary description
Gravy Index[219]	The sum of the hydropathy values of all the amino acids divided by the number of residues in the sequence, where hydropathy is a measure of the hydrophobic(+)/hydrophilic(-) nature of the side chain.
Instability Index[220]	This predicts instability by taking pairs of amino acids in the sequence and determining if these are more common in known unstable proteins.
Predicted transmembrane helices[215]	Parts of a protein are categorised into different regions, and uses hydrophobicity, charge bias, helix lengths and grammatical constraints, with a HMM to predict transmembrane helices.
Insertion Score	In the multiple sequence alignment, the percentage of gaps existing in homologous structures.
NetSurfP[217]	Predicts the surface accessibility (the accessibility of a particular amino-acid in the sequence to solvents), and forms a Z-score based on the mean of the surface accessibility.
PSIPRED[60,209,210]	(covered in main text)
DISOPRED2[212]	(covered in main text)

Table 2-1: List of parameters from XtalPred

The *XtalPred*[213–216] server was used to evaluate the full CRLF3 sequence, which provided more useful information for construct design (Figure 2-3). It showed clearly a prediction for coiled coils in the first part of the protein. The chances for crystallisation were assessed as poor. The old expert pool measure determined that CRLF3 was in the lowest chance class of getting good crystallisation (class 5 out of 5). The newer RF algorithm method was more optimistic (class 5 out of 11).

This analysis then allowed the choice of constructs to be made. It was clear that the protein was separated into two regions in the homology searches. It was also apparent that the first fifteen residues of the protein had a high probability of disorder. Therefore four constructs were designed (Figure 2-3). Construct 1 was the full length protein with just the first fifteen residues removed (residues 16-442) as it was felt that this might be enough to allow the protein to crystallise. Construct 2 (residues 38-442) removed more of the disordered region at the start, though there was a fear it might stop the first region from folding. The remaining two constructs were designed to crystallise each region of the protein separately. Construct 3 (residues 174-442) was the region covered by most of the homology models. It is also advantageous that this region did not have any of the disordered regions or those that were likely to form coiled coils. Construct 4 (residues 16-173) was the interesting region in that there is very little homology to it, but with the prediction of coiled coils was likely to be difficult to crystallise.

To determine whether any of these constructs were more likely to crystallise, they were run through *XtalPred*[213–216] server (Table 2-2). The expert pool older algorithm predicted that all of them were more likely to crystallise than the native protein, putting all the constructs in the same class (class 4 out of 5), but this is still a poor score. In contrast the Random Forest method produced far more varied results, which to some extent disagreed with the expert pool. For construct 1, the removal of the first fifteen residues improved the crystallography class (up to class 4 / 11 from the original class 5 / 11). Intriguingly, as more residues were removed from the N-terminus, the estimated chances of crystallisation actually decreased back again to those of the native sequence (class 5 out of 11).

	EP-Class	RF-class	Length	Gravy Index	Instability Index (II)	Isoelectric point (pI)	Coiled Coils
Construct1	4	4	427	-0.38	46.85	5.12	31
Construct2	4	5	405	-0.35	46.29	5.03	0
Construct3	4	3	269	-0.33	47.52	5.17	0
Construct4	4	11	158	-0.48	45.63	5.07	31
Native	5	5	442	-0.38	47.04	5.01	45

	Longest disorder region	Percentage of coil structure	Trans-membrane helices (TM)	Signal peptides (SP)	Insertion score	Homologs in NR (clustered to 60% seq. ident.)	Homologs in PDB
Construct1	17	41	No	No	0.15	484	57
Construct2	20	42	No	No	0.13	416	63
Construct3	4	49	No	No	0.14	273	31
Construct4	19	19	No	No	0.01	11	0
Native	33	39	No	No	0.14	496	60

Table 2-2: The abbreviated output from XtalPred server

This shows the differences in the fragments clearly. Using the RF algorithm, fragment 3 is the most likely to crystallise, with fragment 4 the least likely. The EP-Class and RF-class colouring follows the colour schemes used for those classes in the *XtalPred* server. In other columns, if a value leads to particularly poor prognosis for crystallisation, it is highlighted.

Considering the other constructs, it is found that construct 3 is predicted to have a very good chance of crystallising (class 3 out of 11), and looks like a good candidate. This is perhaps unsurprising as all the regions of disorder, and regions that are likely to cause coiled coils, have been removed. Finally, construct 4 is predicted as highly unlikely to crystallise, being in the bottom class (class 11 out of 11).

2.6 Stage two: First attempts to obtain a crystal structure

2.6.1 CRLF3 Protein Preparation

The production of the cytokine receptor like factor-3 (CRLF3) protein constructs was undertaken by Dr. Cavan Bennett in Dr. Cedric Ghevaert's group at the Department of Haematology in the NHS Blood and Transplant Centre at the University of Cambridge. The crystallisation process was undertaken by Dr Yahui Yan, Cambridge Institute for Medical Research, University of Cambridge, observed by the author to familiarise himself with the overall protein structure determination process. Details of their work are described here for completeness.

The CRLF3 constructs were first amplified through PCR (using the Platinum High Fidelity Taq Polymerase (Invitrogen) along with appropriate primers) of *in vitro* cultured murine MK cDNA. The PCR products were isolated through the use of a 1.5 % agarose TAE (40mM Tris, 20mM acetic acid, 1mM EDTA) gel. The specific band for this DNA was extracted using a QIAquick Gel Extraction Kit (QIAGEN). The DNA was cleaned using a QIAquick PCR Purification Kit (QIAGEN), and then digested using BamHI and NotI (New England Biolabs).

These constructs were then inserted into a pGEX-6P-2 vector (GE Healthcare) by digesting the vector with the same restriction enzymes as above (BamHI and NotI), and then ligating the plasmid to the constructs with the T4 Ligase (New England Biolabs). This vector with the constructs was then placed by heat-shock into BL21-CodonPlus^{TM}-(DE3)-RP competent *E. coli* cells (Stratagene), allowing the production of CRLF3 protein covalently linked to glutathione-*S*-transferase (GST). This later allows for the CRLF3 protein and its constructs to be purified.

These BL21 cells with the vector were then grown in 2xTY media at 37 °C, and were induced (after reaching appropriate cell density ($OD_{600} \approx 0.8$)) to start producing protein with 1mM isopropul-1-thio-β-D-galactopyranoside (IPTG) (Sigma-Aldrich). The cells were left at 30 °C for at least five hours. This method was consistent with previous literature[221].

The protein was then extracted from pelleted cells using a mixture of 1 mg ml^{-1} lysozyme (Sigma-Aldrich), 10 mM $MgCl_2$, 10 U ml^{-1} DNase I (Sigma-Aldrich) and one tablet of Protease Inhibitor Cocktail (Roche).

To purify the protein it was separated from the major fragments of the lysed cell by spinning at 10,000 rpm (40 min, 4 °C), with the supernatant stored at -80 °C until it was needed. The CRLF3 protein or the construct was purified by capturing the Glutathione-*S*-transferase (GST fusion protein) through affinity chromatography. GSTrap FF 5 ml columns were used, allowing all non-GST tagged proteins (i.e. all other protein) to be washed away; CRLF3 protein or construct was then removed by incubating the resin with PreScission Protease (GE Healthcare) for 16 hours at 5 °C. It was then run through size-exclusion column chromatography (HiLoad 16/600 Superdex 75pg column (GE Healthcare)) (Figure 2-5). The protein was concentrated (VivaSpin 2 column with a 10 kDa cut off (Satorius Stedim Biotech)) in a buffer (10 mM Tris pH 7.4, 150 mM NaCl), and the concentration was measured through the use of a Nanodrop (Thermo Scientific) or bicinchoninic acid (BCA) assay using Pierce® BCA Protein Assay Kit (Thermo Scientific).

Unfortunately, attempts to manufacture constructs 1, 2 and 4 and more native protein resulted in poor yield (0 mg, 0.7 mg, 0.8 mg and 0.6 mg for 1l of bacterial cells respectively) so crystallisation trials were not attempted for these constructs. However, construct 3 produced a high yield (7.0 mg) in 1l of media, so it was concentrated (11.7 mg ml^{-1}) in 10mM Tris pH 7.4, 150mM NaCl and taken on to crystallisation trials.

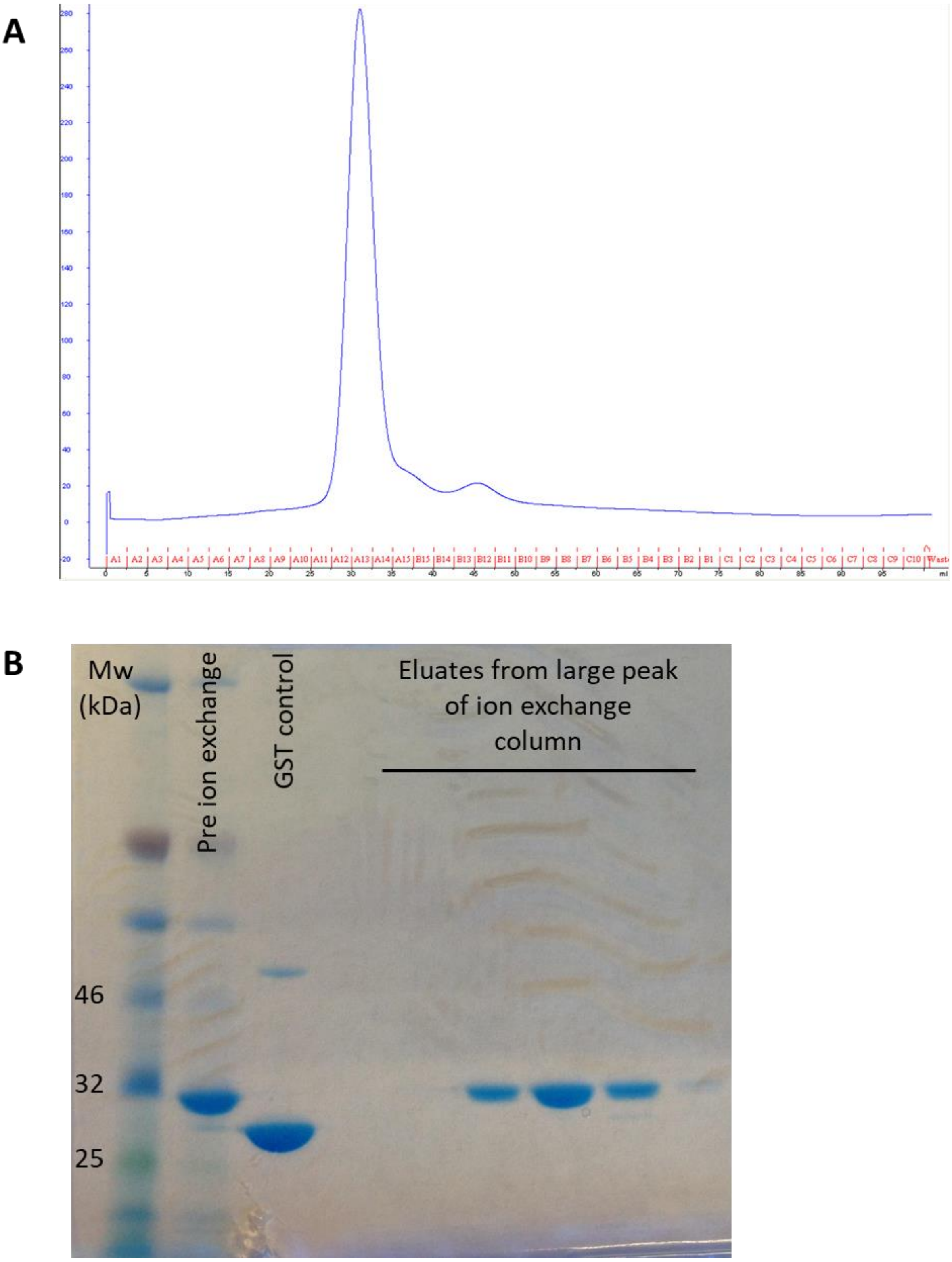


Figure 2-5: CRLF3 protein purification down an ion exchange column

(A) The blue trace shows the construct 3 protein eluting cleanly from the column, with little other protein emerging. (B) SDS-PAGE confirms this observation of pure protein from the column fraction, with Coomassie stain showing the protein present. (Image taken and labelled by Dr. Cavan Bennett).

2.6.2 CRLF3 Crystallisation

Crystallisation screens were done by using a vapour diffusion method. A Nanodrop Screenmaker 96+8 (Innovadyne Technologies) was used to set up these plates, with the drop consisting of 200 nl protein (or construct) together with 200 nl screening solution. Vapour diffusion crystallisation was set up on these 96-well plates with a 70 µl screening solution reservoir. The plates were maintained at 20 °C and imaged with the Rock Imager Crystal Screening System (Formulatrix). These screens had been done for native protein in previous work, but little success was observed. For construct 3, crystals were obtained under many conditions. The following screens were used for these crystallisation trials. PEG/Ion Screen, Peg/Ion 2 Screen (Hampton Research), Wizard 1, Wizard 2, Wizard 3, Wizard 4, Cryo I, Cryo II (Emerald BioSystems), Structure Screen 1, Structure Screen 2, PACT Premier, Stura Footprint Screen, MacroSol and JCSG-Plus (Molecular Dimensions).The following crystallisation conditions were successful (Table 2-3):

	Successful crystallisation conditions
1	12 % PEG 3,350, 0.1 M sodium acetate
2	20 % PEG 3,350, 0.2 M sodium formate pH 7.0
3	25 % Ethylene glycol
4	25 % PEG 3,350, 0.2 M ammonium acetate, 0.1 M BIS-TRIS pH 5.5
5	0.14 M calcium chloride, 30 % glycerol, 14 % 2-propanol, 0.07 M sodium acetate
6	23 % PEG 3,350, 0.2 M sodium formate
7	22 % PEG 3,350, 0.2 M sodium formate
8	9 % 2-propanol, 0.1 M sodium cacodylate, 0.2 M zinc acetate, pH 6.5
9	20 % PEG 3,350, 0.2 M sodium acetate pH 7
10	12 % PEG 3,350, 0.2 M sodium acetate pH 7
11	6-16 % PEG 3,350, 0.01 M sodium acetate pH 7
12	2 % 1,4-Dioxane, 10 % PEG 20,000, 0.1 M BICINE pH 9
13	20 % PEG 8,000, 0.2 M calcium acetate 0.1 M MES, pH 6
14	10 % 2-propanol, 0.1 M sodium cacodylate, pH 6.5, 0.2 M zinc acetate

Table 2-3: List of successful crystallisation conditions for CRLF3 and its constructs

Crystals were picked from the successful trials and cryoprotected (25 % ethylene glycol diluted in the crystallisation solution), and were stored at approximately 77 K.

2.6.3 CRLF3 X-ray diffraction experiments

Remote data collection was then undertaken at Diamond Light Source on the I02 beamline with a Pilatus 6MF (Pixel Array Detector)[222], and crystals were kept cool by the localised flow of cold nitrogen gas at approximately 100 K. We thank Diamond Light Source for access to beamline I02 (proposal number MX8547-174 - first visit) that contributed to the results presented here. This beamline allowed varying wavelengths to be used (0.9795 Å for native sets and 1.70007 Å for S-SAD). However, it has the limitation that the goniometer was only single axis, which prevented ideal data collection strategies in S-SAD.

The data recorded were indexed and scaled using the *xia2* pipeline[223]. For each crystal that was exposed, a grid scan was first undertaken to determine the best beam location. Three exposures were then taken at widely spaced angular orientations to determine the optimal data collection strategy. For a native set, one set of data was then collected. For SAD, a peak wavelength data set was collected.

xia2 is a pipeline which permits the preparation of data for molecular replacement or experimental phasing. It attempts to obtain a value for the intensities/amplitudes for all Miller indices that can be seen on the diffraction pattern, and so reducing the information from a huge number of pictures to a single reflections file. Thus *xia2* provides an automatic way to extract the data from images, allowing the indexing of the reflections on an image, and merging of the data from the different reflections. *xia2* allows the use of a variety of packages to do the data reduction, including *HKL/HKL-2000*[224], *MOSFLM/SCALA*[103], *XDS/XSCALE*[102] and *d*trek*[225].

The two datasets that were taken forward were both taken from crystals obtained in crystallisation conditions of 12 % PEG 3,350, 0.1 M sodium acetate and cryo-protected in 25 % EG. These are labelled later as CRLF3 construct 3 original native and CRLF3 construct 3 final model 2.

2.6.4 Initial attempts to solve CRLF3 structure

2.6.4.1 Attempts with the CRLF3 construct 3 original native dataset

Initially, one component molecular replacement was undertaken using *phaser*[4] in the *Phenix*[226] package using the native dataset (for statistics, see Section 2.9). In order to get a starting model, *HHPRED* was re-run using just the sequence for construct 3 so that the best possible model for molecular replacement could be found. The initial starting model selected was 1UC6[227], and molecular replacement was not successful with that search model.

An improved starting model was then created using *Sculptor*[228] and *Ensembler*[228] and the following models (Table 2-4). First, the top hits from *HHPRED* were taken and merged together to form an ensemble using *Ensembler*. This program superimposes multiple models so that they can act as an ensemble (multiple starting models) for molecular replacement. *Sculptor*[228] was then used to trim any loops where the models did not match, as these are the regions where there is the least agreement and it has been found that better results are obtained when they are excluded. This new model was then used for molecular replacement but still no solution was found.

PDB ID	Chain	Protein Name	Resolution Å	Function
1UC6[227]	A	C-terminus domain of the ciliary neurotrophic factor	NMR – RMSD is 0.37 ± 0.04 Å for backbone atoms.	A cytokine that enhances the survival of neuronal cells, and a major protective factor in demyelinating central nervous system disease. To exert its biological functions, CNTF first binds its non-signalling, specific receptor.

1BQU[229]	A	Cytokine-binding region of GP130	2	Transmembrane receptor, required for interactions with cytokines. Implicated in adult tissue systems, including haemopoesis, nervous system, bone, heart, adipose tissue, testes, liver and muscle, and continuous activation leads to cardiac hypertrophy.
1CD9[230]	B	GCSF-receptor complex	2.8	Cytokine, and is a growth factor regulating maturation, proliferation and differentiation of precursor cells.
1I1R[231]	A	Extracellular gp130 cytokine receptor signalling complex	2.4	This is a shared signal-transducing receptor for a family of cytokines which share a common four-helix bundle fold.
2B5I[232]	C	quaternary complex of interleukin-2 with its alpha, beta, and gamma receptors	2.3	Interleukin-2 is a cytokine promoting proliferation, differentiation and survival of T-cells, and promoting cytolitic activity of natural killer cells.
2E3V (no publication)	A	first fibronectin type III domain of neural cell adhesion molecule splicing isoform from human muscle culture lambda-4.4	1.95	Cell to cell adhesion, as well as communication between different neuronal cells.

2HAZ[233]	A	first fibronectin domain of human NCAM1 (neural cell adhesion molecule)	1.7	Axon guidance and pathfinding, neurite outgrowth, synaptic plasticity and cell migration in the central nervous system.
2IC2[234]	A	First fibronectin III (FN3) Domain of Ihog (interference hedgehog)	1.3	Involved in signalling pathways that mediate key tissue-patterning effects during animal development.
2B83[235]	A	Mutated Clostridium beijerinckii alcohol dehydrogenase	2.25	Enzymes involved in the interconversion between alcohols and aldehydes or ketones.
3N1F[236]	C	Signalling domain of Indian Hedgehog IhhN bound to membrane-proximal type III fibronectin of the cysteine dioxygenase (CDOFn3)	1.6	Hedgehog is a signalling protein mediating key cell differentiation and tissue patterning events during animal development, and cysteine dioxygenase interacts with this in a calcium dependant way.
3T04[237]	D	Monobody 7c12/abl1 Src Homology 2 (sh2) domain complex	2.1	Functionally a kinase involved in signalling and implicated in chronic myelogenous leukaemia.

3V6O[238]	A	Leptin Receptor-antibody complex	1.95	Leptin is involved in pathways regulating energy management, fertility and the immune system. This receptor is on those pathways.
3WIH[239]	A	Third fibronectin domain (FN3) of human ROBO1 in complex with the Fab fragment of murine monoclonal antibody B2212A	1.701	Axon guidance receptor, and is a member of a neural cell adhesion molecule family.
4LSD[240]	A	Myokine structure	2.28	Cytokine released by muscle cells in response to muscular contractions.
4O00 (no publication)	A	Titin A-band domain A3	1.853	Giant protein, responsible for the passive elasticity of muscle.
4S0S[241]	D	Human pregnane X receptor ligand binding domain with Adnextin-1	2.8	Nuclear receptor to foreign substances which are not recognised by the body. This receptor goes on to interact with steroid receptor co-activator-1 (SRC-1) to exert its effect.
4U3H[241]	A	Fibronectin type III domain consensus protein (FN3con)	1.98	Artificial protein designed to look at protein stability. Consensus of all fibronectin type III domain.
4WTW[242]	A	Third Fibronectin type III domain (FN3) domain of integrin beta4	1.606	Involved in stable anchorage of epithelial cell to underlying basal membrane.

4WTX[242]	A	Fourth Fibronectin type III domain (FN3) domain of integrin beta4	1.5	Involved in stable anchorage of epithelial cell to underlying basal membrane.

Table 2-4: List of homologues similar to construct 3 as found by HHPRED

phenix.MRage[243] was then used to try to carry out automated molecular replacement. Here the results from *HHPRED* were directly fed to the program, which now acts as an automatic pipeline. It makes decisions based on what combination of input models is most likely to create the best possible model and the best possible molecular replacement results. It separates out all the steps in molecular replacement being run in *phaser*[4] so that it can allocate computing resources appropriately and adequately follow the different branches. However, this also did not give a clear result.

The results from the *HHPRED* server were used in *phenix.MR_Rosetta*[244]. This takes the homology models, along with 3-residue and 9-residue fragments for loop-building that are appropriate for the CRLF3 protein. The homology models are optimised using *Rosetta* without taking into account the X-ray diffraction pattern to produce the best model possible for molecular replacement within *phaser*[4]. Molecular replacement in *phaser*[4] is then undertaken, and the refined density-modified electron density maps fed back into *Rosetta* to rebuild the model. A final step is to rebuild the model in *phenix.autobuild*. However, these trials also failed to produce a solution.

Use of a protein prediction server, called *I-TASSER*[245–248], was then attempted. This server takes the protein sequence and uses multiple threading alignments and iterative structural assembly simulations to build the best model possible. It has regularly performed well in Critical Assessment of protein Structure Prediction (CASP) experiments[249] and so the five output models that were generated were fed into *phaser*[4]. They were also put into the *phaser.MRage* pipeline[243], as there seemed to be a conserved core in the five models, and with this pipeline the non-similar part of the five models is pruned away. Unfortunately, all these attempts also failed to lead to a correct solution for CRLF3.

2.6.4.2 Attempts with the S-SAD dataset

With the failure of molecular replacement using the native dataset, structure solution attempts were also made using a S-SAD dataset (for statistics, see Section 2.9, referred to as CRLF3 construct 3 final model 2). An anomalous signal exists at the wavelength used to record these measurements. It was hoped this would be sufficient to determine the location of the sulphur atoms (the S substructure). The first attempts used an automated pipeline *phenix.autosol*[250]. This starts by checking the data using *phenix.xtriage*[251], then moves on to obtaining the substructure of S atoms using *HySS*[252], phasing with *phaser*[4], statistical density modification and model building with *RESOLVE*[161,253], before a final evaluation of map quality using real-space measures is made. Unfortunately, this too failed to produce a solution.

An exhaustive search was then carried out to find the S substructure using *HySS*[252], exploiting the full range of features available in the program. This did produce a more complete S substructure, though with not a particularly high figure of merit score, suggesting that it was not completely correct. Attempting to obtain the phases using this substructure in *phaser*[4] did not yield a successful solution.

Further efforts were made to improve the S-SAD data to yield better phases, including re-merging the data using *XDS* and *XSCALE*[102,254]. This included attempts to leave out parts of the dataset that showed signs of radiation damage, and also attempting zero dose extrapolation[255]. However, these attempts ultimately proved to be unproductive and a solution was not found.

2.6.4.3 Result of initial attempts using the initial datasets

All the attempts to try and resolve the structure from the data already recorded failed, so other ways of doing experimental phasing were deemed necessary. However, the data collected seemed to be of relatively good quality, with high resolution and good completeness. In particular, the native set seemed to be of very good quality and so it was very likely that the models for molecular replacement were too poor to be used. The S-SAD dataset was recorded close to the optimal wavelength, but the goniometer limitation to a single axis of rotation made it difficult to collect truly multiplicitous data. Furthermore, there was a slow degradation in the signal which is ascribed to radiation damage, hence the attempts to prune the data. Despite the evidence of radiation damage in the sample, the data collected were still of an acceptable quality for S-SAD structure determination.

The S-SAD dataset was not ideal. It exhibited low completeness of around 84% (see Section 2.9 for data statistics) due to the limited rotations available on the beam line.

2.7 Stage three: Further attempts to solve the crystal structure

The limitations evident with the data achieved using a single-axis goniometer led to the decision to use a line on the Diamond Light Source that offered three-axis experimentation. Crystals were still available, and a number were subjected to a heavy atom soak to allow the experiments to be extended to the use of SAD, based on these heavy atoms. Additional native sets were to be acquired if possible.

2.7.1 CRLF3 Protein Preparation

Selenomethionine was not used as it was felt that protein production would be even more difficult for a version of this construct (3) containing the altered amino acid. This fragment only contained a single methionine, and so whilst the substitution will provide some signal (up to 6 % as predicted in the optimistic case on Ethan Merritt's website[256]), it is limited. There was also little experience with this amino acid in the laboratory, and it was felt that other methods would be successful and should be attempted first. As there was limited time available from the collaborator, the more immediately available option was taken. If the protein structure had not been solved, this approach would have been attempted.

2.7.2 CRLF3 Crystallisation

Heavy metal incorporation into a second batch of crystals prepared as described earlier (example of crystals seen has already been shown (Figure 1-5)) was achieved by soaking in different heavy metal solutions for different amounts of time (Table 2-5). Mercury and iodine were selected as the anomalous scatterers for this experiment, as they are known to give good anomalous signal, and they have frequently been used successfully.

Serial	Condition	Heavy Atom	HA soaking time	Cryo-protectant	data collection
1	20 % PEG 3,350, 0.2 M sodium formate pH 7.0	10 mM thimerosal	overnight	25% EG	Hg-SAD/MAD
		10 mM thimerosal	4h	25% EG	no data
2	25 % Ethylene glycol	N.A	N.A		native and S-SAD
		0.5 KI	2-30min	25% EG	I-SAD
3	25 % PEG 3,350, 0.2 M ammonium acetate, 0.1 M BIS-TRIS pH 5.5	N.A	N.A	PFPE	native
4	23 % PEG 3,350, 0.2 M sodium formate	10 mM thimerosal	50 min	25 % EG	Hg-SAD
		10 mM thimerosal	4 h	25 % EG	Hg-SAD/MAD
		10 mM thimerosal	Overnight	25 % EG	Hg-SAD
5	22 % PEG 3,350, 0.2 M sodium formate	10 mM $Hg(OAc)_2$	overnight	25 % EG	poor diffraction
6	9 % 2-propanol, 0.1 M sodium cacodylate, 0.2 M zinc acetate, pH 6.5	20 mM $Hg(OAc)_2$	6 days	PFPE	poor diffraction
7	20 % PEG 3,350, 0.2 M sodium acetate pH 7	0.5 KI	5-20 min	25 % EG	poor diffraction
8	12 % PEG 3,350, 0.2 M sodium acetate pH 7	N.A.	N.A	25 % EG	native and S-SAD
9	6-16 % PEG3350, 0.01 M sodium acetate, pH 7	20 mM thimerosal	6 days	25 % EG	poor diffraction
		20 mM thimerosal	6 days	25 % EG	poor diffraction
		5mM $HgCl_2$	6 days	25 % EG	poor diffraction
		10mM $Hg(OAc)_2$	6 days	25 % EG	poor diffraction
		10mM thimerosal	6 days	25 % EG	poor diffraction
		10mM thimerosal	overnight	25 % EG	poor diffraction

10	2 % 1,4-Dioxane, 10 % PEG 20,000, 0.1 M BICINE pH 9	N.A	N.A	PFPE	poor diffraction
11	20 % PEG 8,000, 0.2 M calcium acetate 0.1 M MES, pH 6	N.A	N.A	PFPE	poor diffraction
12	10% 2-propanol, 0.1 M sodium cacodylate, pH 6.5, 0.2 M zinc acetate	20 mM $Hg(OAc)_2$	overnight	PFPE	poor diffraction

Table 2-5: Details of the results of heavy metal soaks and subsequent diffraction performance

where

EG – Ethylene Glycol
MES - 2-(N-morpholino)ethanesulfonic acid
PEG – polyethylene glycol
PFPE – perfluoropolyether oil
N.A. – Not Applicable (as no soak undertaken with these particular crystals)

2.7.3 CRLF3 X-ray diffraction experiments

Further remote data collection was then undertaken at Diamond Light Source on the I04 beamline with a Pilatus 6MF (Pixel Array Detector)[222], and crystals were kept cool by the localised flow of cold nitrogen gas (approximately 100 K). We thank Diamond Light Source for access to beamline I04 (proposal number MX11235-49 – second visit) that contributed to the results presented here. This beamline allowed varying wavelengths to be used (0.9795 Å for native sets, 1.0060 Å for Hg-SAD and 1.70007 Å for S-SAD). This line incorporates a Mini-Kappa multi-axis goniometer, which allows the axes of rotation for effective collection strategies in SAD. Datasets were recorded for the native crystals, crystals that had been soaked in mercury, and those soaked in iodine (Table 2-5). A similar procedure to that described previously was used (Section 2.6.3). The data recorded were then indexed and scaled using the *xia2* pipeline[223], as before, with various collection strategies (Table 2-6) for the different crystals (Table 2-7).

The Hg SAD dataset that lead to phasing was obtained from crystals grown in 20 % PEG 3,350, 0.2 M sodium formate pH 7.0, which had then been soaked with 10 mM thimerosal overnight and then cryo-protected in 25 % EG. The dataset that led to CRFL3 construct 3 final model 1 was crystallised in 25 % PEG 8,000, 0.2 M calcium acetate, 0.1 M MES and cryo-protected in PFPE.

2.7.4 Further attempts to solve CRLF3 structure

The datasets from the SAD experiments were placed into *phenix.autosol* with a similar flow as described above for the SAD cases, although the phasing is done with *SOLVE*[257–260]. This program generates a first estimate of phases, and then proceeds to generate a trial model structure (Table 2-8). This was then refined by using a combination of *phenix.refine*[167] and manual model building with *Coot*[23]. The mercury-soaked crystals provided the break-through needed to progress this project, as this program successfully created a model from the data generated using these crystals. The iodide crystal data set, however, did not yield results using this analysis. Results gave a poor anomalous signal, so it is likely that insufficient iodine was incorporated in the crystals under the soak conditions selected. Later analysis revealed the presence of three iodide sites in the crystal structure, but these did not yield sufficient signal. Further S-SAD experiments were again inconclusive and were not taken forward.

At this point, it became possible to re-analyse all available results from both sets of diffraction experiments using this initial set of phases and model in order to create a set of models for further consideration.

2.8 Stage four: Generation of structural models

As noted in Section 2.7.4, structure determination with the Hg-SAD dataset was successful and allowed a model to be built. It was then used as the initial model for molecular replacement for all other datasets, leading to the solution of the structure. The final structure was obtained (Figure 2-6) and carefully refined (Table 2-9). Rigid-body refinement could also have been attempted, but would have given essentially the same answer.

From these initial solutions model building was achieved through *phenix.autobuild*[161]. Refinement was then done using a combination of *phenix.refine*[167] (3 cycles, with refinement strategy being the default (XYZ coordinates, Real-space, Occupancies, Individual B-factors)) and manual model building with *Coot*[23]. Structure validation was done performed with *MolProbity*[168,169] and various statistics including Ramachandran Plots (Figure 2-7), Polygon reports (Figure 2-8 and Figure 2-9), together with the secondary structure elements highlighted (Figure 2-10).

In all cases, there was one molecule found in the asymmetric unit. The crystals were approximately isomorphous.

2.9 Conclusions of structural determination research

Four models that have been discussed above were solved – from the first data collection the CRLF3 Construct 3 Original Native Model and the CRLF3 Construct 3 Final Model 2 (which was originally the model used for S-SAD experiments), and from the second data collection the CRLF3 Construct 3 Hg SAD Model which provided the necessary phase information to phase all models, along with the CRLF3 Construct 3 Final Model 1 (Table 2-8 and Table 2-9).

Two models were selected for further use. The first was from a data set generated on the second visit, taken at 0.9795 Å, which produced good statistics and a well-refined model (Table 2-9). The second was from a dataset generated on the first visit, which produced a model with good refinement statistics (Table 2-9). However, in the latter case the disulphide bridge was broken, indicating radiation damage (Figure 2-11 and Figure 2-12). The two domains of the protein are linked together by this disulphide bond. There is evidence of a slight shift between the two domains if the models for the two cases are compared (Figure 2-13). The first model is therefore assumed to be a more accurate representation of the construct 3.

	CRLF3 Construct 3 Final Model 1	CRLF3 Construct 3 Final Model 2	CRLF3 Construct 3 Hg SAD Model	CRLF3 Construct 3 Original Native Model
Ω Start	36.0°	186.0°	55.0°	159.0°
Ω Osc	0.20°	0.15°	0.50°	0.15°
Ω Overlap	0°	0°	0°	0°
No. Images	900	3600	720	1200
Max Resolution	1.59Å	1.90Å	1.80Å	1.60Å
Wavelength	0.9795Å	1.70007Å	1.0060Å	0.9795Å
Exposure	0.200s	0.040s	0.200s	0.100s
Transmission	20.00%	7.58%	9.99%	30.02%
Beamsize	43x30µm	43x30µm	43x30µm	70x28µm

Table 2-6: Data collection strategies for the different data sets

	CRLF3 Construct 3 Final Model 1	CRLF3 Construct 3 Final Model 2	CRLF3 Construct 3 Hg SAD Model	CRLF3 Construct 3 Original Native Model
Crystal appearance	Plate	Plate	Plate	Plate
Crystal length	360 µm	290 µm	250 µm	270 µm
Crystal width	210 µm	50 µm	190 µm	70 µm
Crystal thickness	10 µm	10 µm	10 µm	10 µm

Table 2-7: Size and appearance of crystals used in diffraction experiments

	CRLF3 Construct 3 Hg SAD Model	CRLF3 Construct 3 Original Native Model
Data collection		
Synchrotron station (DLS)	I04	I02
Wavelength, Å	1.006	0.9795
Space group	C 1 2 1	C 1 2 1
Cell dimensions		
a,b,c; Å	87.25, 40.94, 76.29	88.09, 40.81, 76.49
α, β, γ;°	90.00, 98.74, 90.00	90.00, 99.62, 90.00
Resolution, Å	27.31 - 1.97 (2.02 - 1.97)	75.42 - 1.7 (1.78 - 1.7)
Rmerge	0.067 (0.714)	0.056 (0.428)
Rmeas	0.079 (0.845)	0.078 (0.581)
<1/σ (I)>	16.6 (2.1)	6.8 (1.0)
CC1/2	0.999 (0.829)	0.998 (0.796)
Multiplicity	6.7 (6.7)	3 (2.9)
Completeness, %	99.7 (96.0)	98.8 (98.1)
Mosaicity, °	0.14	< 0.005
Figure of merit	0.354	-
Refinement		
Rwork	0.2752 (0.3300)	0.2046 (0.3293)
Rfree	0.2969 (0.3403)	0.2360 (0.3706)
No. of reflections	128603 (8993)	89090 (11095)
No. of unique reflections	19156 (1351)	29360 (3848)
No. of atoms	1702	2265
Average B-factors, Å^2	30.95	26.02
RMS deviations		
Bond lengths, Å	0.009	0.008
Bond angles,°	1.01	1.12
Ramachandran favoured region, %	96.43	97.98
Ramachandran outliers, %	0.00	0.00
MolProbity score	1.82	1.13

Table 2-8 Data collection and refinement statistics of the Hg-SAD data set and the native data set from the first diffraction experiments

The parentheses show the statistics for the high resolution shell.

	CRLF3 Construct 3 Final Model 1	CRLF3 Construct 3 Final Model 2
Data collection		
Synchrotron station (DLS)	I04	I02
Wavelength, Å	0.9795	1.7007
Space group	C 1 2 1	C 1 2 1
Cell dimensions		
a,b,c; Å	88.03, 40.38, 76.53	87.80, 40.60, 76.66
α, β, γ; °	90.00, 99.94, 90.00	90.00, 99.80, 90.00
Resolution, Å	37.69 - 1.61 (1.65 - 1.61)	32.15 - 1.74 (1.79 - 1.74)
Rmerge	0.045 (0.764)	0.057 (0.778)
Rmeas	0.063 (1.053)	0.064 (0.913)
<1/σ (I)>	13.3 (1.4)	20.0 (2.1)
CC1/2	0.999 (0.504)	0.999 (0.492)
Multiplicity	3.4 (3.6)	9.3 (6.4)
Completeness, %	96.7 (94.9)	88.4 (49.2)
Mosaicity, °	< 0.005	< 0.005
Refinement		
Rwork	0.1951 (0.3250)	0.1786 (0.3738)
Rfree	0.2163 (0.3464)	0.2023 (0.3943)
No. of reflections	114959 (8369)	227050 (6424)
No. of unique reflections	33326 (2356)	24310 (1000)
No. of atoms	2127	2177
Average B-factors, Å^2	33.14	35.73
RMS deviations		
Bond lengths, Å	0.009	0.013
Bond angles, °	1.06	1.08
Ramachandran favoured region, %	96.37	97.98
Ramachandran outliers, %	0.00	0.00
MolProbity score	1.53	1.15

Table 2-9: Statistics of the data collection and phasing of two fully refined models of CRLF3 Construct 3.

The parentheses show the statistics for the high resolution shell.

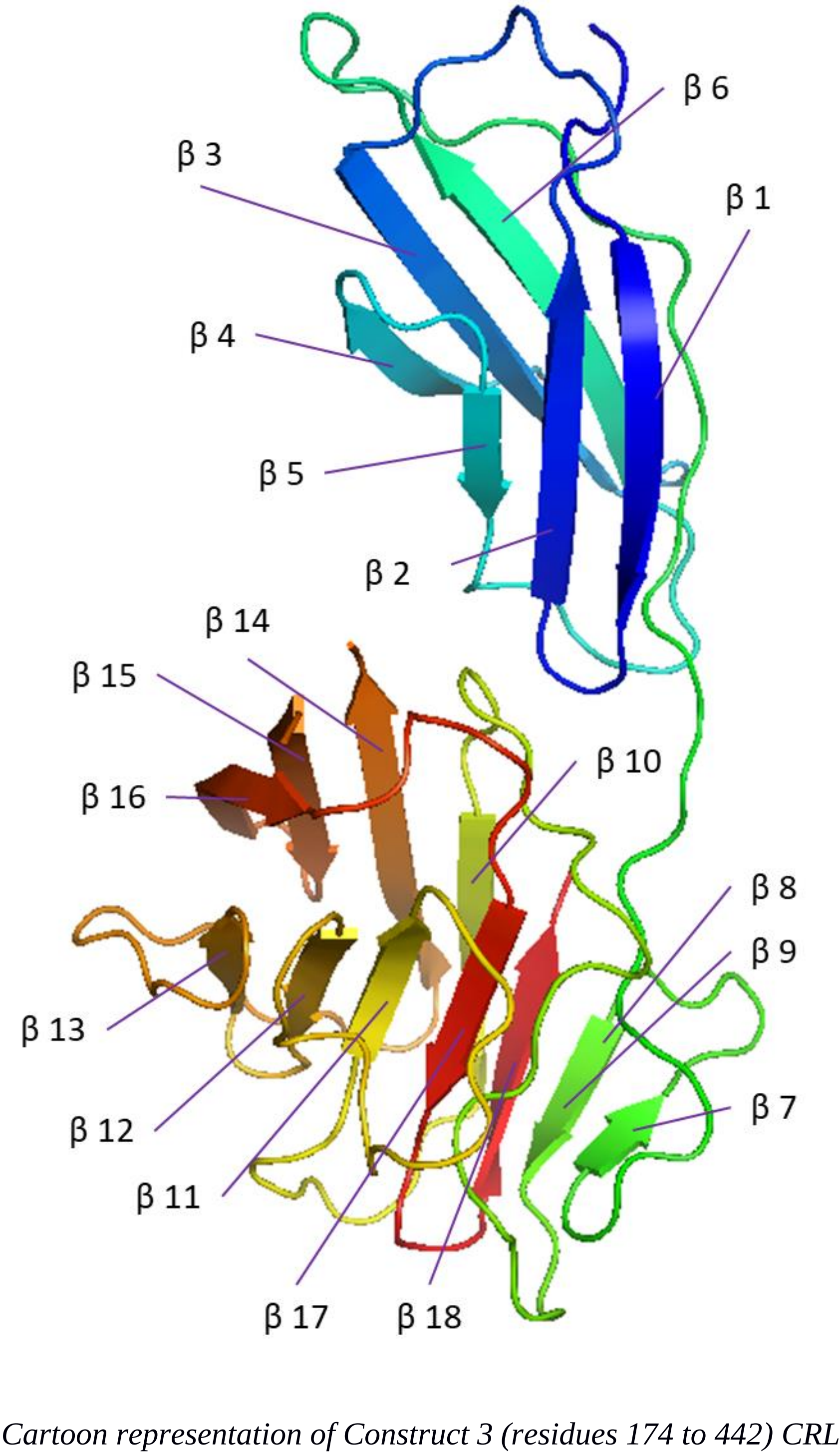


Figure 2-6 Cartoon representation of Construct 3 (residues 174 to 442) CRLF3

This representation has been taken from the best available model, using rainbow colouring, with blue being the N-terminus and red the C-terminus. There are clearly two distinct domains in this construct. (Figure created using *PyMOL*[261]).

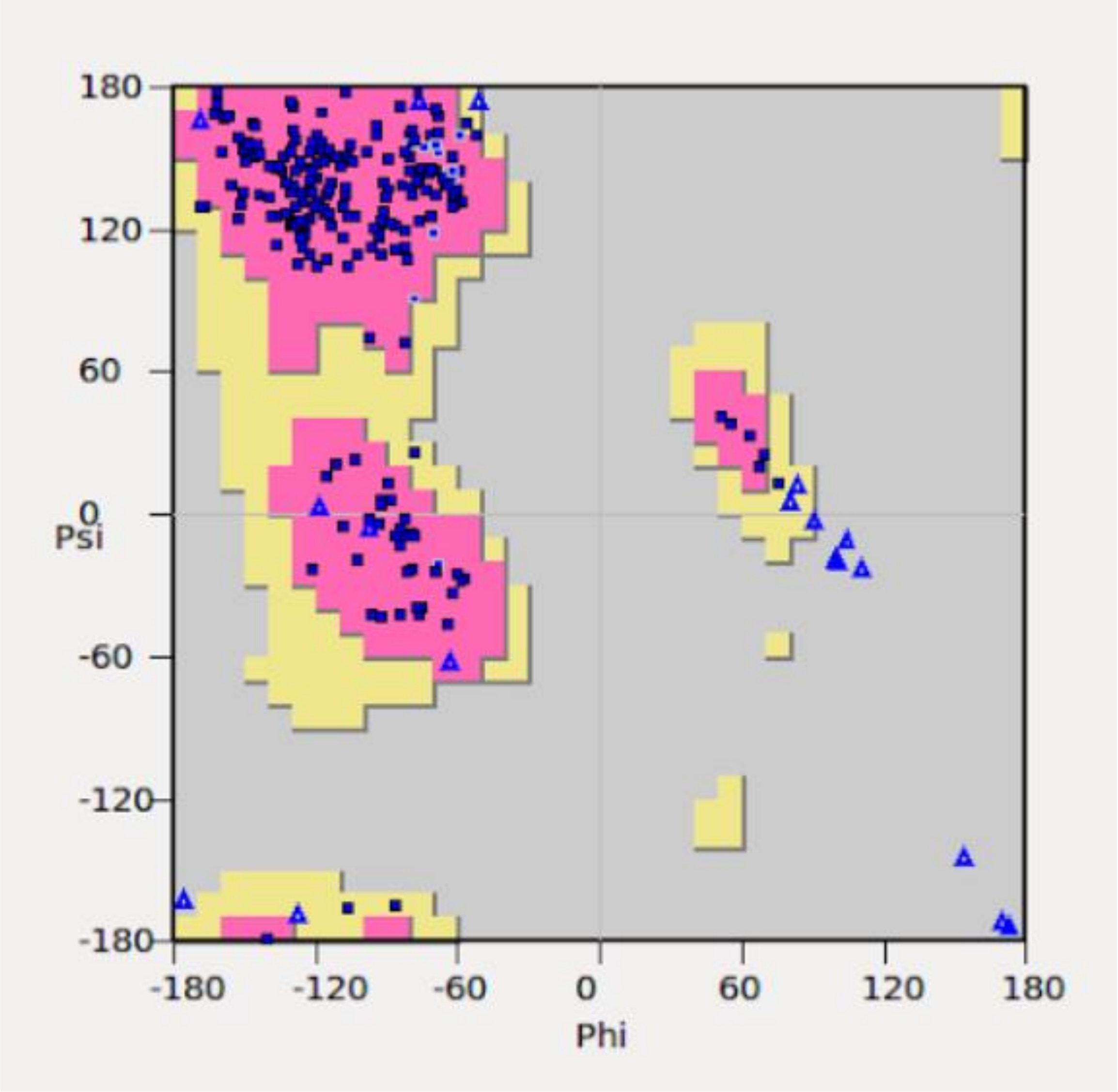


Figure 2-7: A Ramachandran plot for the CRLF3 protein model 1

This figure shows the φ and ψ angles for all modelled residues in the protein as dark blue triangles (showing glycine residues) or squares (all other residues). The pink region shows the favourable region for a glutamine residue, and yellow the allowed region. (plot generated using *Coot*[23]).

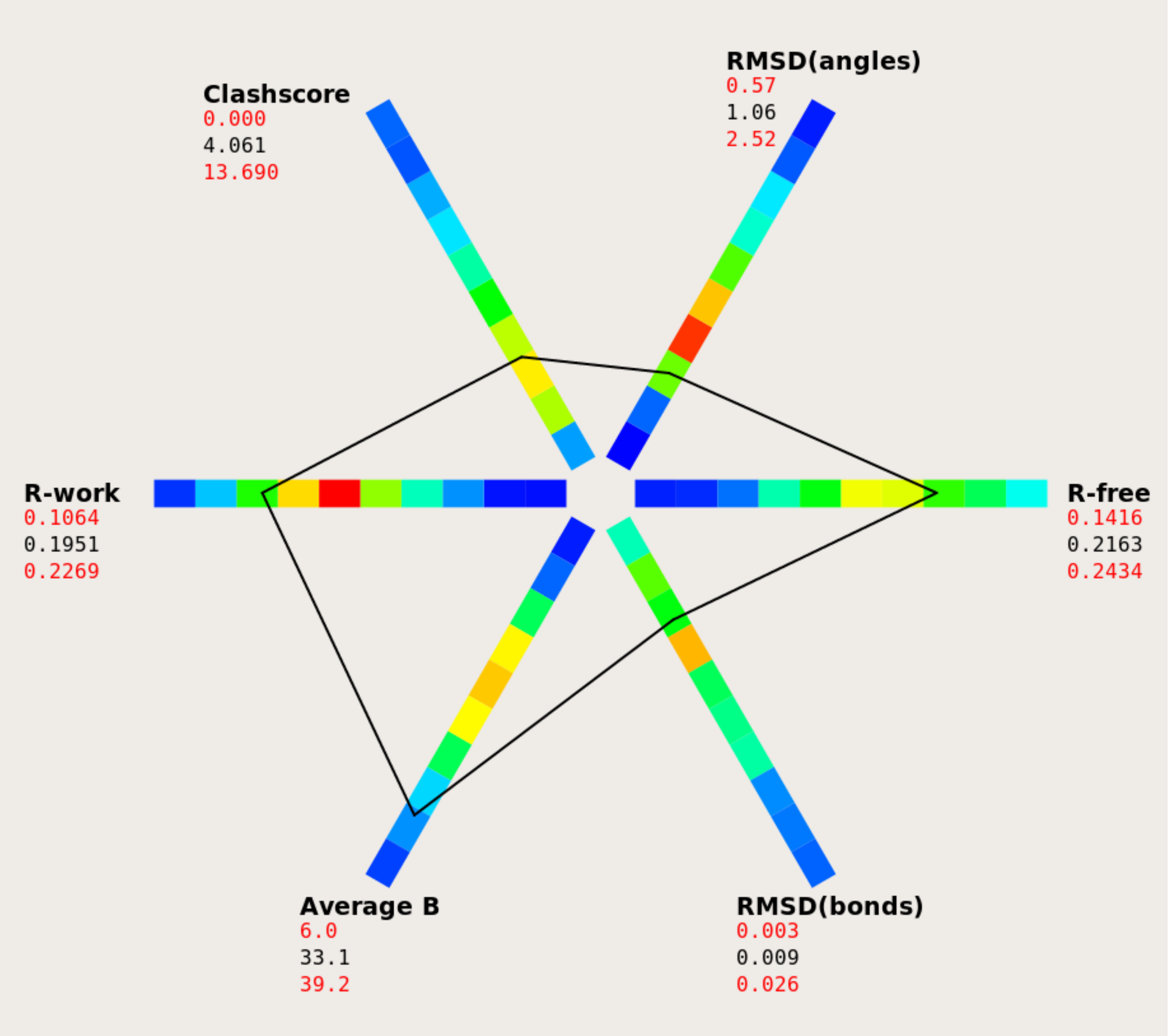


Figure 2-8: Polygon report for CRLF3 construct 3 final model 1

This polygon report[262] image was generated using *MolProbity*[168,169] through *phenix*[226]

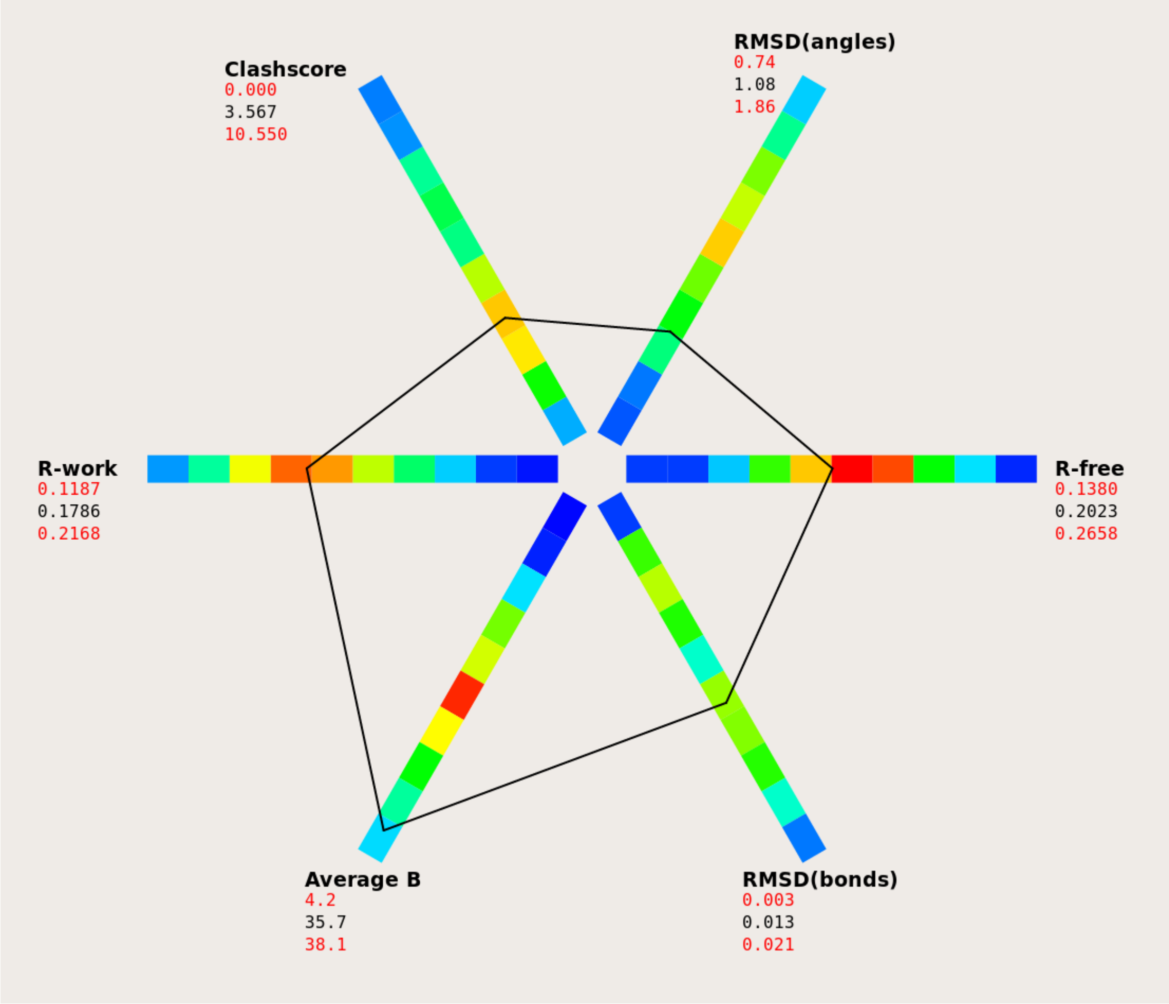


Figure 2-9: Polygon report for CRLF3 construct 3 final model 2

This polygon report[262] image was generated using *MolProbity*[168,169] through *phenix*[226]

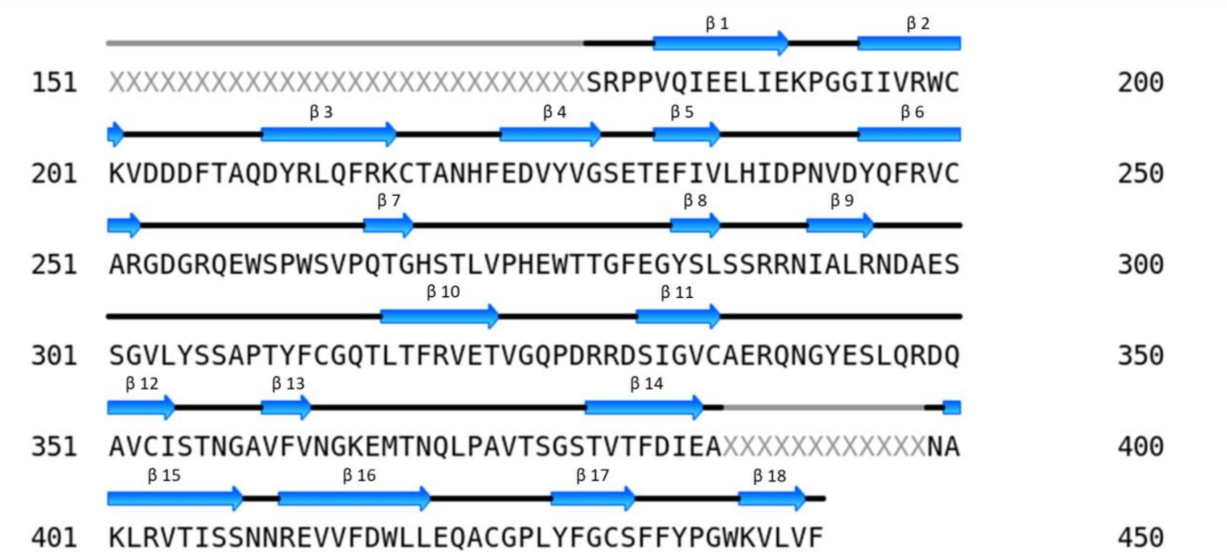

β 1
β 2
151 XXXXXXXXXXXXXXXXXXXXXXXXXXXXSRPPVQIEELIEKPGGIIVRWC 200
β 3
β 4
β 5
β 6
201 KVDDDFTAQDYRLQFRKCTANHFEDVYVGSETEFIVLHIDPNVDYQFRVC 250
β 7
β 8
β 9
251 ARGDGRQEWSPWSVPQTGHSTLVPHEWTTGFEGYSLSSRRNIALRNDAES 300
β 10
β 11
301 SGVLYSSAPTYFCGQTLTFRVETVGQPDRRDSIGVCAERQNGYESLQRDQ 350
β 12
β 13
β 14
351 AVCISTNGAVFVNGKEMTNQLPAVTSGSTVTFDIEAXXXXXXXXXXXXNA 400
β 15
β 16
β 17
β 18
401 KLRVTISSNNREVVFDWLLEQACGPLYFGCSFFYPGWKVLVF 450

Figure 2-10: Secondary structure elements from CRLF3 construct 3 final model 1

This image was generated using *MolProbity*[168,169] through *phenix*[226]. The letters above show the amino acids that were successfully placed in the electron density. X shows the places where the model was not built, as the electron density was not able to support the placing of those residues. The construct only started from 174, so only 5 residues were missing from the beginning of the construct.

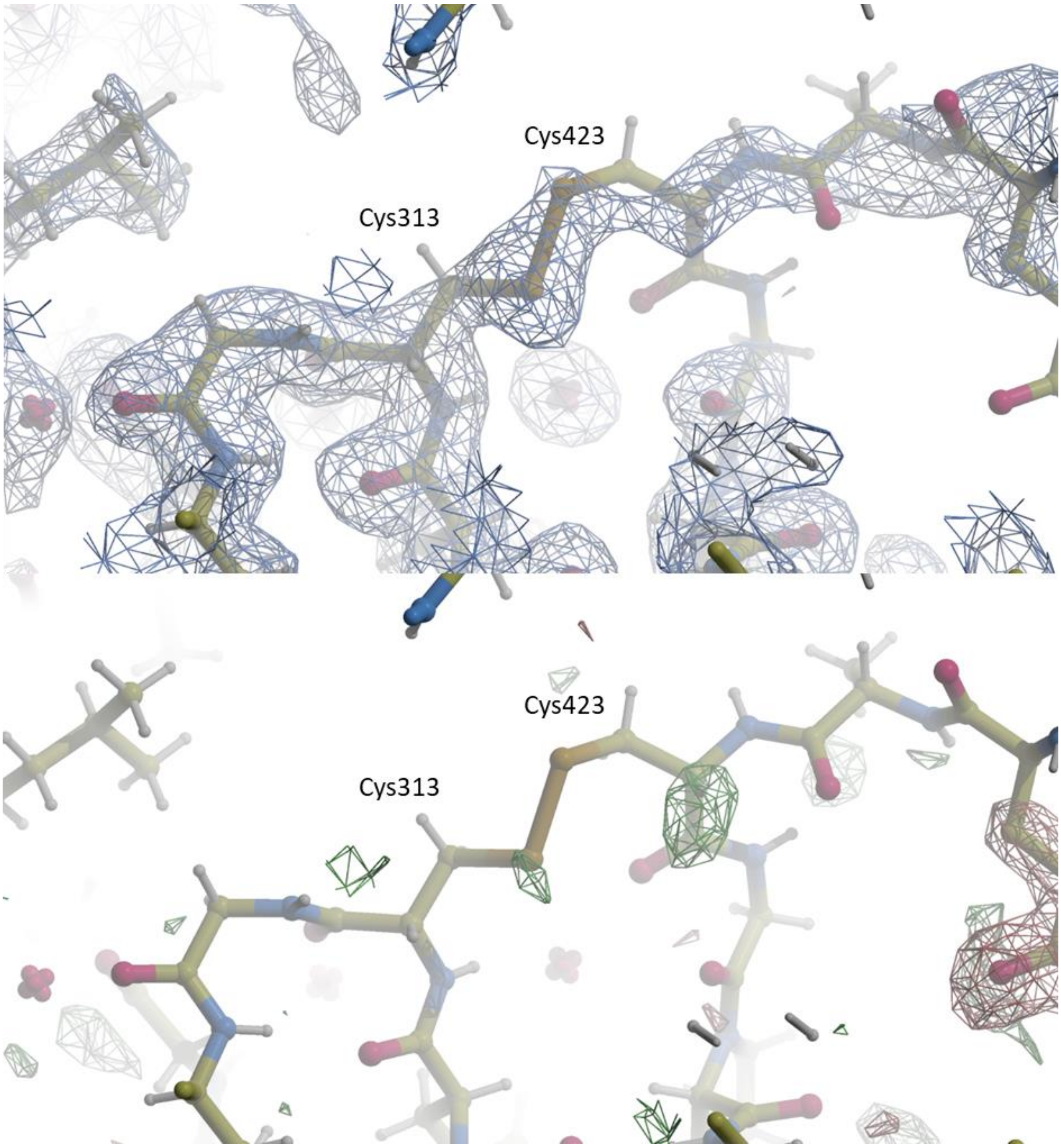


Figure 2-11: Electron and difference density maps of CRLF3 construct 3 final model 1 around the disulphide bond

Diagram indicates that the disulphide bond is present in model 1. The electron density (a $2F_O - F_C$ map) is contoured at the 1σ level, and the difference density (a $F_O - F_C$ map) at the 3σ level (images generated in *Coot*[23] and *RASTER3D*[24]).

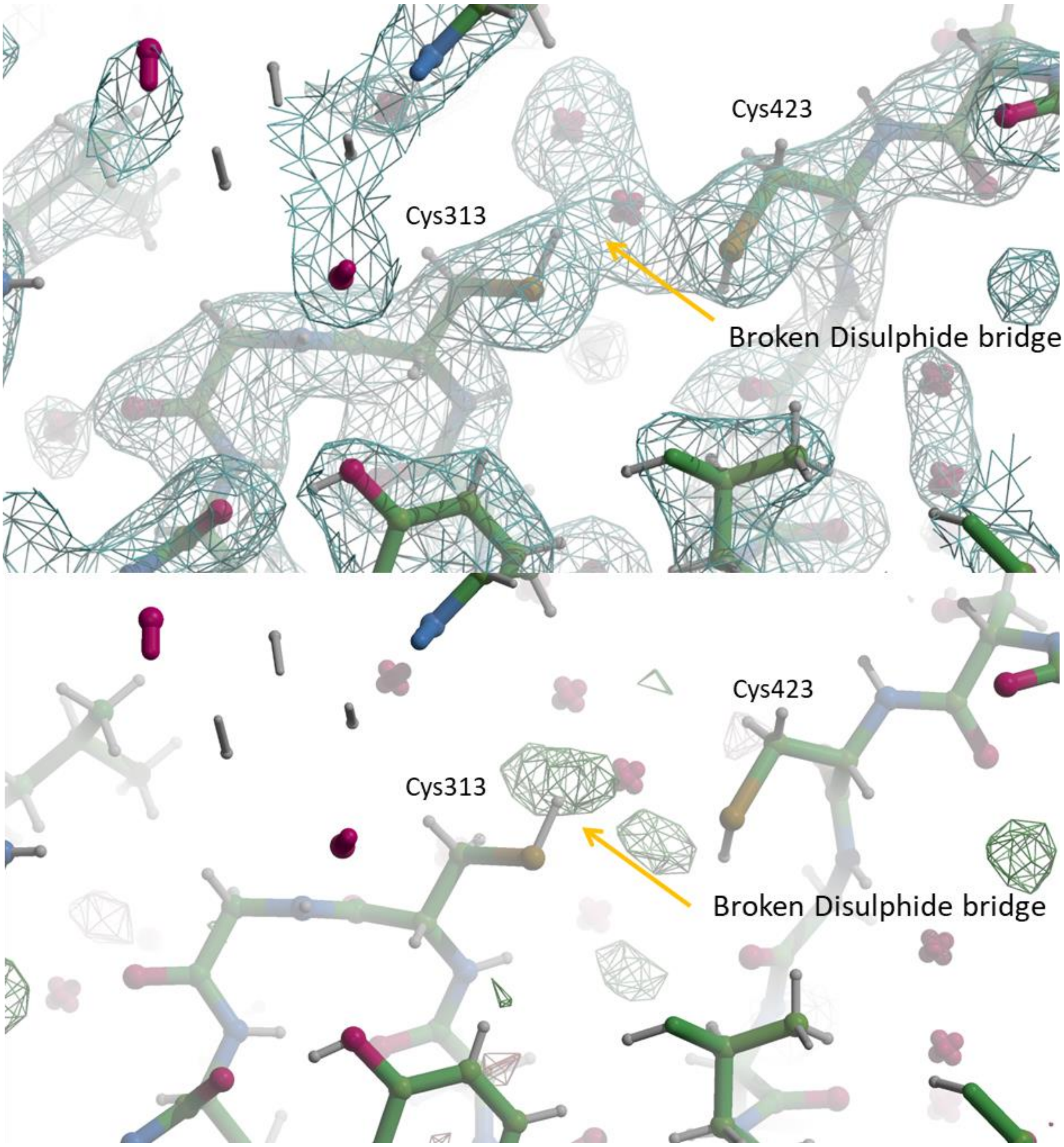


Figure 2-12: Electron and difference density maps of CRLF3 construct 3 final model 2 around the disulphide bond

Diagram indicates that the disulphide bond is broken in model 2, indicative of possible radiation damage. The electron density (a $2F_O - F_C$ map) is contoured at the 1σ level, and the difference density (a $F_O - F_C$ map) at the 3σ level (images generated in *Coot*[23] and *RASTER3D*[24]).

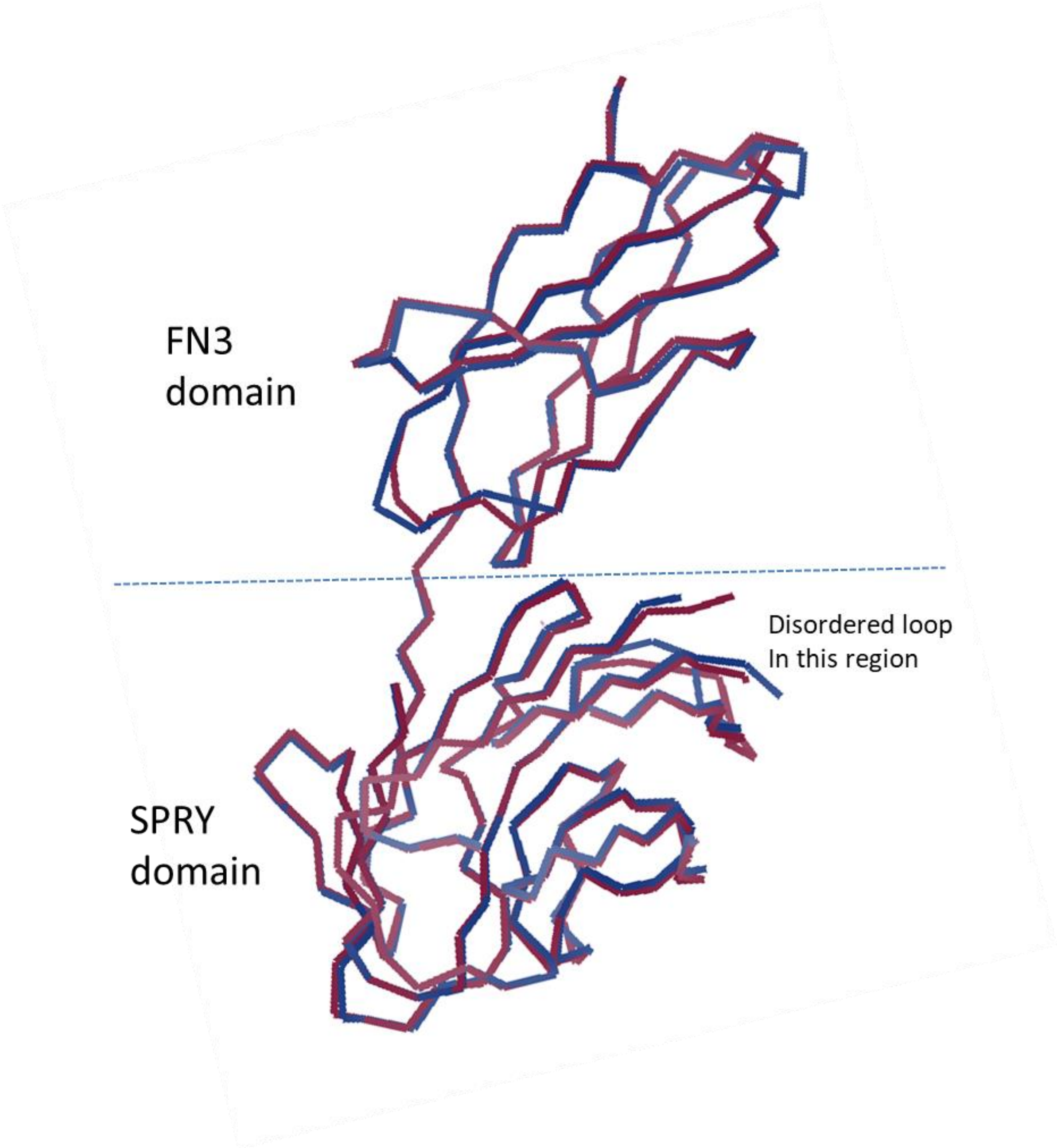


Figure 2-13 Showing the main difference between CRLF3 construct 3 final model 1 and final model 2

This view shows superimposed images of the two models. The image from the first data model is shown in dark blue, and that for the second model in dark red. A shift in the model, leading to a slightly different angle between the two domains, can be seen, and is ascribed to this difference in the disulphide bonding. In the FN3 domain, the first data model (red) is predominantly above the second (blue), and the converse is true in the SPRY domain. (image generated in *Coot*[23] and *RASTER3D*[24]).

It is unsurprising that the second model showed evidence of greater experimental damage. The results were taken at a longer wavelength, where the sulphur atoms, which scatter anomalously, absorb more of the incoming energy (Table 2-6). This increases the likelihood that this bond will be broken in the proteins in the crystal.

Another observation was that the substructure of sulphur atoms calculated after the exhaustive search with *HySS*[252] seemed to be in credible positions. However, the sulphur substructure did not provide enough phase information within the diffraction pattern to create an effective initial estimate, making it impossible to build a model.

The polygon report[262] of both structures (Figure 2-8 and Figure 2-9) is typical of structures with similar resolution. The average B factors are quite high. The R-work, R-free and clash score are all slightly worse than average. However, the root-mean-square deviation (RMSD) of the bonds and angles from the ideal are both better than average. CRLF3 Construct 3 final model 1 was the best structure that could be obtained and so it was the structure that was taken forward to be analysed.

The CRLF3 construct 3 protein has two clear domains. The N-terminus domain is a fibronectin type 3 (FN3) domain. Such domains have been implicated in cell adhesion, cell morphology, thrombosis, cell migration, and embryonic differentiation[263]. This domain is found in a significant number of proteins (some stating that it was found in up to 2 % of proteins[264]), its role has been ascribed as primarily functional, but its loops can determine interactions with other proteins[265]. There is no further general information, but a hypothesis can be proposed that this domain might be interacting with the tubulin in micro-tubules, and so destabilising their structure.

The second domain is a SPRY domain, so called as this structure comes from SPla and the RYanodine Receptor. The function of this domain is unknown. SPRY containing proteins have been shown to be important in the platelet count in the protein in previous literature[266,267], though no clear functional role has been ascribed as yet to the SPRY domain.
The above two domains were identified in the CRLF3 Construct 3 structure. The other part of the native protein has very few homologous models, and they are primarily coiled coils, which are hard to crystallise, and it is hard to predict exact roles. However, given that the

homology model is linked to ubiquitination, it suggests that the first half of the protein might have a role in this function.

Now that a structural model has been obtained, some of the previous models that were used in molecular replacement can be re-examined and compared. The molecular replacement models covering both domains were not a good match (for example 3T1W and 1FNF), as they did not predict the geometry between the two domains well. This led to some misalignment. Additionally, the backbone trace was also a poor match, and helps to explain the failure of the molecular replacement solution (Figure 2-14).

Molecular replacement models that covered only one domain were also used (1UC6 and 1CD9). These matched up reasonably to the final structure, but again not well enough to allow for a solution (Figure 2-15).

The *I-Tasser*[245–248] model correctly predicted the domains of the structure, but the trace was not close enough to provide enough correct phase information for molecular replacements (Figure 2-16).

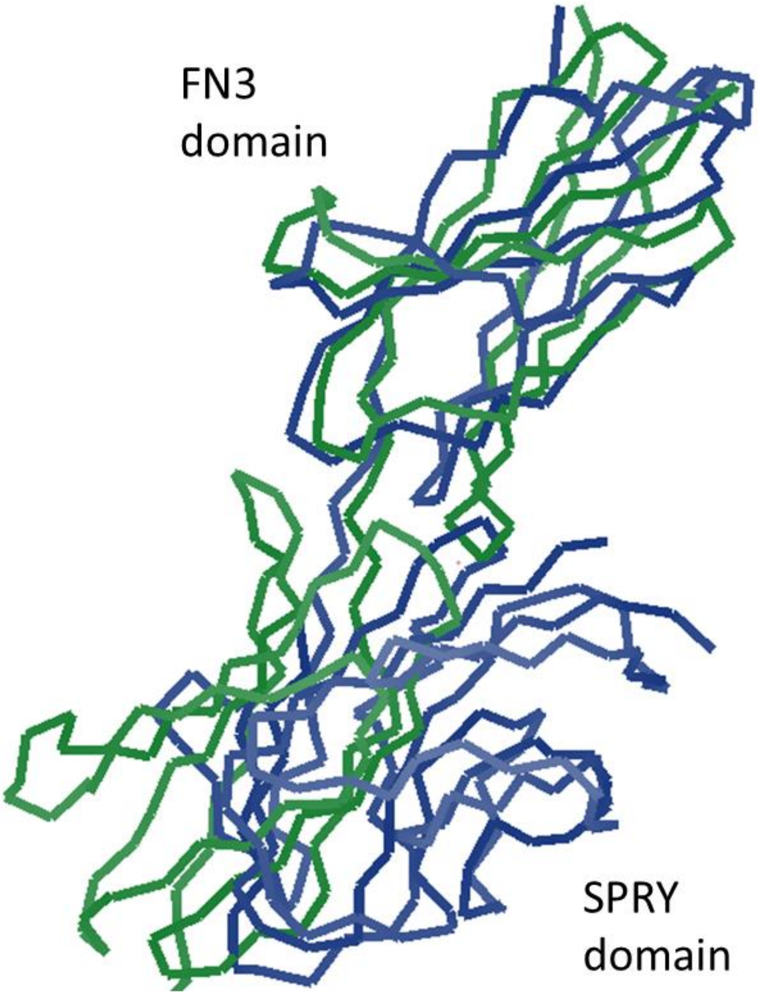


Figure 2-14: Comparison of the 3T1W molecular replacement model (green) versus the final model 1 (blue) of the CRLF3 construct 3

Core RMSD calculated is 5.4 Å. (Image generated using *Coot*[23] and *RASTER3D*[24]).

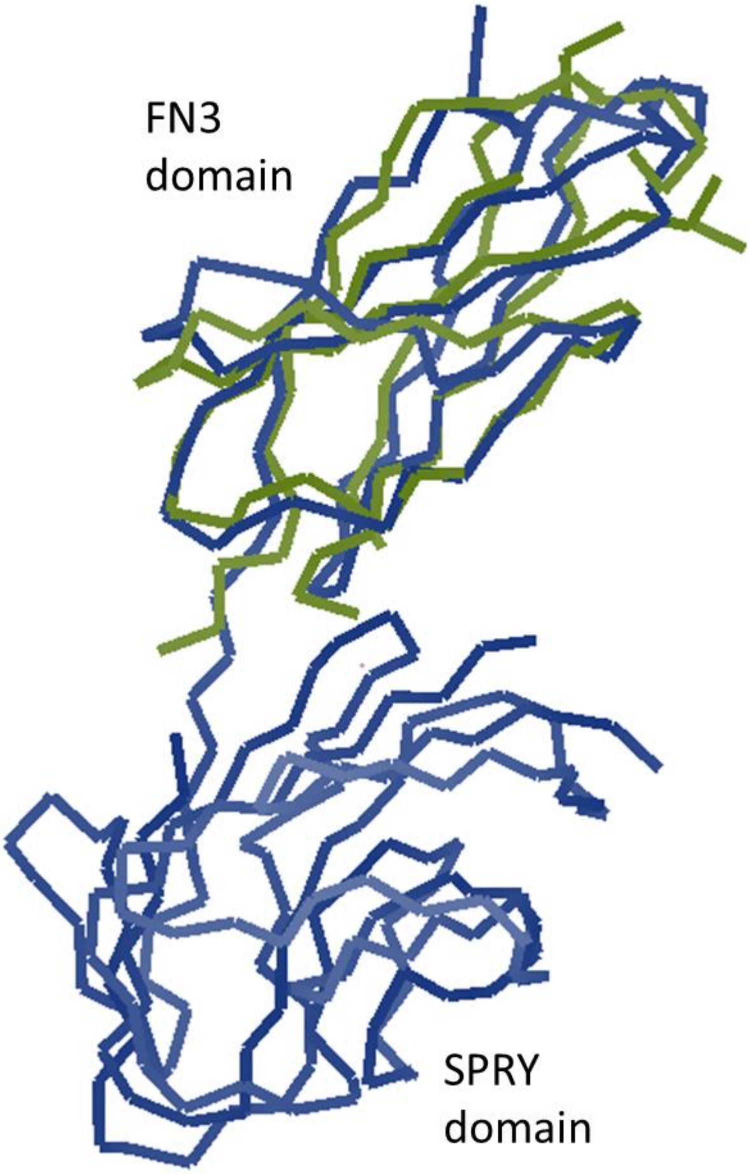


Figure 2-15: Comparison of the 1UC6 molecular replacement model (green) versus the final model 1 (blue) of the CRLF3 construct 3

Core RMSD calculated is 2.0 Å. (Image generated using *Coot*[23] and *RASTER3D*[24]).

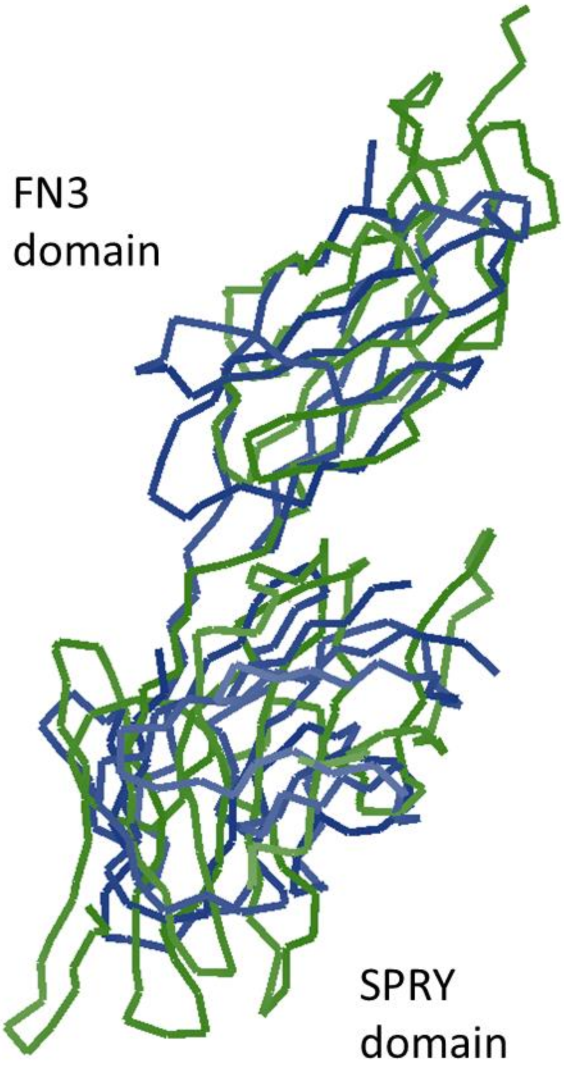


Figure 2-16: Comparison of the I-Tasser predicted model (green) versus the final model 1 (blue) of the CRLF3 construct 3

Core RMSD calculated is 4.8 Å. (Image generated using *Coot*[23] and *RASTER3D*[24]).

2.10 Analysing the model with the results from ExAC

CRLF3 Construct 3 final model 1 was then used to investigate the effect of mutations. The aim was to inform the assessment of the mutations that had been identified as possibly leading to platelet and bleeding diseases.

The BRIDGE consortium, which is the organisational body of the NIHR BioResource funded Next Generation Sequencing (NGS) projects, has as its two main aims the discovery of novel sequence variants that lead to rare genetic diseases, and the determination of the success of NGS approaches in re-identifying rare disease-causing variants that are already known to exist. It is run by Mr Antony Attwood, Department of Haematology, University of Cambridge. The clinical cases for BRIDGE in the UK are placed in the NIHR BioResource, with clinical cases from outside the UK being accepted as well. Sequence information together with disease information is released for other scientists through the European Genome-phenome archive at the European Bioinformatics Institute (EGA).

One of the projects in the BRIDGE consortium was exome-sequencing with patients with platelet and bleeding diseases. A number of potential mutations (polymorphisms) in the Crlf3 gene were seen in some patients, which were then further explored. These polymorphisms may or may not affect the function or stability of the protein.

First, the significance of these mutations was examined. This was achieved using the database created by the Exome Aggregation Consortium (ExAC)[268]. This project involved combining the genome data from 60,706 humans with an easy to use web explorer. It was assumed that if mutations to the Crlf3 gene were found in ExAC, which is a project not looking specifically at any bleeding or platelet disease, then they are likely to be mutations that do not harm the protein and are likely to be tolerated. Therefore, reviewing the mutations in ExAC should act as a helpful screening process.

As can be seen (Table 2-10 and Figure 2-17 to Figure 2-33), there is a huge variety in the number of matches found in ExAC. The residues that have zero matches in ExAC are very interesting, as these are likely to be residues that will be significant in causing disease.

The mutations that have only a few matches in ExAC might also be interesting. The people with these rare mutations may have had more significant diseases which mask the presentation of symptoms, or may have had other mutations in the protein which affect the expression of the noted mutations. Finally, the mutations that have a huge number of matches, such as Pro309Leu (Figure 2-27), are highly unlikely to be important, as many people have this mutation without a significant pathology in clotting diseases.

Mutation	ASA (%)	B/E	No. of mutations in ExAC	Environment
Cys430Ser	0.6	B	0	Very buried hydrophobic residue – polar residue in this position will stabilise the denatured state, so the overall protein is destabilised.
Asn410Asp	96.6	E	73	Completely exposed residue, can accommodate mutation completely.
Gly398Arg	-	-	1	Disordered loop, so no information.
Asn394Ser	-	-	12	Disordered loop, so no information.
Thr392Ile	-	-	1	Disordered loop, so no information.
Val387Met	-	-	7	Disordered loop, so no information.
Phe382Ser	0.0	B	0	Loss of hydrophobic interactions in the core of the protein with Val335, Ile354, Leu317 and Val380.
Asp349Tyr	66.5	E	0	Loss of interaction with Gln350, leading to destabilisation of the protein.
Thr323Ile	46.9	E	0	Bulkier group may cause steric clash. Not clear.
Cys313Phe	20.1		0	Eliminates di-sulphide bridge.
Pro309Leu	40.2	E	445	Exposed, so can accommodate mutation.
Ala308Ser	62.7	E	1	Plenty of space to accommodate mutation.
Asp297Asn	88.7	E	3	No residues near it, so no evidence of issue.
Val236Ile	2.4	B	12	Although buried, there is space for mutation to be accommodated.
Ile235Thr	59.7	E	55	Mutation can be accommodated, so unlikely to affect the protein.
Gly229Arg	20.0		0	Plenty of space, but backchain is in an unfavourable

				φ and ψ for Arg residue, so this might cause protein to stay in the denatured form
Val228Ile	39.3		4	Plenty of space to accommodate the potential clash of sidechains.
Val226Leu	10.2	B	3	Possible steric clash with Phe234, though likely to be accommodated.
Arg212Ser	27.8		0	Loss of hydrophobic interactions between the long side chain and the two phenyl rings and other hydrophobic groups. Loss of interactions with Asp225.
Val202Met	34.5		12	Can easily accommodate mutation – likely not to cause a problem.
Arg180His	67.3	E	43	Unclear – at start of crystallised chain. High number of mutations in ExAC, so unlikely to cause disease.

Table 2-10 Mutations found in CRLF3 by the BRIDGE consortium

The BRIDGE consortium is searching for mutations in proteins that cause blood diseases. The percentage accessible surface area (ASA) to solvent was calculated using the program GetArea[269,270]. Using the standards of previous papers in the lab[271,272], if the ASA is predicted to be 10 % or less then it is considered to be buried (B). If greater than 40 %, then it is considered to be exposed (E). The ExAC database[268] contains the genome data for 60,706 humans, from a large variety of sources. Therefore matches here most probably suggest that the mutation does not cause blood disease. Finally, there is text which explains the environment of each residue.

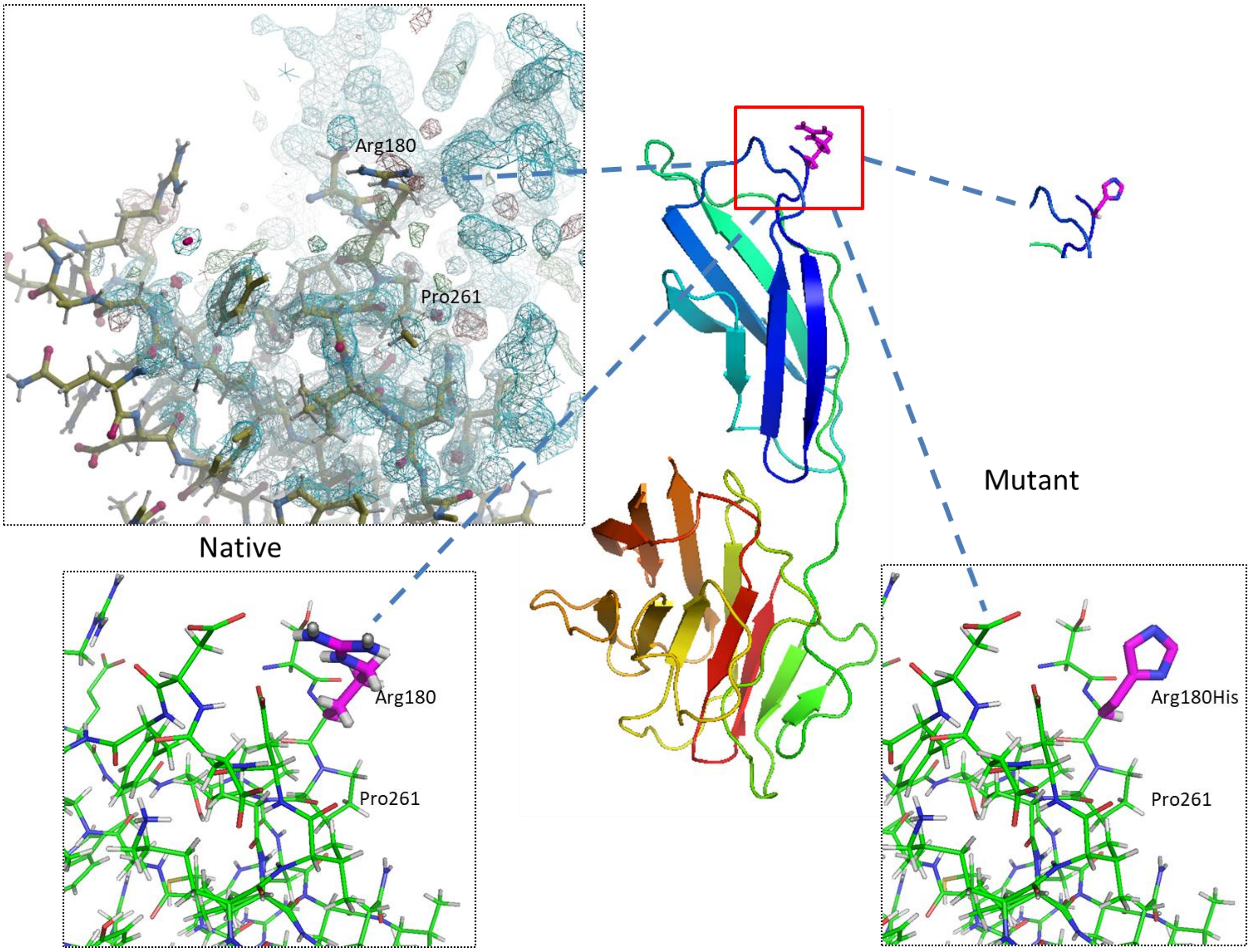


Figure 2-17: Mutation Arg180His: Location, native structure electron density and model of mutant

The location of the mutated region is shown on a cartoon representation of the native CRLF3 construct 3, with a stick model of the relevant residue displayed. The equivalent section (red box) with the residue associated with the mutation is shown on the top right diagram. The image on the top left is the electron density map of this region. The two lower images show the stick diagrams for the locale of the native (left) and mutated (right) variants, with neighbouring residues labelled. The mutated diagram shows the side chain in a probable configuration with no backbone reorganisation, using *Coot*[23] to change the residue in this way. The upper left, bottom left and bottom right images are all in the same orientation, but the central figure and the expanded box show the overall context. The effect of this mutation is unclear; it is at the start of crystallised chain. There are a high number of mutations in ExAC, so it is unlikely to cause disease. The electron density map ($2F_O - F_C$) [blue surface] was generated in *Coot*[23], and represents the electron density at the 1σ level, and the difference density ($F_O - F_C$) at the 3σ level [green surface for positive density, and red surface for negative density]. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

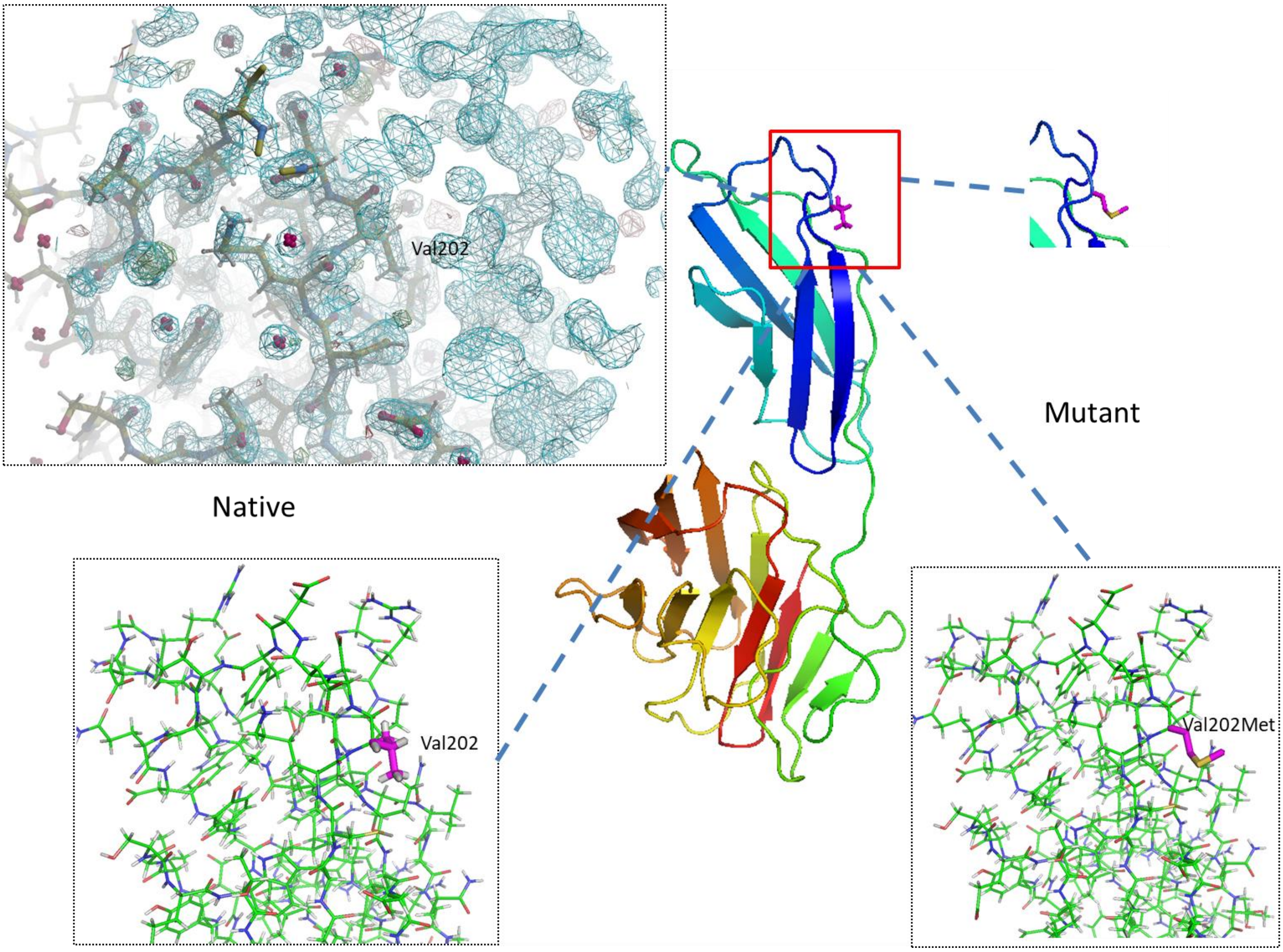


Figure 2-18: Mutation Val202Met: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). The structure can easily accommodate the mutation, so it is unlikely to cause disease. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

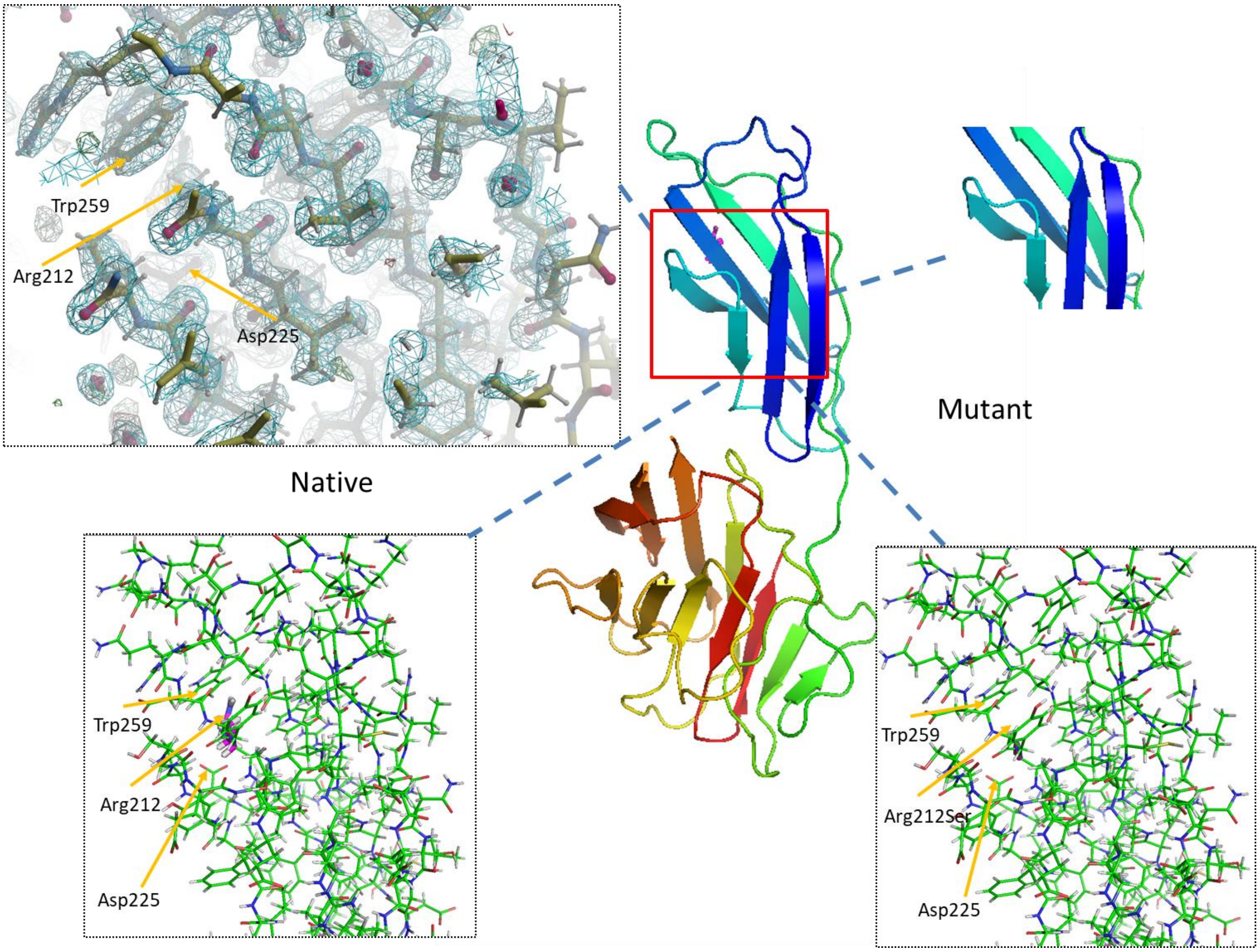


Figure 2-19: Mutation Arg212Ser: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This mutation shows a loss of hydrophobic interactions between the long side chain and the two phenyl rings and other hydrophobic groups, in particular a loss of interactions with Asp225. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

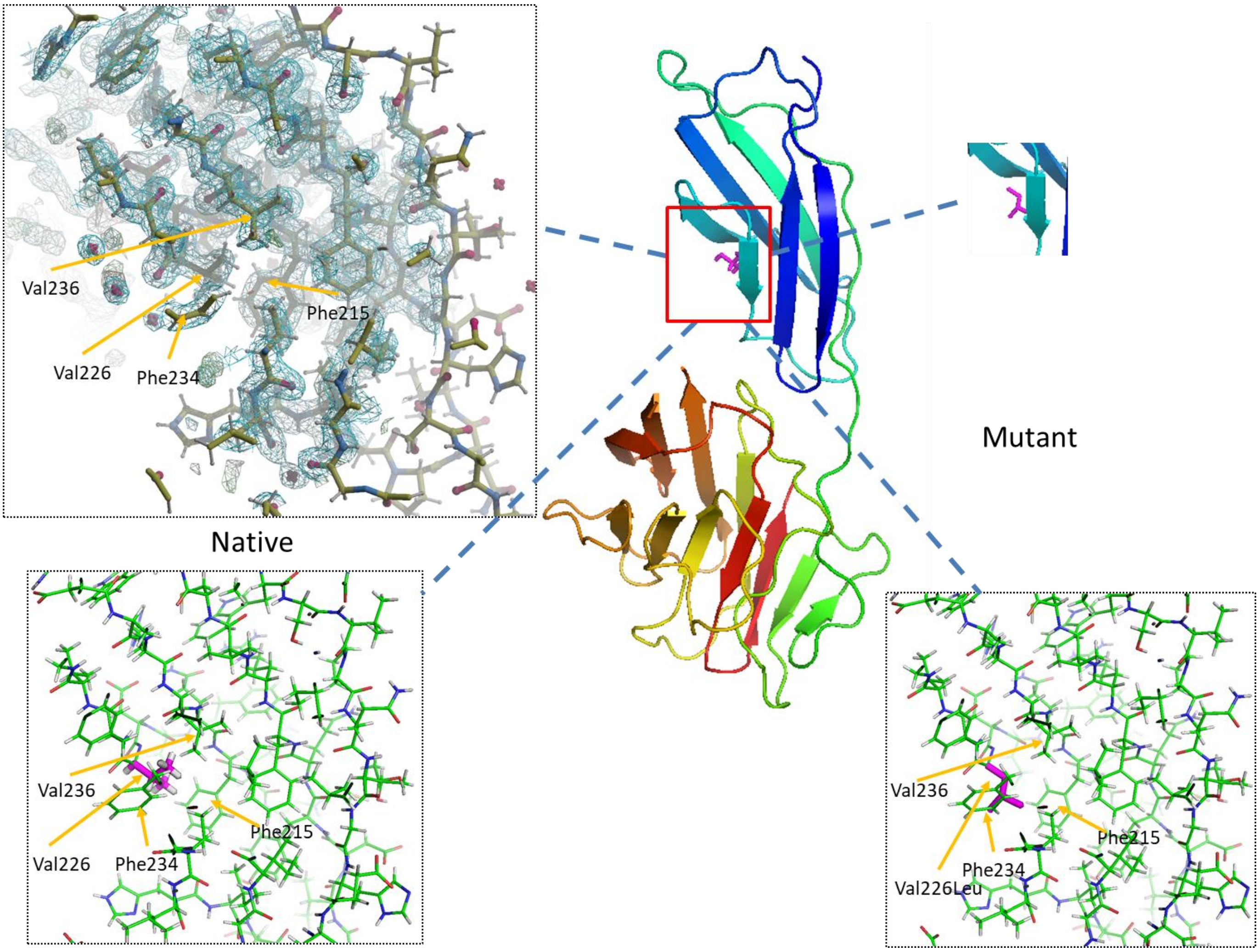


Figure 2-20: Mutation Val226Leu: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This mutation shows a possible steric clash with Phe234, though it is likely to be accommodated. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

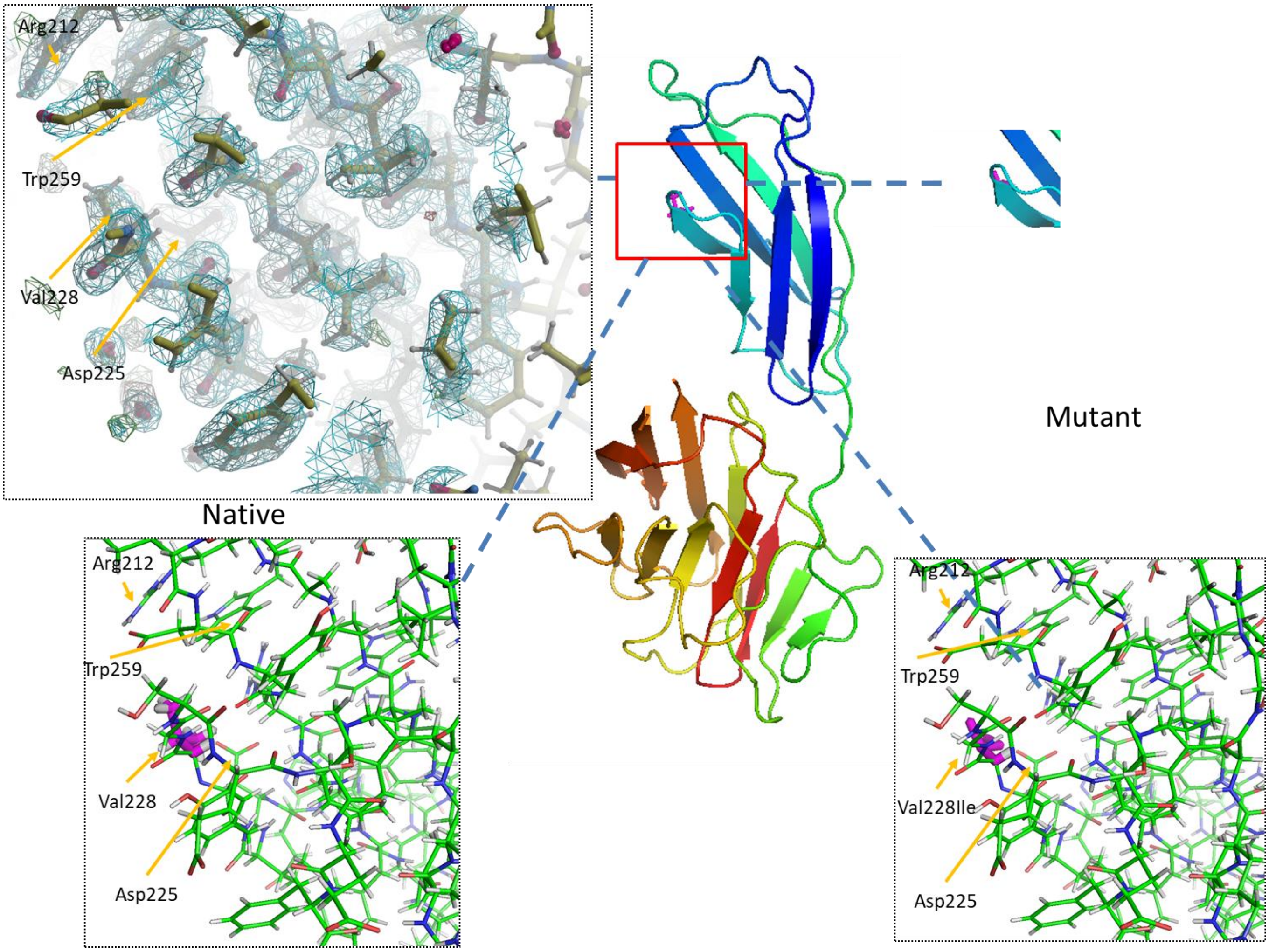


Figure 2-21: Mutation Val228Ile: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). The region has plenty of space to accommodate any potential clash. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

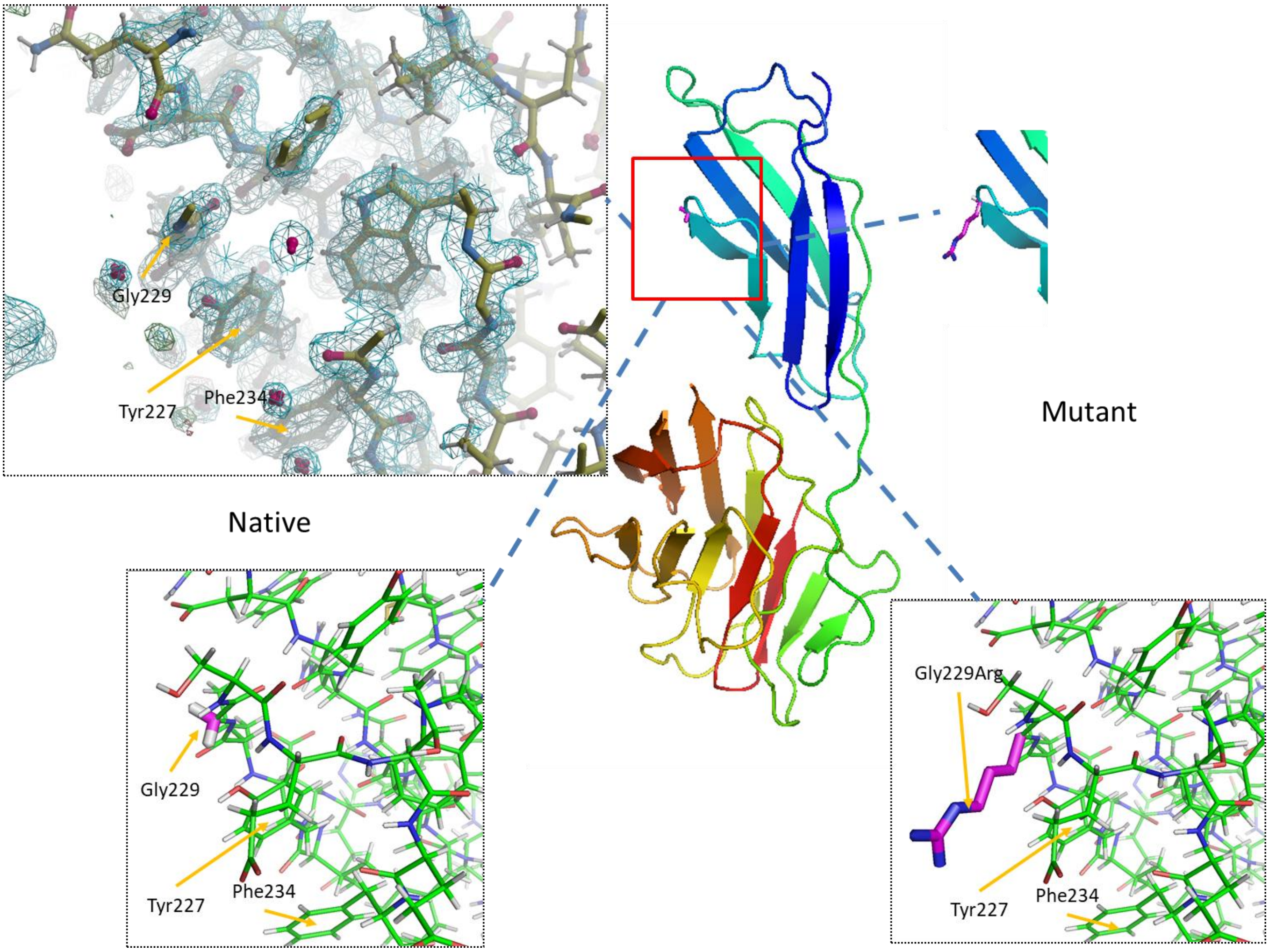


Figure 2-22: Mutation Gly229Arg: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). The region has plenty of space, but the backchain is in an unfavourable φ and ψ for the Arg residue, so this mutation might cause the protein to fall apart. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint u*sed to label the figures.

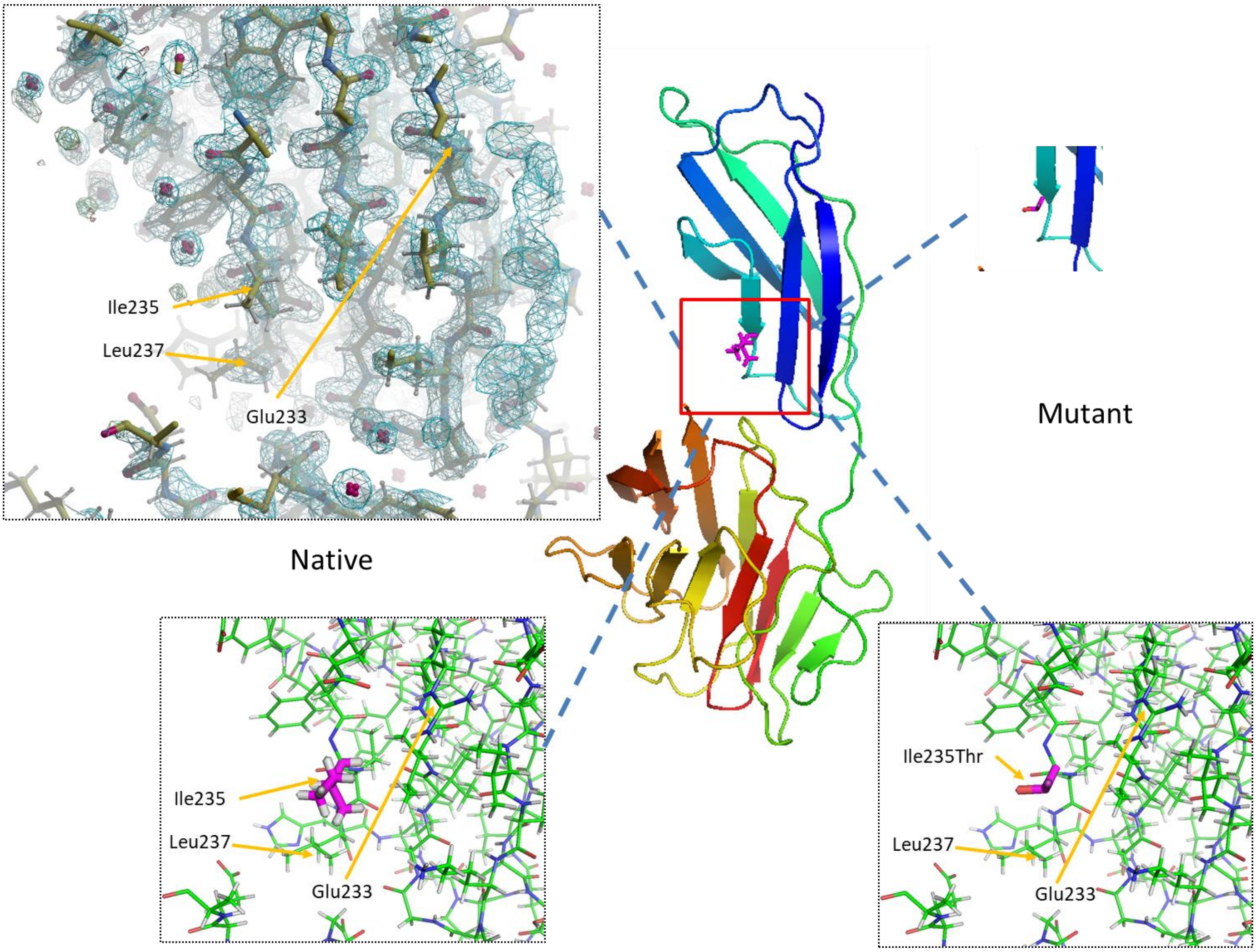


Figure 2-23: Mutation Ile235Thr: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). The mutation can be accommodated, so it is unlikely to affect the protein. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

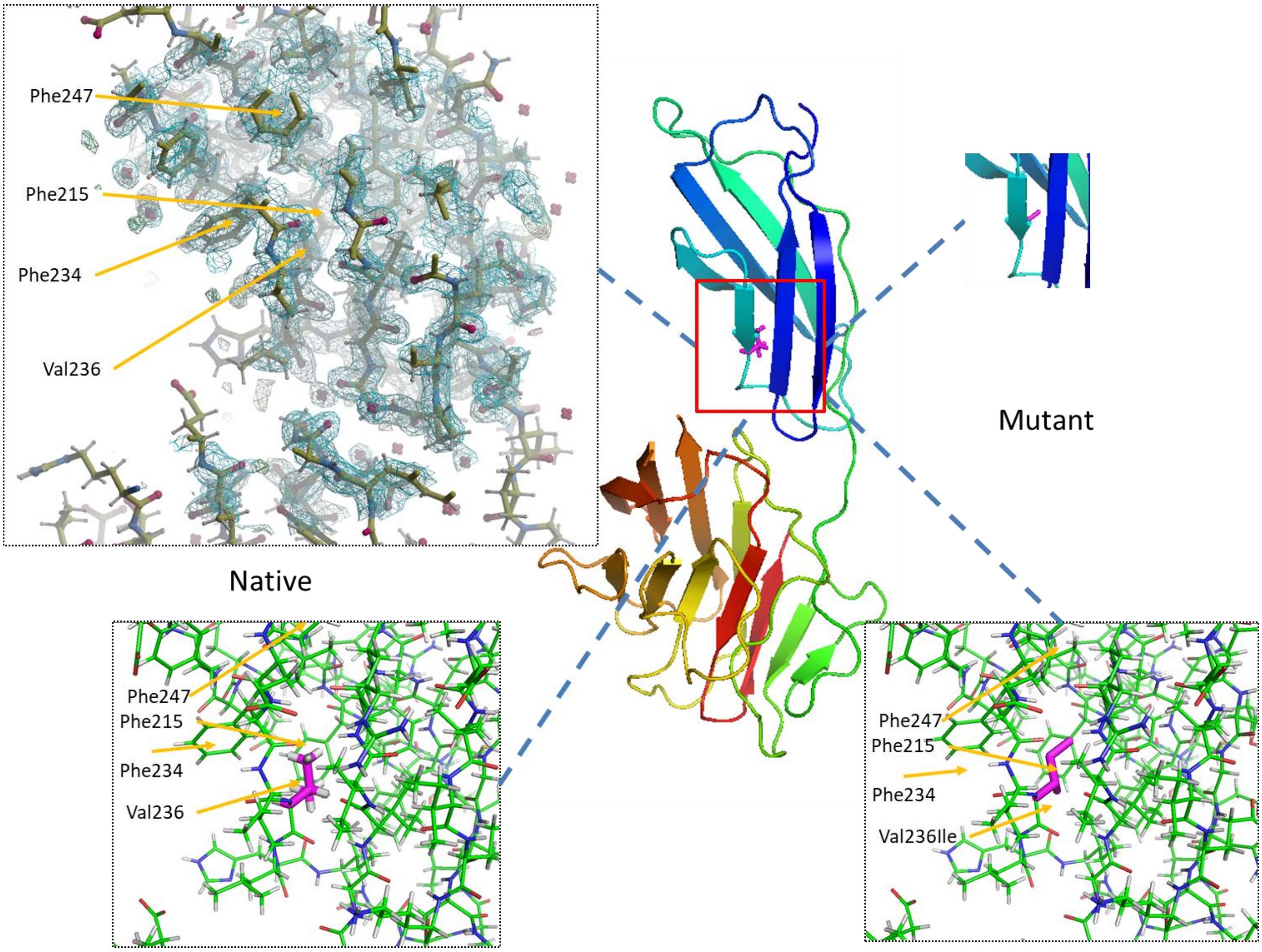


Figure 2-24: Mutation Val236Ile: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). Although buried, there is space for the mutation to be accommodated. This suggests that the mutation will not cause issues in itself. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

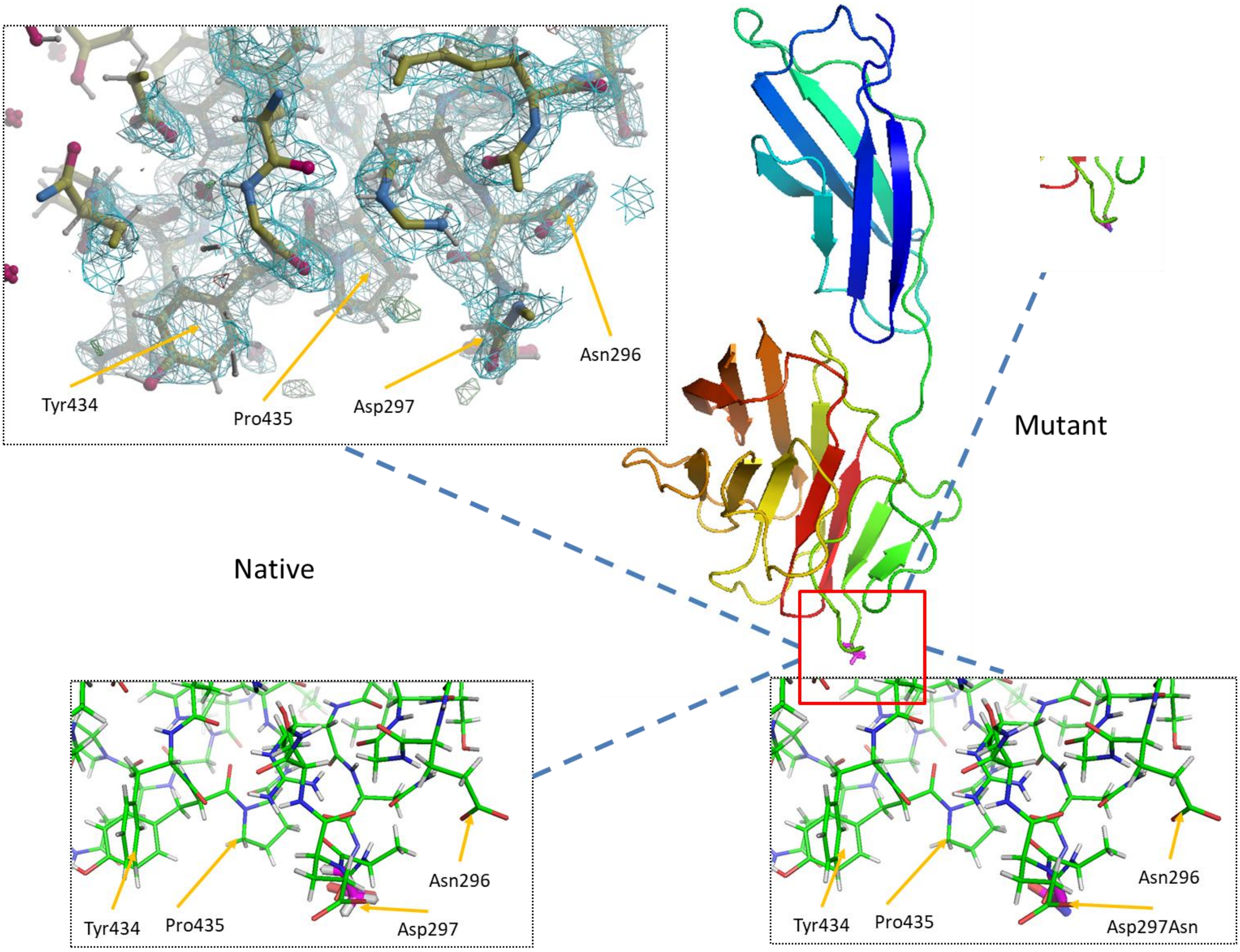


Figure 2-25: Mutation Asp297Asn: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). There are no residues near the mutation, so there is no evidence of any issue arising. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

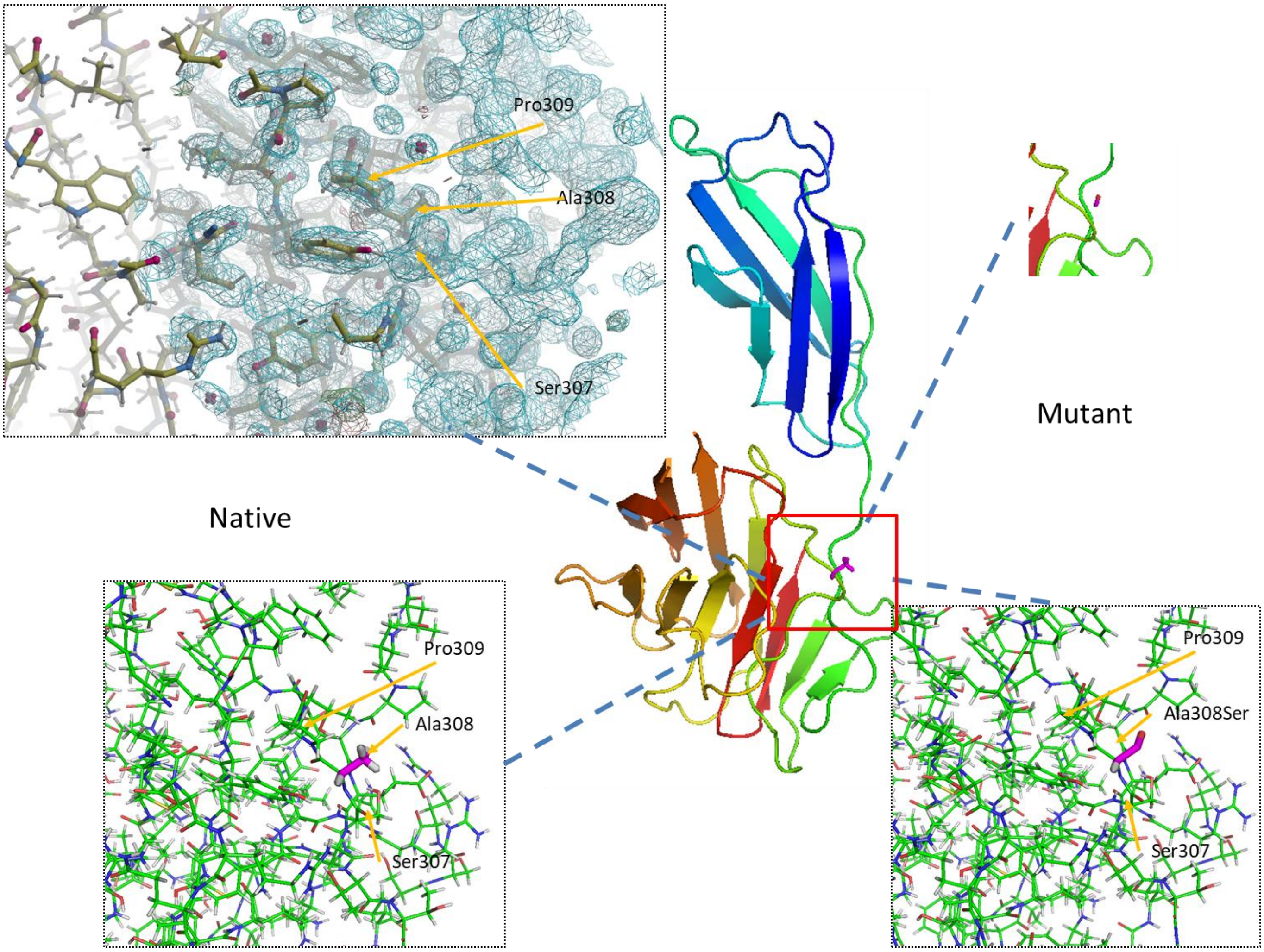


Figure 2-26: Mutation Ala308Ser: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). The region provides plenty of space to accommodate the mutation. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

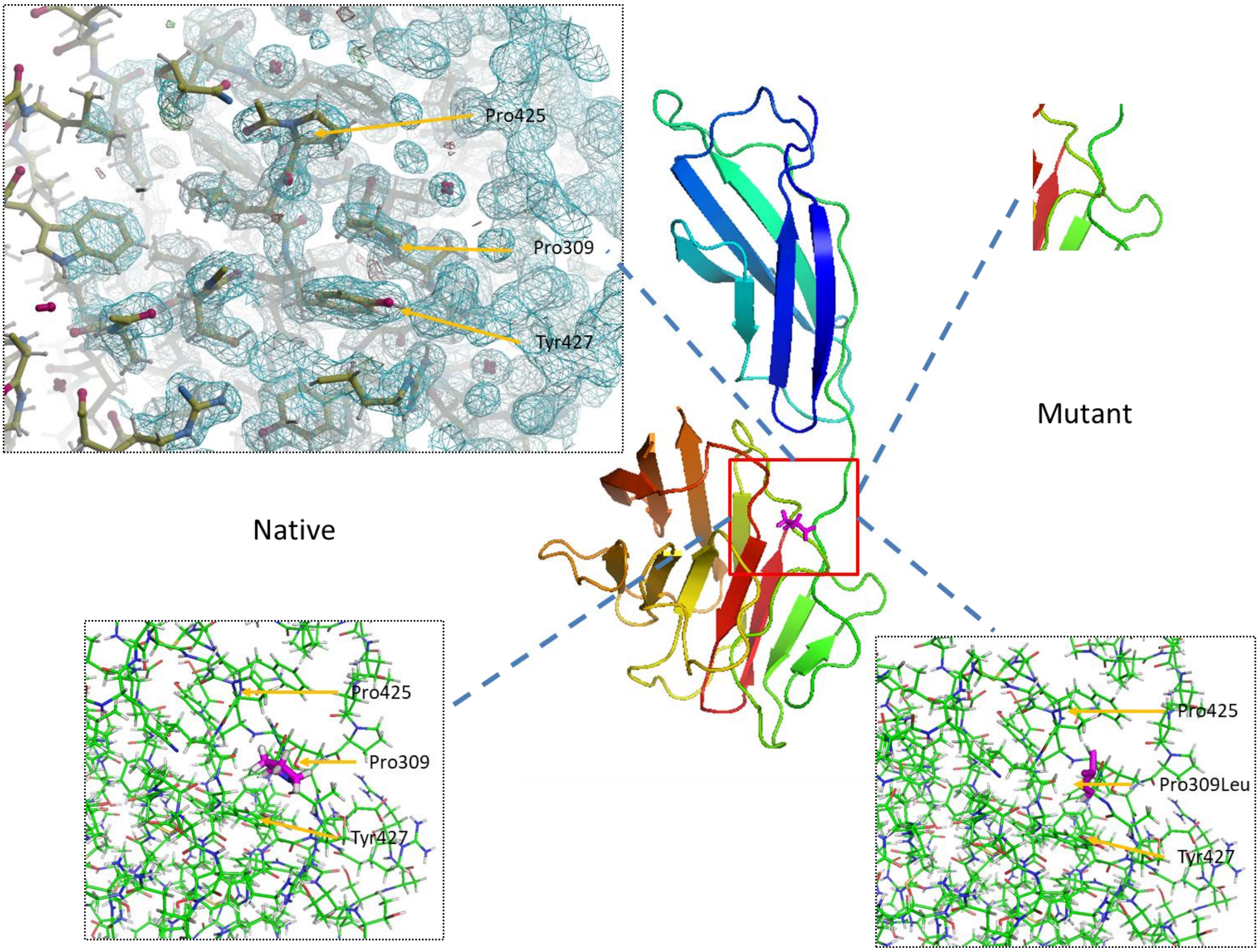


Figure 2-27: Mutation Pro309Leu: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). The area is exposed, so it can accommodate mutation easily. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

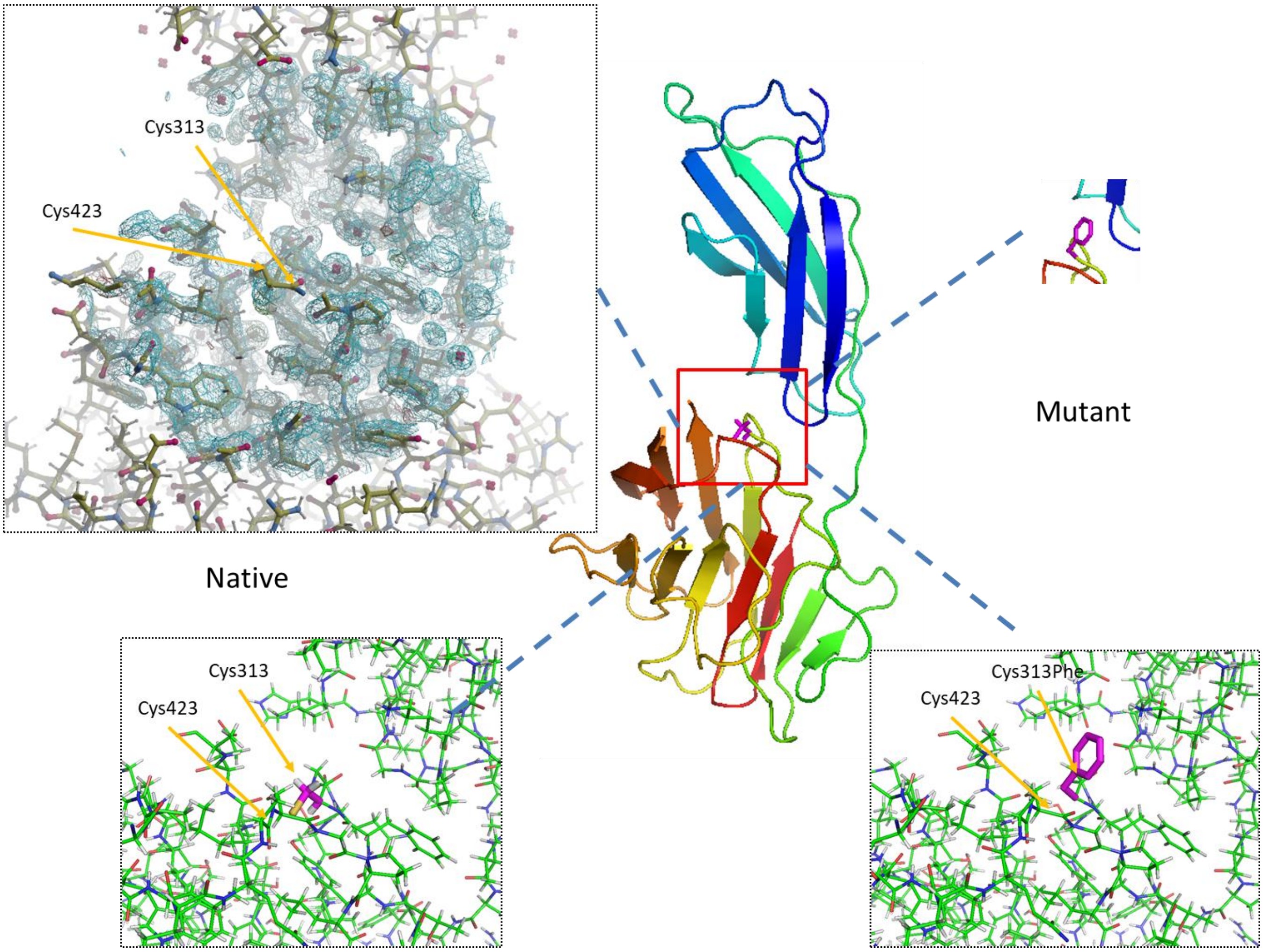


Figure 2-28: Mutation Cys313Phe: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This mutation eliminates di-sulphide bridge, so it is likely to have a major effect. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

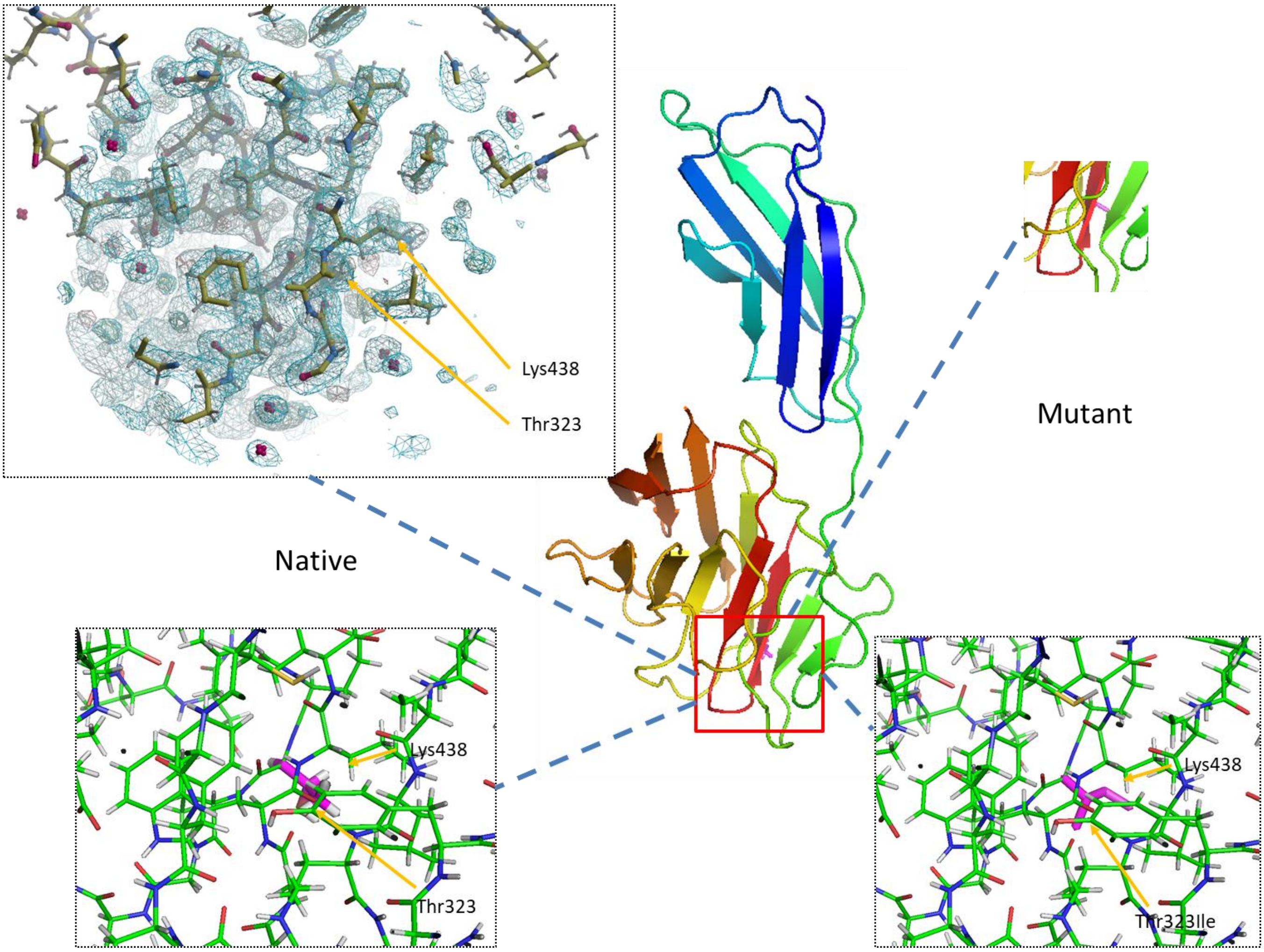


Figure 2-29: Mutation Thr323Ile: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This is a bulkier group that may cause steric clash. The impact is not clear. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

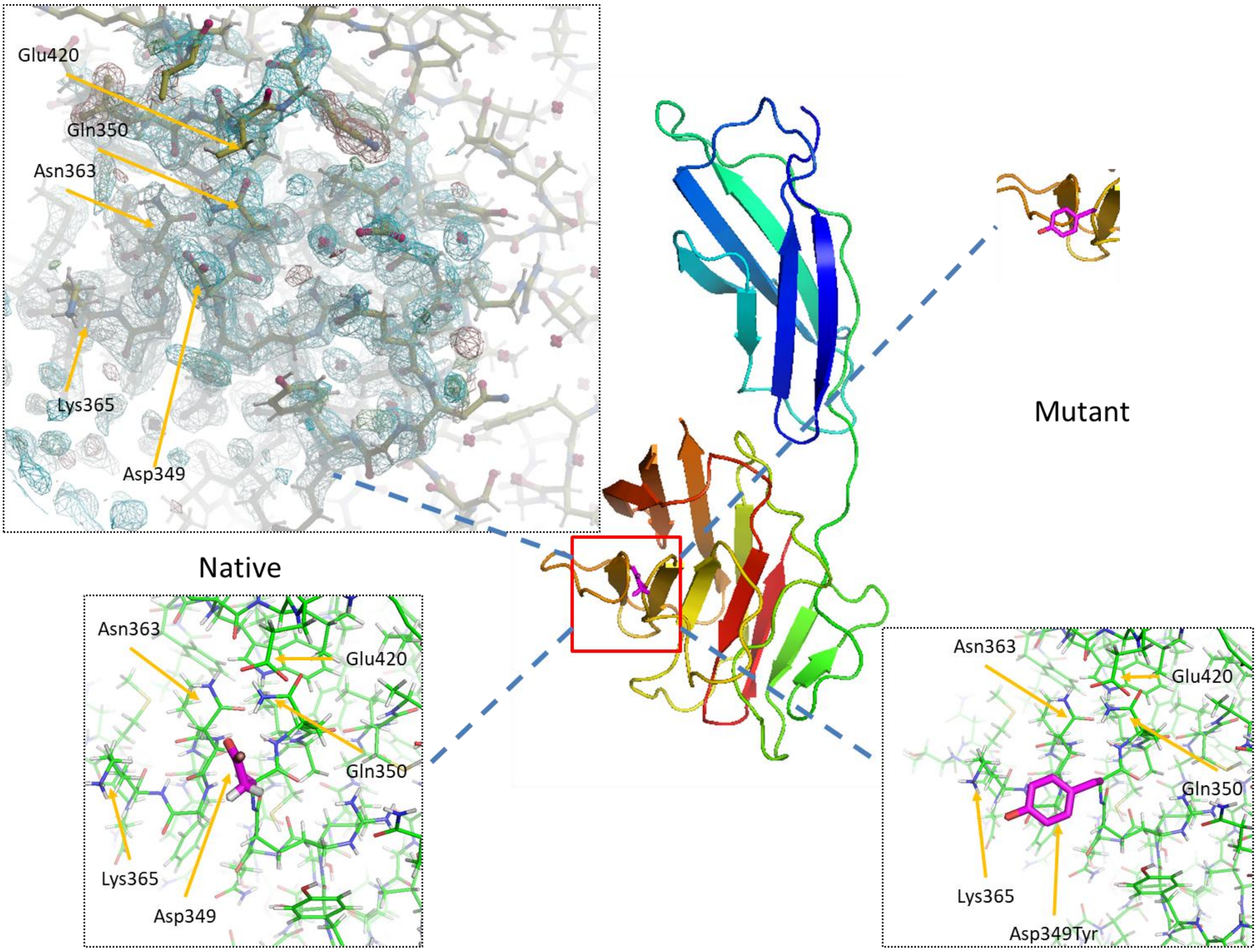


Figure 2-30: Mutation Asp349Tyr: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This mutation shows a loss of interaction with Gln350, leading to destabilisation of the protein. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

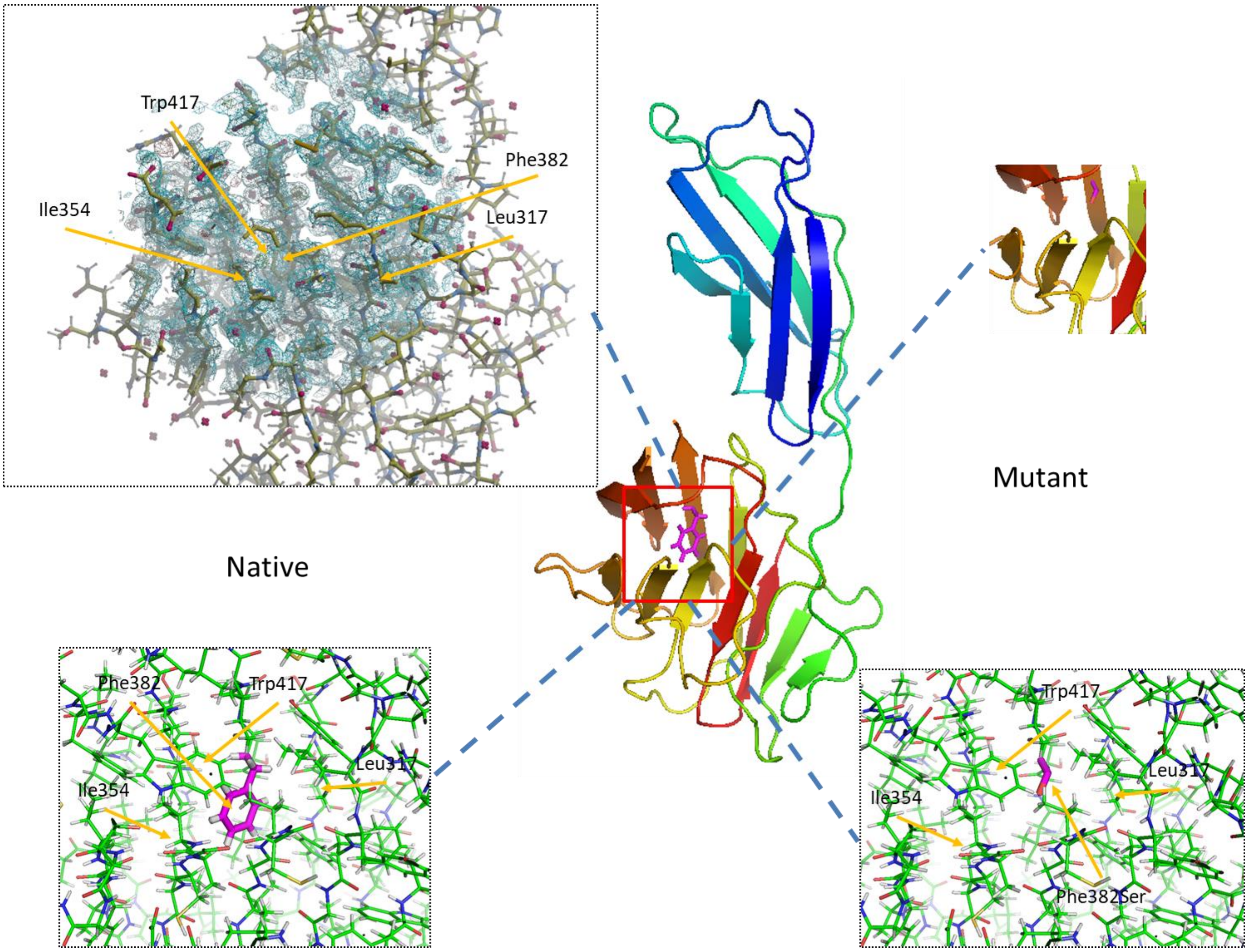


Figure 2-31: Mutation Phe382Ser: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This mutation causes a loss of hydrophobic interactions with Val335, Ile354, Leu317 and Val380 in the core of the protein. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

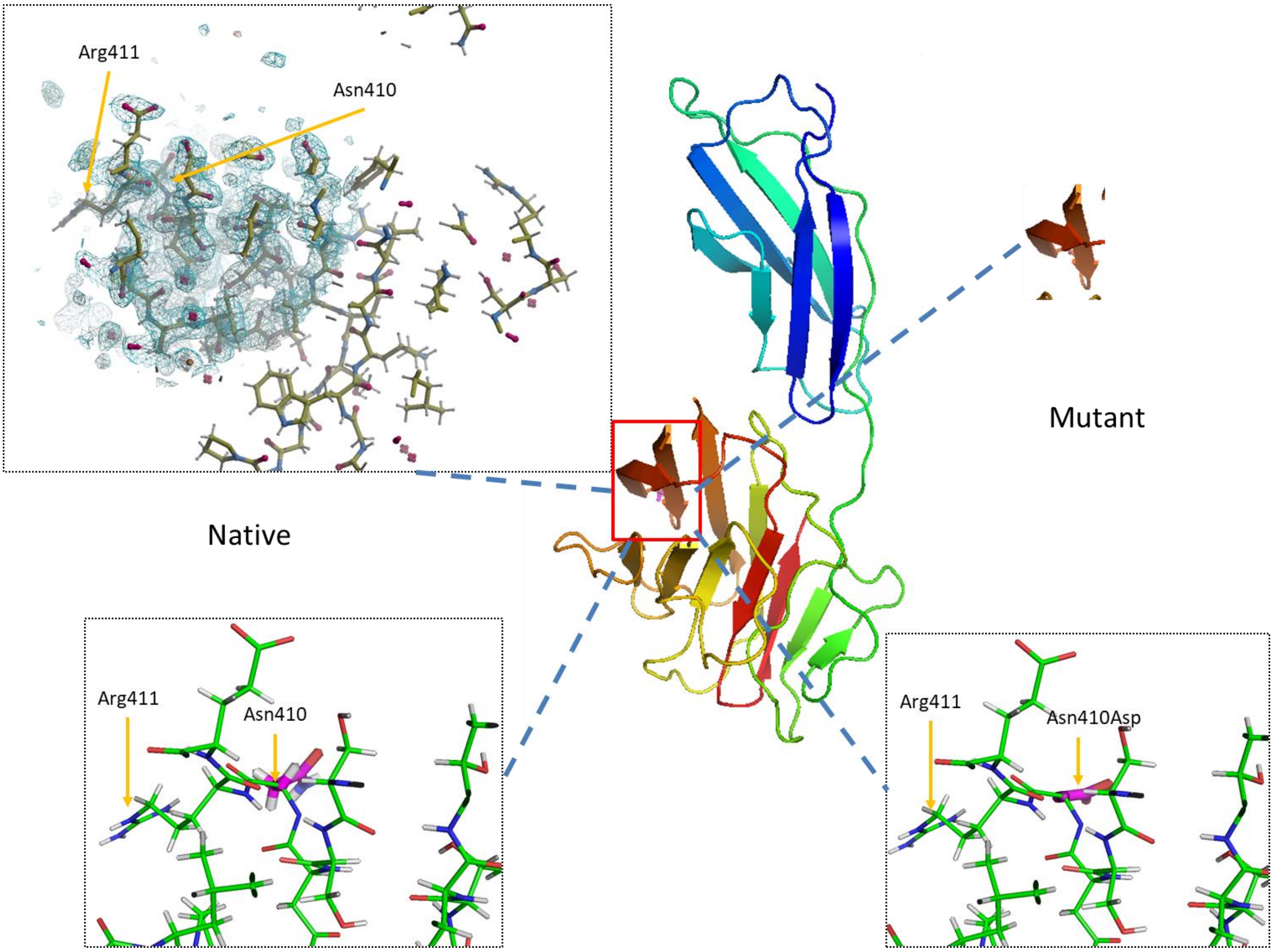


Figure 2-32: Mutation Asn410Asp: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). Here we see a completely exposed residue, which can accommodate the mutation completely. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

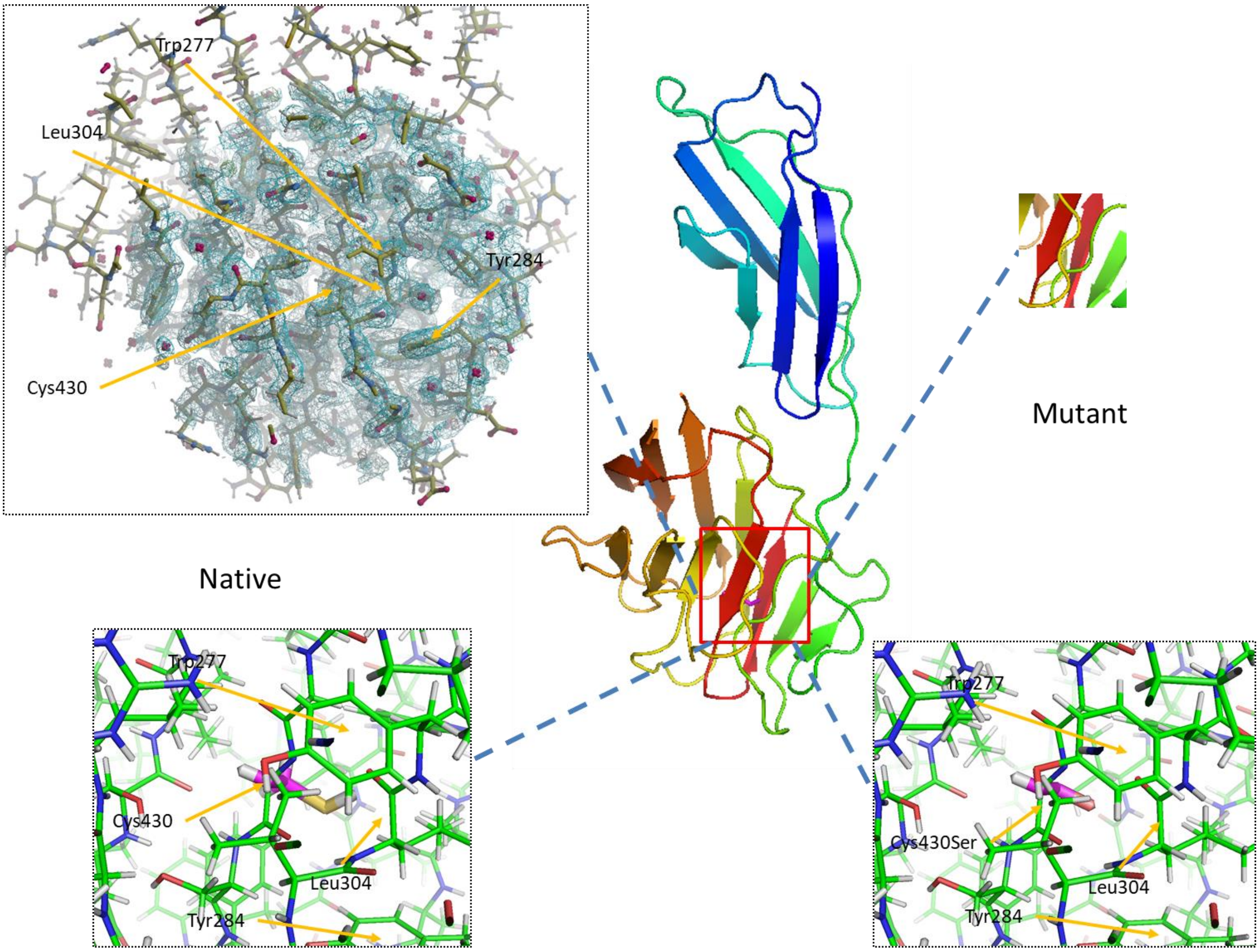


Figure 2-33: Mutation Cys430Ser: Location, native structure electron density and model of mutant

The images are described in a previous legend (Figure 2-17). This is a highly buried hydrophobic residue; the polar residue in this area will stabilise denatured state, so protein is destabilised by the mutation. The electron density map was generated in *Coot*[23] and *RASTER3D*[24], and represents the electron density at the 1σ level, and the difference density at the 3σ level. The stick models and cartoon representation were generated using *PyMOL*[261], with *Microsoft PowerPoint* used to label the figures.

Another measure that was used to determine the importance of a mutation was the solvent accessible surface area of each residue on the protein. This is a way of quantifying how exposed each residue is to the solvent. A low value suggests that the residue is buried deep within the protein. A high score conversely implies that the residue is exposed to the solvent. This score therefore gives an indication of the position of the residue in the folded protein.

This measure is important as the different positioning of the mutation can have quite different effects on the protein. A buried residue is deep within a protein, usually in a hydrophobic core. There are hydrophobic interactions with other hydrophobic residues, which stabilise the native state of the protein. In the protein's denatured state, the solvent forms unfavourable fixed networks around the hydrophobic residues, which reduces the entropy of the solvent molecules. This destabilises the denatured state, hence increasing the stability of the protein, as the degree of stability is proportional to the Gibbs' free energy difference between the native and denatured state.

A good example of a mutation causing the loss of hydrophobic effect in CRLF3 is Phe382Ser mutation (Figure 2-31). The phenylalanine residue is large and hydrophobic, found deep within the protein, forming good hydrophobic interactions in the native state and decreasing the entropy of the denatured state. When this residue is mutated to serine, if there is no reorganisation, a large hole would emerge which would be energetically unfavourable. If this mutant of the protein were able to fold it is likely that there would be either a large reorganisation or water trapped in the hole with little entropy. Both of these will be energetically unfavourable and so, as the native state is higher in energy, the protein stability is reduced. In this mutation of the phenylalanine residue to serine – a small polar group - the new residue interacts well with solvent in the denatured state. In the native state it either interacts with a trapped water molecule or it would not be able to form a hydrogen bond. This is therefore considerably unfavourable. Thus the buried nature of this residue, combined with the zero matches in ExAC, shows that the Phe382Ser mutation is highly likely to lead to the destabilisation of the protein. It therefore might explain the bleeding and platelet disease seen in the patient.

The mutation of a buried hydrophobic residue to a polar residue is further seen with Cys430Ser (Figure 2-33). Here the cysteine is in a hydrophobic pocket fully buried within the protein and the mutation to the polar serine will not be favourable. This could explain why a platelet disease is seen in this case.

There are other ways in which a mutation to a buried residue in a protein might lead to the protein not being able to carry out its function. If a large and bulky residue is inserted in the protein in place of a small residue, it might be prevented from folding because of steric clashes. In CRLF3, Thr323Ile (Figure 2-29), even though it is an exposed residue, might have this problem, though this is unclear from a first examination of the structure.

It should be added at this point that, whilst there are no examples in this particular set of mutations for CRLF3, if there is a mutation leading to charge buried within the protein, this will highly destabilise the protein. However, such a mutation is likely to be extremely damaging to the viability of the person, and so are unlikely to be seen.

Next, there are those mutations that are exposed on the surface. These mutations tend to be more difficult to interpret. Usually, mutations here can be tolerated more, as there is space allowing for the change in residues. An example is Asn410Asp (Figure 2-32), which has enough free volume to accommodate the mutation, as shown by the large number of hits in ExAC. However, some are still unfavourable due to the steric clashes that they may cause (e.g. Thr323Ile, which may clash with the residues around it).

However, there are some exposed residues which seem to be critical because of the side chain interactions. Asp349Tyr (Figure 2-30) is an excellent example. Whilst it is completely exposed, one of the carboxylate oxygens on the aspartate interacts with the N-H on Gln350. Given there are no mutations of this residue in ExAC, this suggests that it is indeed important for the correct folding of the protein into a functional form, so might lead to clotting disease if mutated.

There are other mutations which would be expected to cause wholesale disruption of the protein, such as Cys313Phe (Figure 2-28). Here the disulphide bridge is completely disrupted and blocked with a large bulky phenyl ring. This will either cause the incorrect folding of the protein or no folding at all. Thus this will affect the functioning of this protein and so could be a cause of the clotting disease found.

The stability of mutations of the protein can be inferred from stearic clashes or consideration of the interaction with neighbouring sidechains. However, the effect of the mutation on the overall structure including the part of the protein that was not crystallised cannot be estimated from this work. Interactions with other proteins will occur through surface sidechains, and these interactions may occur in regions that appear to have plenty of available space for larger residues, but which are constrained by the partner protein. As an example, the Thr323Ile mutation may be in the binding site of another protein and hence might cause disease.

Hence it is clear that knowledge of the structure around a residue is useful in understanding the effect of the mutations. In this section the use of knowledge of the structural model, along with other bioinformatics (ExAC), to explain which mutations are likely to be relevant to the causes of clotting diseases has been explored.

2.11 Overall conclusions

A successful model of the second section of the protein (residues 174-442) was generated by using experimental phasing of a crystal of one of the four constructs that were proposed and investigated. Two slightly different fully-refined models were derived from the different crystals, with different statistics. The CRLF3 Construct 3 final model 1 was used for the mutation studies, with the future aim to deposit the structure in the PDB, leading to a publication, pending further discussions with our collaborators. Crystallisation techniques were observed in the initial trials, and the challenges in crystallising this protein or some of the proposed constructs were considered from knowledge of the possible protein structure.

S-SAD did not provide enough information about phases to allow determination of the structure, as the anomalous signal was too weak. The molecular replacement (MR) analysis of the data was unsuccessful, as the MR models did not superimpose well enough with the CRLF3 Construct 3 structure.

Naturally occurring mutations of the CRLF3 protein were considered in the light of the structure that was generated. The BRIDGE consortium provided information on the single point mutations that led to blood clotting disorders, and these were analysed with reference to the proposed structure. Some mutations cause a clear structural effect and can be understood in the light of destabilising the protein, others do not, suggesting interactions with other molecules or the parts of the protein that were not modelled.

Throughout this chapter various different computational techniques have been applied to the problem of gaining structural insight into CRLF3. A number of techniques were investigated, and an understanding of the use of *phaser*[4] and *Phenix*[226] was gained and exploited in other parts of the research reported in this thesis.

Chapter 3 <u>NaCelleS</u>

3.1 <u>Rationale</u>

This project was developed to improve the identification and location of non-crystallographic symmetry (NCS) in density maps, with initial application in X-ray diffraction. In poor electron density maps, if the NCS can be discovered, then regions related by NCS can be averaged together, enhancing the signal-to-noise ratio, and increasing the chances of building a successful initial model. If this process can be improved, the discovery of solutions to difficult crystallographic structures should be facilitated.

To achieve this aim, ways to improve the program *phenix.find_ncs_from_density*[2] using *phaser*[4] were explored. The novel program, developed in this work, which exploits some of the capabilities within *phaser*[4], has been named *NaCelleS*. The underpinning hypothesis driving this development is that specific volumes of electron density within the protein can be exploited to identify the protein, and hence be used to locate other positions of this protein within the unit cell. If a sphere of electron density distribution within the protein is cut out from the input electron density map, then locating another similar sphere with the same electron density can be used to identify another copy of the protein and define the NCS operators, providing that there is NCS in the crystal. Sufficient signal-to-noise ratio to allow this independent location to be found was assumed.

Why would this be an improvement on the current approach? The original *phenix.find_ncs_from_density*[2] program attempts to identify NCS using a single sphere cut out around a position. The coordinates of this position are the highest scoring values given by the protocol *phenix.guess_molecular_centres*[2], which returns a list of points that are most likely to be inside the protein, given an electron density map. This likelihood is determined as follows. *phenix.guess_molecular_centres*[2] assumes that in solvent, there are relatively few peaks in the electron density map, but within a region of protein, there are many features. Therefore it looks through the map and decides which regions have the most variation in electron density and therefore features, and then reports back a list of these centres in descending order of RMS feature density. *phenix.find_ncs_from_density*[2] then takes a sphere of defined size (typically 10Å) based on the most likely position, and tries every possible orientation within the asymmetric unit. It uses a 20° grid through the rotations in three dimensions, and runs a phased translation function, looking for any peaks that represent matches of the electron density distribution within the test sphere[1]. A list of NCS operators is reported at the end[2].

There are, however, a variety of ways in which this process could potentially be improved.

Firstly, only a coarse range of rotations is used – a grid of every 20°. Second, there is the problem that, in order to run in a reasonable amount of computer time, *phenix.find_ncs_from_density*[2] restricts the resolution to 4 Å. This, therefore, is not using all the data that are available, and this can lead to a loss of information in maps offering higher resolution. Both of these problems can be overcome by using *phaser*[4] and exploiting its features in a new way.

In the program *NaCelleS*, which uses *phaser*[4], these restrictions can be largely removed and the analyses can be improved, as *phaser*[4] can operate with a finer rotation grid (appropriate to the resolution), and at a higher resolution, if such data are available. By using *phaser*[4] initially to generate a list of likely rotations, it reduces the number of rotations that have to be used for separate phased translation functions. With this highly reduced list, *phaser*[4] can then make an intelligent choice as to how much resolution is needed in order to get an unambiguous signal. Given that far fewer rotations are needed compared to *phenix.find_ncs_from_density*[2], a greater time can be spent on each individual run in the phased translation search, so a higher data resolution can be used. This should improve the signal-to-noise ratio and permit more accurate solutions.

[1] It should be noted that Dr. Tom Terwilliger, at the Los Alamos National Laboratory, USA, has added a new parameter in *phenix.find_ncs_from_density*[2], called *n_center_use*, which then repeats the calculation for other centres (unpublished data, but stated in documentation for program) since this work was done.

Another new concept developed within this project is based on the idea of iteration. With *phenix.find_ncs_from_density*[2], the output either yields operators, or returns none. There is no way of using a partial subset of the operators to find further operators. Therefore a further question was postulated as to whether using knowledge of some of the NCS operators, it is possible to determine further NCS operators.

Finally, an interesting idea arose that was implemented using *phaser*[4] for the case where a list of anomalous scatterer sites exists within the asymmetric unit. In certain instances, a single-sphere approach does not work if the initial estimate of phases was too poor (the electron density map is poor). The resulting signal-to-noise ratio is not sufficient to determine the NCS operators. In this case, heavy atom sites (or anomalous scatterers) can be used to improve this ratio. After cutting out the first sphere as in *phenix.find_ncs_from_density*[2], a second sphere containing the equivalent electron density for a second site is generated. This is inserted into a new map of equivalent size, which is scanned as before by the first sphere. *phaser*[4] will find the best fit solution at the coordinate of the second sphere during the scan, as the rest of the map is completely empty and flat. The resultant Translation Function Z-score (TFZ) score will be very high if the second sphere is very similar to (or the same as) the first sphere (same environment) or lower if the second sphere has captured a scatterer in a different position in the protein, and this procedure is then repeated for every site. The TFZ for a search peak is defined as the log likelihood gain (LLG) of the particular translation and rotation minus the mean of a set of random translations and rotations, divided by the RMS deviation of the random set. The LLG is defined as the difference between the likelihood of the model and the likelihood calculated from a Wilson distribution.

3.2 Algorithms implemented within *NaCelleS*

NaCelleS itself offers two different modes of operation – a one-sphere approach and a two-sphere approach. The former is originally based on the operation of *phenix.find_ncs_from_density*[2], but has been significantly enhanced, and the latter is a new algorithm. A further, new iterative mode of operation using these two core algorithms was also investigated.

3.2.1 NaCelleS one-sphere approach

The *NaCelleS* one-sphere approach has many similarities to the method used in *phenix.find_ncs_from_density*[2], but offers significant enhancements (Figure 3-1). As in the reference program, this algorithm selects a centre from either the most likely result of *phenix.guess_molecular_centers*[2] or from a list of heavy atoms or anomalous scatterer sites. A similar sphere of electron density with a 10 Å radius is cut out around the centre using *phenix.cut_out_density*[226]. In this case, the "for_phaser" option is activated in calling this program to optimise its output for use in *phaser*[4] in the following ways: 1) electron density remains based at the original coordinates as *phaser*[4] is optimised to use this coordinate set, 2) the sphere of electron density is placed in a unit cell three times larger in each linear dimension than the sphere. This allows the sampling in reciprocal space to be fine enough for correct linear interpolation of the structure factors[137], 3) other options which allow improved *phaser*[4] outcomes, such as creating a map that allows *phaser*[4] to use the full resolution of the original data, were set. In maps not generated for *phaser*[4], as used by the reference program, the high resolution is slightly truncated to prevent series termination errors but this causes signal loss. *phaser*[4] manages termination errors associated with series approximations internally, and so the high resolution limit is set internally to allow the molecular transform to be interpolated to the resolution of the original map. Thus all the available resolution is used in this approach, unlike in *phenix.find_ncs_from_density*[2].

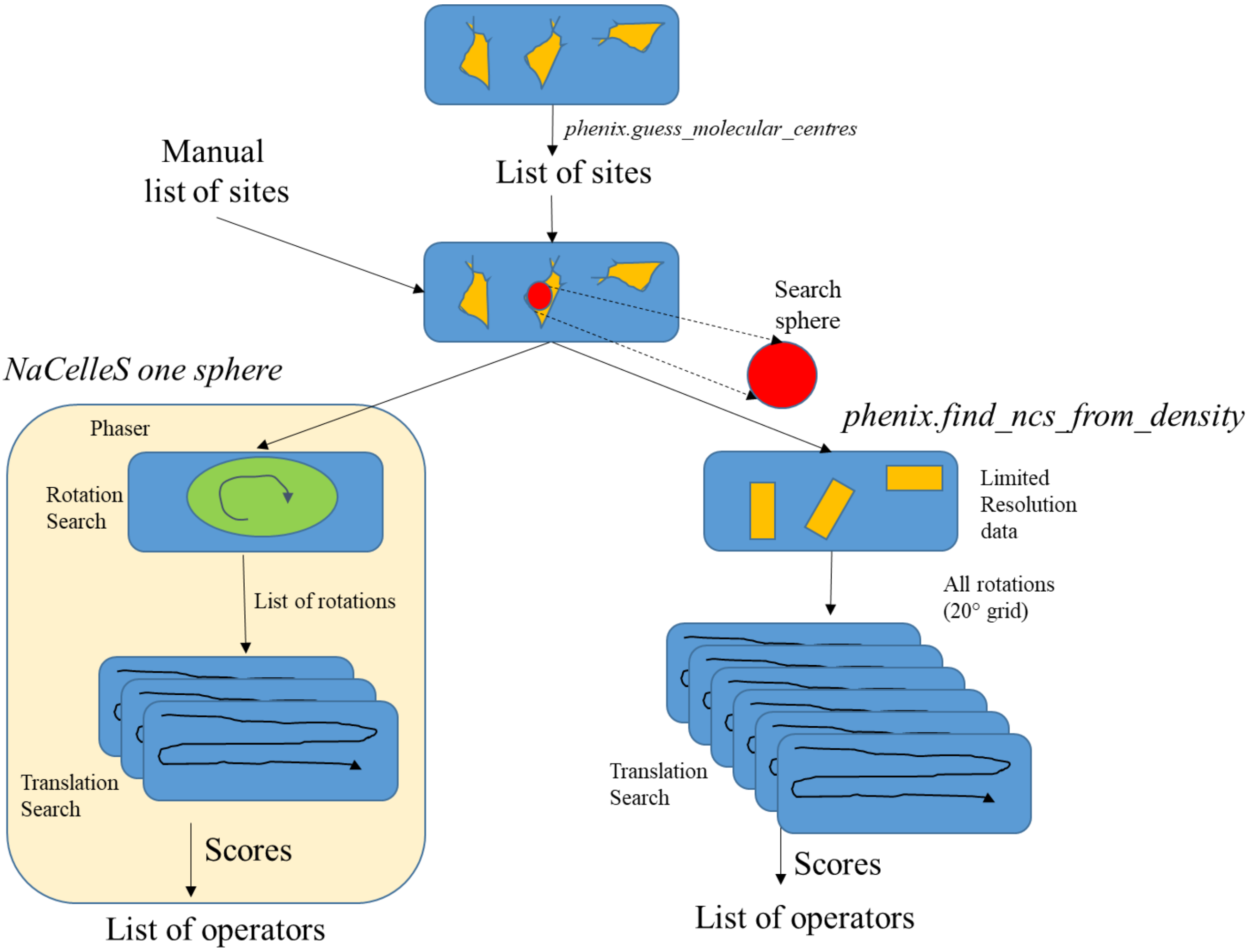


Figure 3-1: One-sphere based algorithms

Flow chart generated using *Microsoft PowerPoint*. All searches were done in reciprocal space.

An initial rotation search is then done with this sphere against the entire map at the full available resolution, generating a list of orientations for use in the subsequent phased translation search, which now also uses the full resolution of the data. Within *phaser*[4], the process is iterated, as after finding each NCS operator, *phaser*[4] takes into account this solution in subsequent searches, leading to an increase in the signal of the rotation function. The number of NCS copies needs to be provided to *NaCelleS*. A list of NCS operators is the end product, and TFZ, LLG scores and elapsed time are reported. The list of NCS operators was then processed in *phenix.find_ncs*, and the overall NCS cross-correlation ρ value recorded. ρ values can range between -1 and +1, although values below zero are extremely rare. The higher the value, the higher the confidence in the NCS operators that have been determined.

3.2.2 NaCelleS two-sphere approach

The *NaCelleS* two-sphere approach uses the centres of heavy atoms or the anomalous scatterer sites within the asymmetric unit (Figure 3-2). One site is taken, and a sphere with a 10 Å radius is cut out of the electron density map, as before. This search sphere is then inserted in a blank map using *phenix.cut_out_density*, at exactly the same spatial coordinates. A second scatterer site is then taken, and a similar sphere of electron density is cut out in the same manner, and inserted into a further blank map. The first map is used as the search model within the second map, and both LLG and TFZ scores are reported. A second map is then created with a sphere from the next scatterer site and the search repeated, until all sites, including the search site itself, are scanned. The results define potential NCS operators, which are ranked according to the associated TFZ. After cycling the second map through all of the possible permutations, the peaks are plotted and analysed in Microsoft Excel 2010.

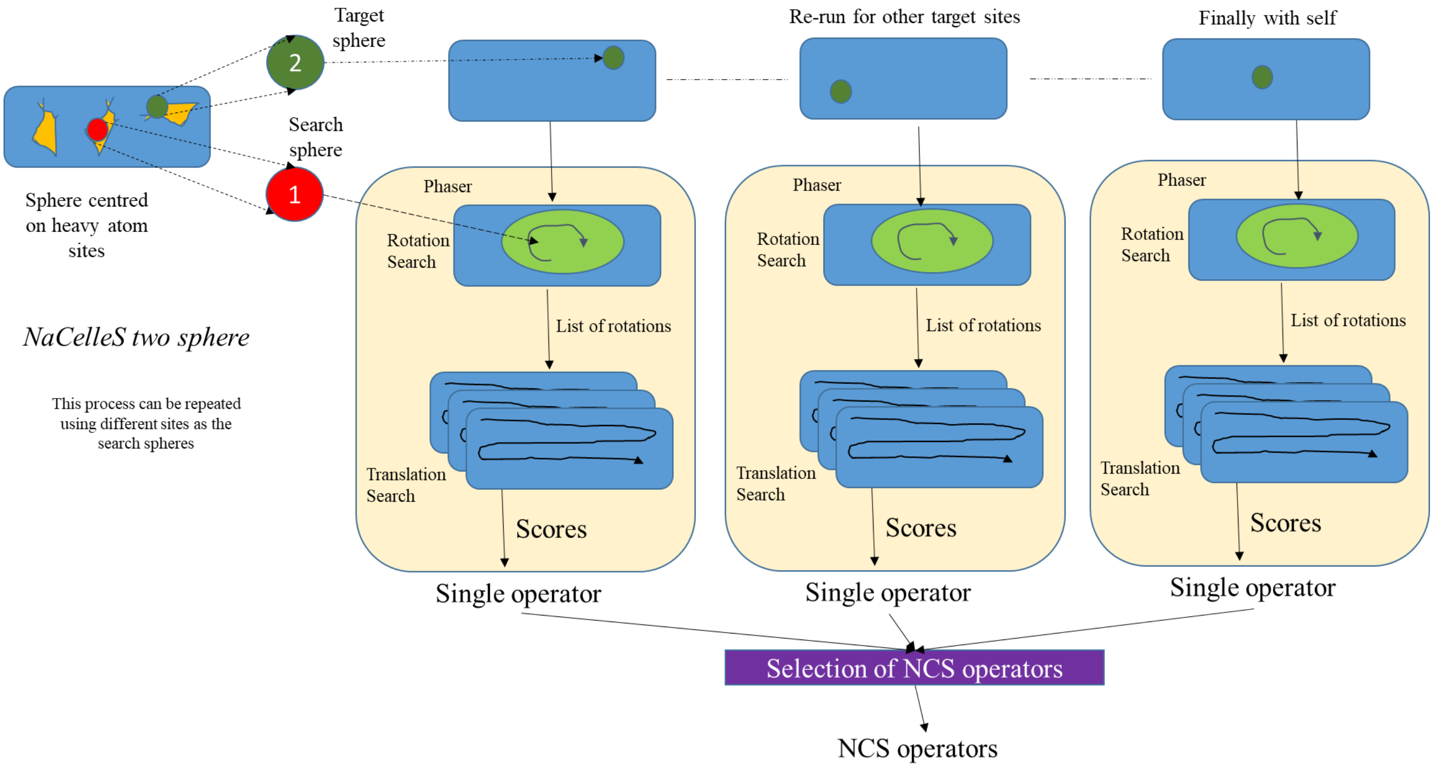


Figure 3-2: NaCelleS two-sphere algorithm

Flow chart generated using *Microsoft PowerPoint.* All searches were done in reciprocal space.

3.2.3 Iterative approach

A novel iterative approach was then developed to take advantage of the features available in *phaser*[4], and the approaches detailed above (Figure 3-3). The NCS operators output from either the one or two-sphere approaches in *NaCelleS* are selected for use in this iteration as follows. In the two-sphere approach, the top *n* operators are taken, where *n* is the number of NCS operators expected. In the one-sphere approach, *phaser*[4] can return one or more self-consistent but different solutions, each of which can have up to the requested number of operators. The most likely solution, with as many operators as were found, is taken. These operators are then input into *phenix.find_ncs*[226], which uses them to produce a map that contains only one monomer of electron density in the asymmetric unit, with flat, featureless density elsewhere. This monomer map has been generated by averaging the electron densities at equivalent regions defined by the NCS operators and is called the NCS-averaged map throughout the chapter.

From this map, a sphere is cut out using *phenix.cut_out_density* so as to cover the entirety of a single copy of the NCS-averaged monomer, and *phenix.cut_out_density* uses the “for_phaser” modifications mentioned above. This is then used as an input model to reiterate *NaCelleS* one-sphere approach. Iterations beyond the second showed little additional improvement but consumed significant processing time, and improvements are required in the generation of the averaged map for additional runs to add significant value (Section 5.2.2).

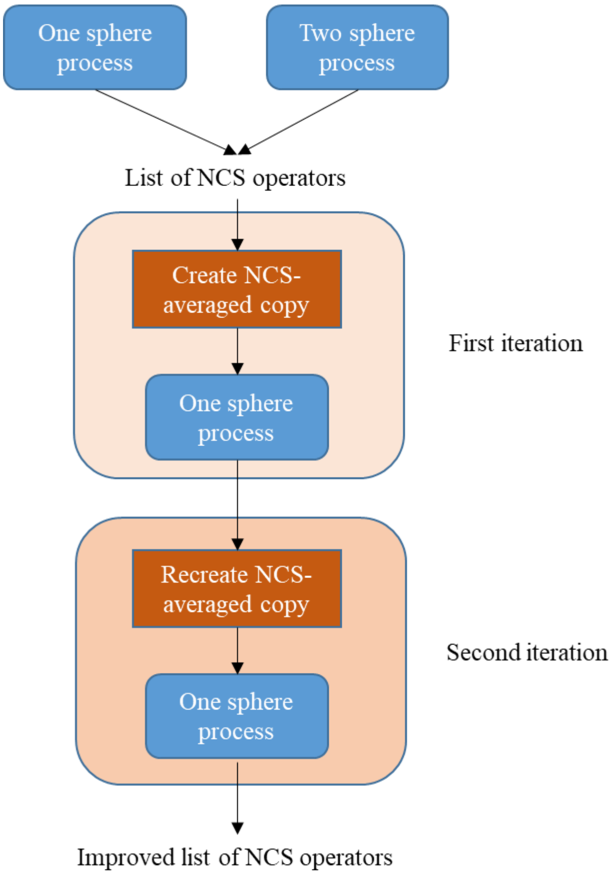


Figure 3-3: NaCelleS iterative algorithm

Flow chart generated using *Microsoft PowerPoint.* All searches were done in reciprocal space.

3.3 Developing Test data

3.3.1 Initial Test Data

When first setting up this project, two test cases were selected to develop the *NaCelleS* software. This selection was done based on an assessment of the test data used in the development of *phenix.find_ncs_from_density*[2], and another data set where the finding of the NCS operators had been challenging.

The first test case, with a PDB code of 1N0E[273] is a protein called UPF0040, space group P2$_1$, unit cell dimensions 78.627 Å, 100.064 Å, 100.851 Å, with angles 90°, 93.30°, 90°, sequence length of 166 amino acids and predicted mass of 19.3 kDa per chain. It is believed to have a role in cellular function of bacterial cell division in *Mycoplasma pneumoniae*. There are eight monomers in the asymmetric unit, which form an eight-fold point group around an axis. However, this axis is only consistent on a local scale, hence it exhibits non-crystallographic symmetry. The data have a high resolution limit of 2.7 Å (Figure 3-4), and were also used in the original *phenix.find_ncs_from_density* paper[2].

The second test case, called GERE, with a PDB code of 1FSE, is a more challenging example. The space group is C2, the unit cell dimensions are 109.019 Å, 61.749 Å, 71.743 Å, with angles of 90°, 97.08°, 90°. The size of the protein is 74 amino acids, with a mass of 8.6 kDa per chain. These data are to a resolution of 2.05 Å, and describe a crystal structure of a DNA-binding protein that regulates gene transcription in the late stages of sporulation in *Bacillus subtilis*[274]. It contains six monomers in the asymmetric unit, formed of three dimers, although the dimers are not quite identical (Figure 3-5).

Both of these data sets contain two anomalous scatterers per monomer, resulting in sixteen sites for 1N0E, and twelve for 1FSE. Conventional approaches to finding the NCS using only the relationship between these sites will struggle because geometrical approaches require three or more sites per monomer. However, the two-sphere approach of *NaCelleS* is well suited to this challenge, and should succeed in identifying the NCS relationships. These are therefore highly useful datasets upon which to train *NaCelleS*.

3.3.2 Extended Test Data

After using the two test sets above to set the appropriate parameters for *NaCelleS*, the program was tested on a large variety of cases. These trial data were selected from the variety of test cases that have been collected by Prof. Paul Adams, Department of Bioengineering, University of California in Berkeley, USA, and that have been used to help develop *phenix*[226]. The full list of test cases that were used to explore the operation of the new program can be seen in Table 3-1.

3.4 Assessing Results

A number of measures to evaluate success were explored (Section 3.6.3 and Section 3.8). Firstly, the number of operators found was returned, as *phaser*[4] usually managed to find the number of solutions that was specified, even if they were wrong solutions. The next measure was the correlation ρ value recorded when the NCS operators were placed into *phenix.find_ncs*[226] and a single asymmetric averaged NCS copy was created and correlated with the individual monomer spheres around the sites (Section 3.7.2). This is directly equivalent to the correlation ρ value that comes out of *phenix.find_ncs_from_density*[2], and so is a very helpful indicator.

The correlation ρ value obtained gives a good idea of the correctness of the operators. If an NCS-averaged single monomer map has a very low correlation score (below 0.2), then it is highly likely to be incorrect. *phenix.find_ncs*[226] uses 0.4 as the usual cut-off for this decision, but a slightly less strict limit of 0.2 can be useful. If an average electron density map cannot be created (internal check within *phenix.find_ncs*), then a correlation value of <0.01 is returned.

The other measures are the TFZ score and the LLG scores.

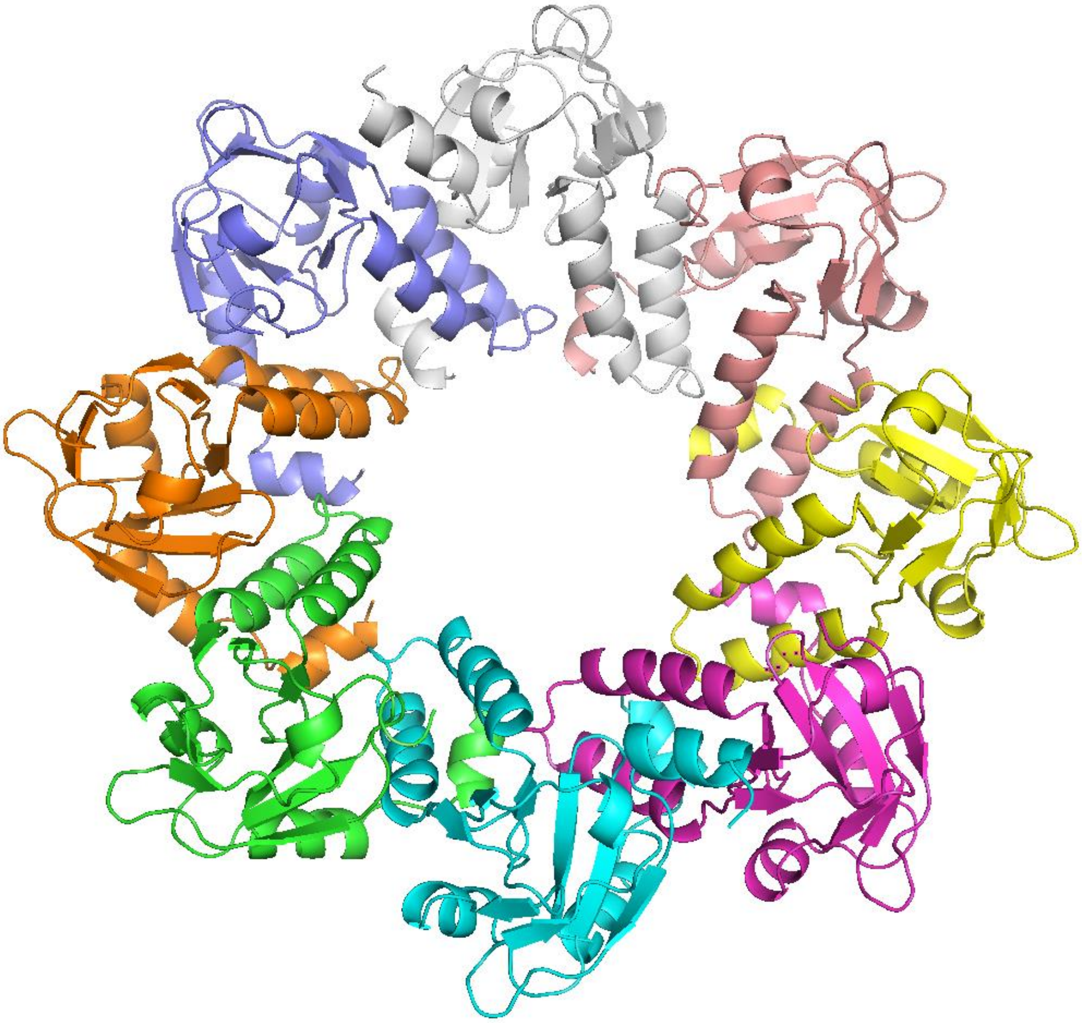


Figure 3-4: Crystallographic structures of the protein 1N0E

There are eight non-crystallographic symmetry copies in a unit cell in 1N0E[273], forming an eight fold axis. Different colours are used for the different chains that exist. Images were created using *Pymol*[261].

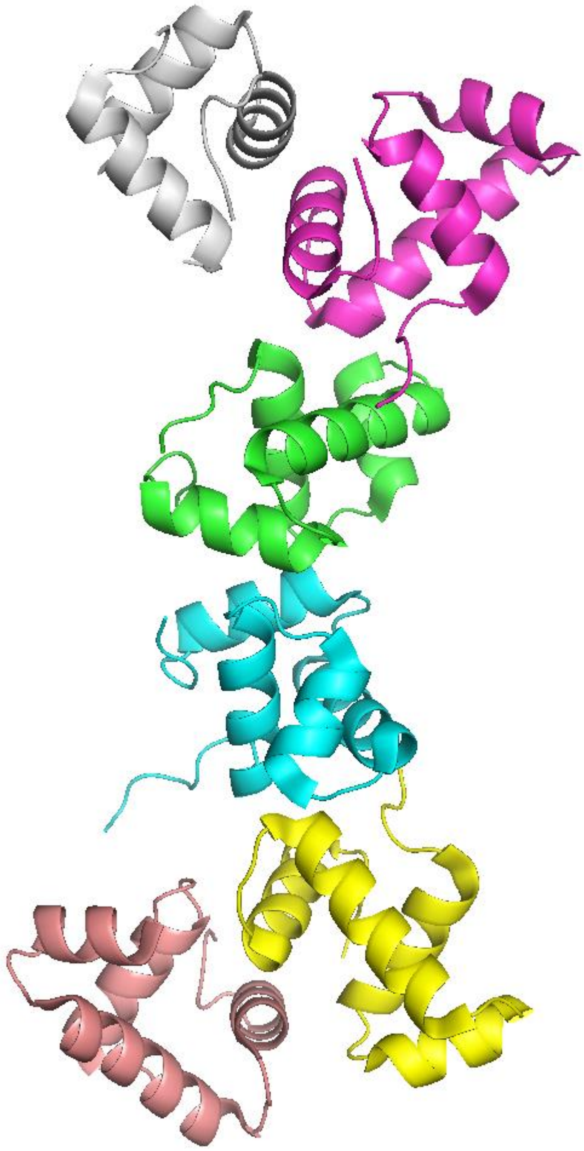


Figure 3-5 Crystallographic structures of the protein 1FSE

There are six non-crystallographic symmetry copies in 1FSE[274], but they are not in a point group[274]. Different colours show the different chains. Images were created using *Pymol*[261].

Name	PDB ID	Dmin (Å)	Space Group	Heavy atom type	Number of monomers in asymmetric unit
1038B[226]	1LQL	3	P212121	Se	10
1071B[275]	1NF2	3	C2	Se	3
cp-synthase[276]	1L1E	2.8	P65	Se	2
s-hydrolase[277]	1A7A	2.8	C222	Se	2
ut-synthase[278]	1E8C	2.8	C2221	Se	2
aep-transaminase[279]	1M32	2.6	P21	Se	6
rab3a[280]	1ZBD	2.6	C2	Se/Zn	1
p32[281]	1P32	2.5	P21	Se	3
cyanase[282]	1DW9	2.4	P1	Se	10
fusion-complex[283]	1SFC	2.4	I222	Se/I	3
vmp[284]	1L8W	2.3	P2	Se	4
gpatase[285]	1ECF	2.25	C2221	Se	2
pdz[286]	1KWA	2.1	C2221	Se	2
1167B[287]	1S12	2	P21	Se	4
synapsin[288]	1AUV	2	P3221	Se	2
ic-lyase[289]	1F61	1.8	P6522	Se	2
mbp[290]	1YTT	1.8	P212121	Yb	2

Table 3-1 List of all test data sets for NaCelleS used in the more extended test case

3.5 One-sphere approach

3.5.1 One-sphere approach using the same centre as phenix.find_ncs_from_density

A direct emulation of *phenix.find_ncs_from_density*[2] was undertaken to investigate possible weaknesses and potential improvements in the methods used in this program. In this case, using the most likely centre predicted from *phenix.guess_molecular_centres*[2], a sphere of electron density was cut out around the centre, and this sphere was then used to search the entire map.

All eight copies of the monomer were found for the 1N0E test case (Figure 3-6). In the 1FSE case, all six monomers were also found (Figure 3-7). The approach taken in *NaCelleS* is successful, but *phenix.find_ncs_from_density*[2] works equally well, also finding all the NCS copies for both of these initial cases when using the *phenix.guess_molecular_centres*[2] centres.

3.5.2 One-sphere approach using the anomalous scatterer sites

To investigate performance under more difficult conditions, the starting site for the electron density to be cut out was set to be the first anomalous scatterer site in the list of anomalous scatterers. These sites tend to be found on the edges of proteins and a search sphere is likely to include significant amounts of solvent, and hence will contain less signal. Using these sites will permit a more challenging comparison of the performance of *NaCelleS* and *phenix.find_ncs_from_density*[2], and an assessment of progress in achieving the aim of this work.

In this case, *phenix.find_ncs_from_density*[2] works poorly, finding the original position and two other NCS copies in the 1N0E case, and the original and one other NCS copy in the 1FSE case. The initial run using *NaCelleS* resulted in a similar limited outcome as *phenix.find_ncs_from_density*[2] – in addition to finding the original site of the electron density, in 1N0E four other copies were found, and in 1FSE one other copy was found (Figure 3-8).

Further work highlighted that the limitation in identifying the further sites stemmed from setting too great a restriction on the list of rotations. Accordingly, *NaCelleS* was rerun using the anomalous scatterer site, but a longer list of rotations was kept from the rotation search by setting the keyword "PACK CUTO 40". This increases the percentage of allowed packing clashes to 40 % instead of 10 %. This improved the results considerably – for 1N0E, *NaCelleS* now manages to find all eight operators (Figure 3-9), and for 1FSE, *NaCelleS* manages to find four sites of six (Figure 3-10). This indicates that the restriction in *phaser*[4] where it does not use phase information in scoring rotations is significant, and that an improved scoring method would add value. However, the last two site were not found, even if every possible orientation of the sphere was used. This may be because this anomalous scatterer site was on the edge of the copy of the protein. This means that the sphere cut out around that centre could incorporate electron density from another copy of the protein, as the site on one copy might be very close in space to another anomalous scatterer site. The correct placing of the two spheres of electron density will exhibit a lot of overlap, so *phaser* rejects this placement on grounds of too much steric clash. As a result the correct NCS is not found.

3.5.3 One-sphere approach with extended test cases

The one-sphere approach was then exercised using a larger selection of test cases (Table 3-2). TFZ and LLG were assessed as scores of success with the one-sphere approach, but they were of limited value in predicting whether valid solutions were generated. For example: the rab3a second data set with Se shows a TFZ of 13.4 and an LLG of 83.1, but clearly is not solved (ρ = 0.14), whereas 1071B has a TFZ of 13.5 but gives a solution (ρ = 0.82), and cp synthase 3 has a low LLG of 46.6 but again has clearly been solved (ρ = 0.83). Therefore none of these measures seem to be predictive.

However, other cases in which the single-sphere approach in *NaCelleS* was applied with the less restricted rotation list (Section 3.5.2) were more encouraging. 1071B and cyanase (second data set) seem to solve well, identifying all the operators. This had not been achieved by the original *phenix.find_ncs_from_density*[2], showing significant improvement in outcome.

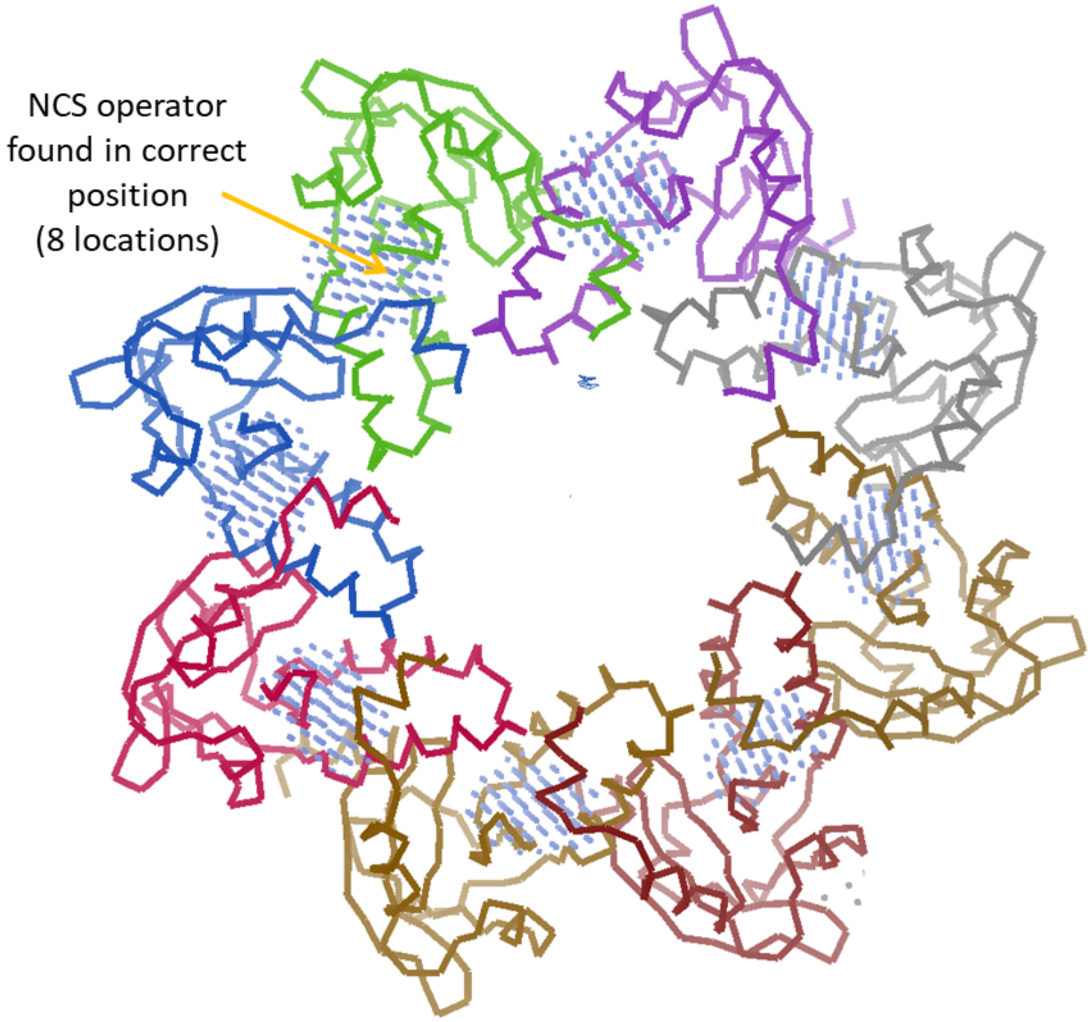
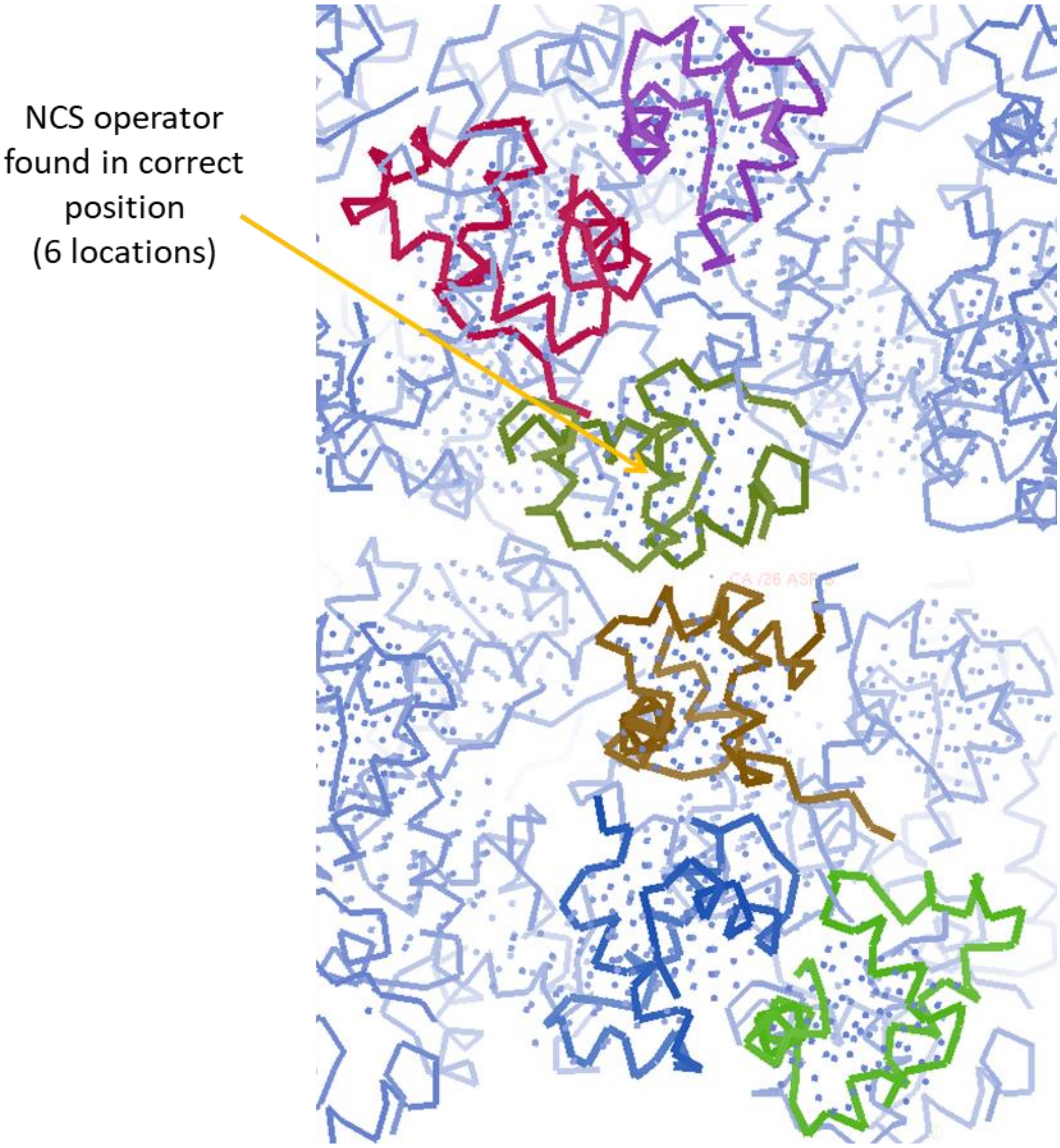


Figure 3-6: Visualisation of successful discovery of NCS in 1N0E using NaCelleS with phenix.guess_molecular_centers centres

The spheres of blue dots here indicate the position selected by *NaCelleS* to place the 10 Å sphere after searching the map. This is shown for the one-sphere run with *NaCelleS*, and is using the *phenix.guess_molecular_centers* centres. Every copy of the protein in the asymmetric unit has a correctly placed search sphere – there are eight in the 1N0E case. The rotation and translation operators are reported as the output of the program. The image was created using *Coot*[23].



NCS operator found in correct position (6 locations)

Figure 3-7 Visualisation of successful discovery of NCS in 1FSE using NaCelleS with phenix.guess_molecular_centers centres

The spheres of blue dots here indicate the position selected by *NaCelleS* to place the 10 Å sphere after searching the map. This is shown for the one-sphere run with *NaCelleS*, and is using the *phenix.guess_molecular_centers* centres. Every copy of the protein in the asymmetric unit has a correctly placed search sphere – six in the 1FSE case. The rotation and translation operators are reported as the output of the program. The image was created using *Coot*[23].

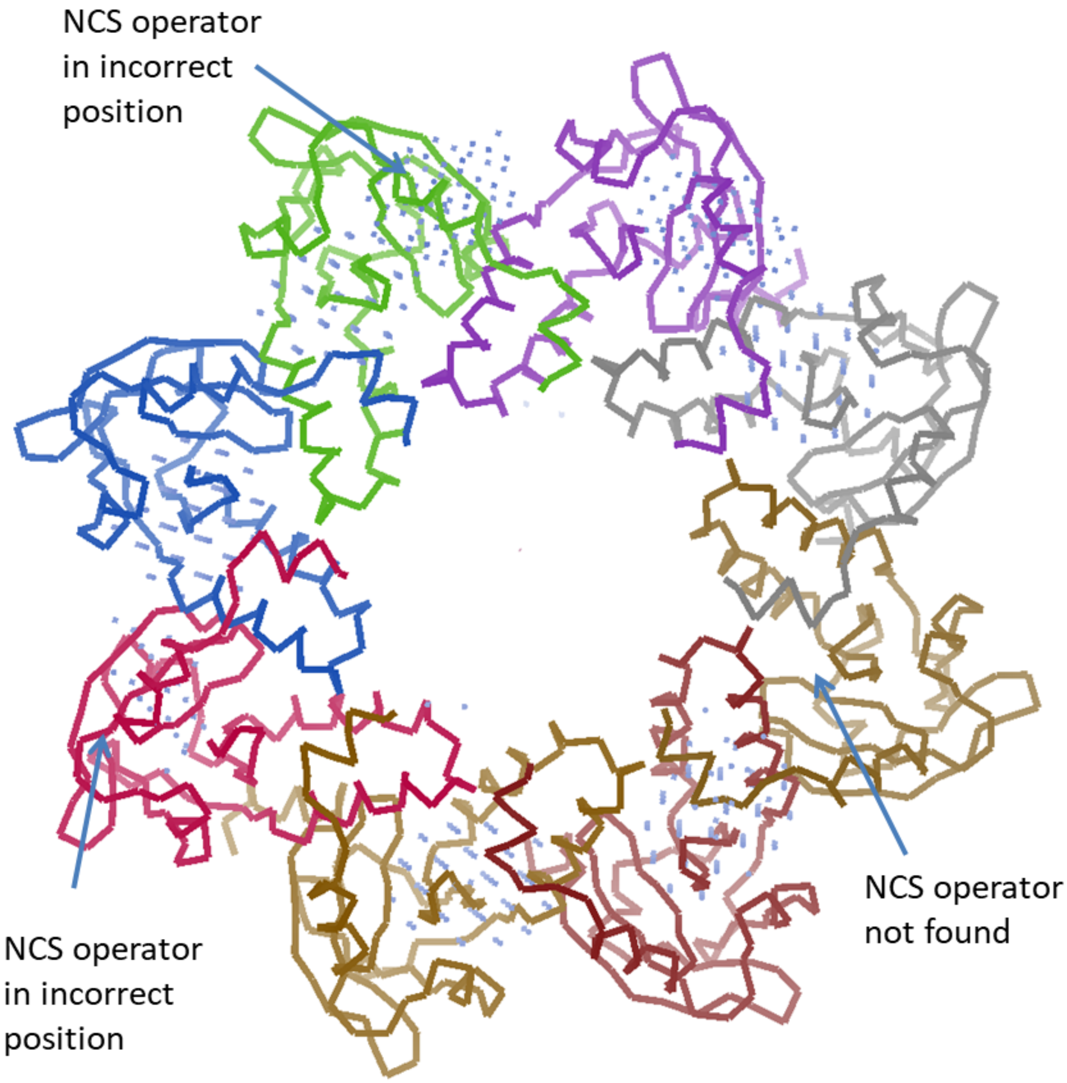


Figure 3-8 Failure to correctly locate NCS in 1N0E using NaCelleS with anomalous scatterer centres

The blue spheres here indicate the positions where *NaCelleS* has placed the 10 Å sphere after searching the map under different conditions. This time, the first anomalous scatterer site is used as the centre around which the search sphere of electron density is cut. The image was created using *Coot*[23].

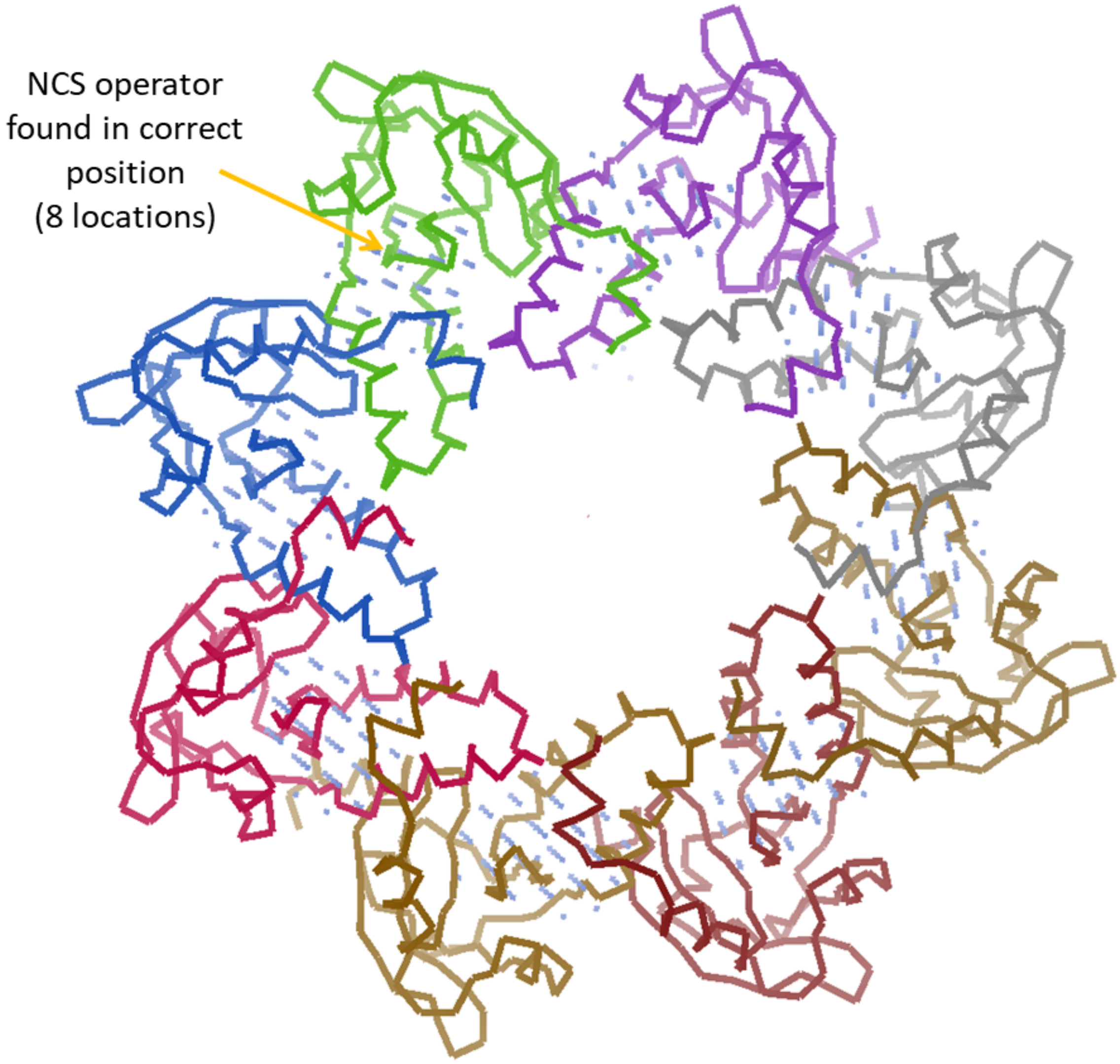


Figure 3-9: Correct location of NCS in 1N0E using NaCelleS with anomalous scatterer centres and expanded rotation list

The blue spheres here indicate the positions where *NaCelleS* has placed the 10 Å sphere after searching the map under different conditions. Again, the first anomalous scatterer site is used as the centre around which the search sphere of electron density is cut. A more extensive list of rotations was used in the phased translation function. The image was created using *Coot*[23].

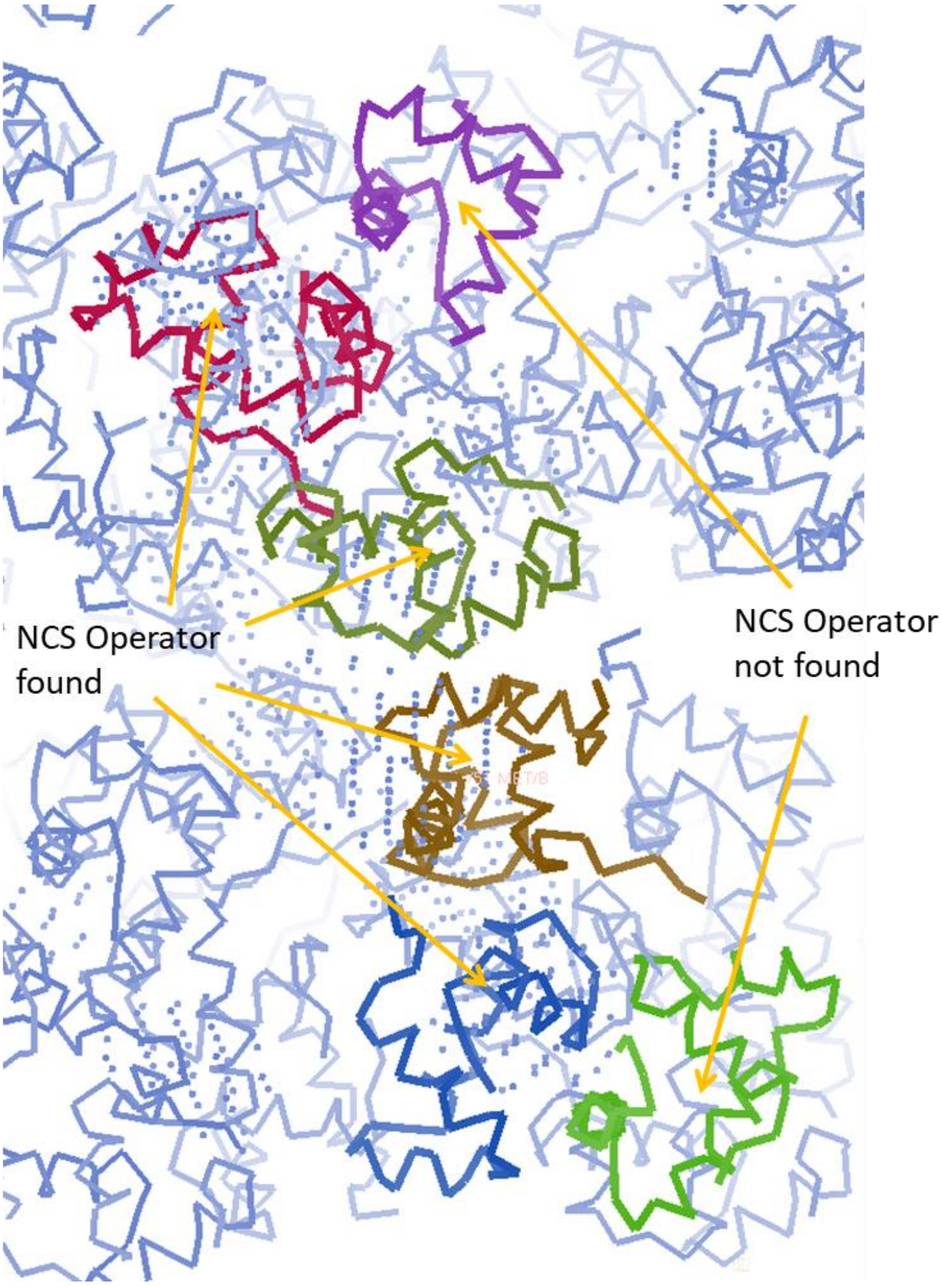


Figure 3-10 Visualisation of unsuccessful discovery of NCS using NaCelleS with anomalous scatterer centres, despite using extended rotation list

The blue spheres here indicate the positions where *NaCelleS* has placed the 10 Å sphere after searching the map under different conditions. The first anomalous scatterer site is used as the centre around which the search sphere of electron density is cut. With 1FSE, four of the six copies were found, but the fifth and sixth copy were not, even with the more extensive list (c). The image was created using *Coot*[23].

Name	Num. of copies expected	Num. of operators found	ρ	TFZ	LLG
02_1038B	10	10	<0.01	4.5	123.3
03_1071B	3	3	0.82	13.5	274.8
04_cp_synthase_1	2	2	<0.01	7.2	30.4
05_cp_synthase_2	2	2	0.05	8.7	32.7
06_cp_synthase_3	2	2	0.83	20.3	46.6
07_s-hydrolase_1	2	2	0.85	37.4	112.1
08_s-hydrolase_		2	0.89	36.4	118.2
09_s-hydrolase_3	2	2	0.89	37.2	120.0
10_ut_synthase	2	2	0.04	8.6	38.8
15_aep_transaminase_1	6	6	0.02	5.4	96.0
18_rab3a_1_Se	1	2	0.09	10.0	75.2
19_rab3a_2_Se	1	2	0.14	13.4	83.1
20_rab3a_3_Se	1	2	0.09	8.3	102.8
21_rab3a_1_Zn	1	2	0.06	8.0	118.5
22_rab3a_2_Zn	1	2	0.09	7.5	180.0
23_rab3a_3_Zn	1	2	0.09	7.7	154.2
24_p32_1	3	3	0.29	9.4	81.5
25_p32_2	3	3	<0.01	9.0	70.0
26_p32_3	3	3	0.24	8.2	82.0
27_cyanase_1	10	10	0.51	3.7	803.2
28_cyanase_2	10	10	0.87	44.5	1184.4
29_cyanase_3	10	10	0.66	8.8	986.4
34_gpatase_1	2	2	0.89	9.7	139.7
35_gpatase_2	2	2	0.88	8.6	145.2
36_gpatase_3	2	2	0.88	9.7	148.3
37_pdz_1	2	2	0.74	35.2	254.6
44_ic_lyase_1	2	2	0.06	10.8	45.5
45_ic_lyase_2	2	2	<0.01	9.6	48.3
46_ic_lyase_3	2	2	<0.01	11.2	46.1
47_mpb_1	2	2	0.85	62.3	1705.2
48_mpb_2	2	2	0.85	56.0	1438.6
50_mpb_4	2	2	0.85	60.3	1645.7

Table 3-2 Results from the one-sphere approach in NaCelleS, using the same centres as phenix.find_ncs (provided by phenix.guess_molecular_centres)

The TFZ and LLG results as reported by *phaser*. A single NCS-averaged map was then created using *phenix.find_ncs* to calculate ρ. A ρ value of less than 0.01 indicates that no NCS average map could be built. The number of copies expected is that which is known to be the correct answer, and was provided as input. A ρ value is highlighted if it is under 0.2, as it is likely that either incorrect NCS operators are found, or no NCS operators were found.

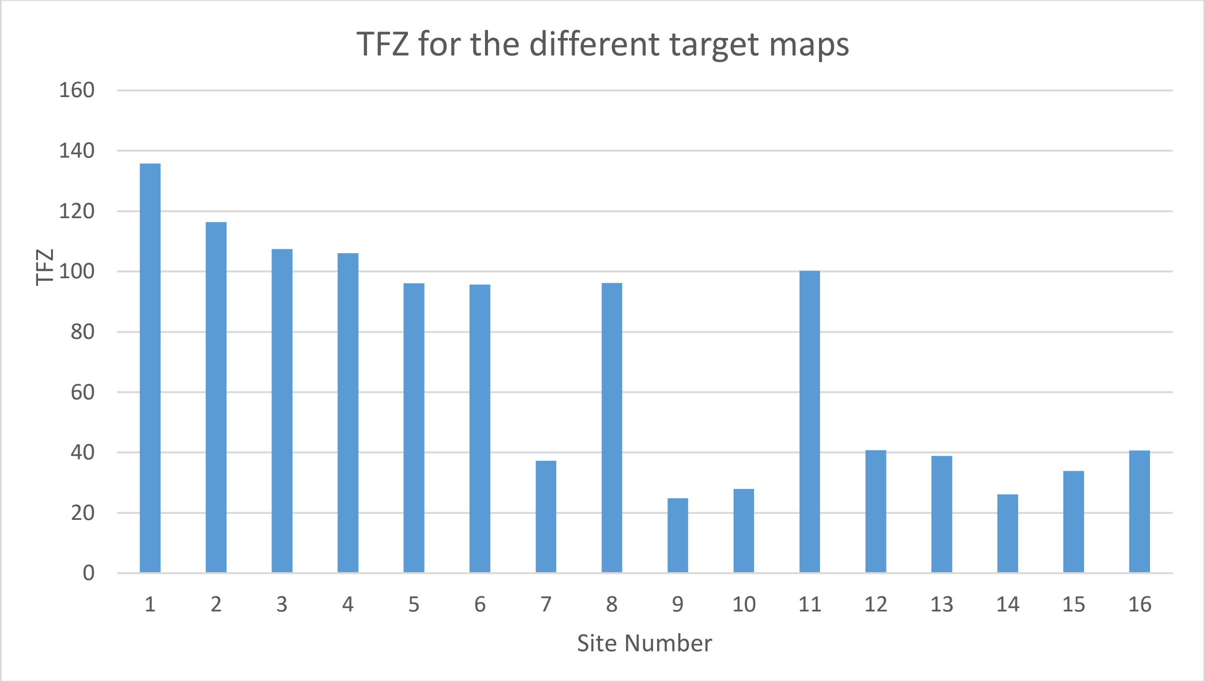
However, there are a number of cases where outcomes are worse with this method. For example, none of the NCS copies in the ic lyase test cases are found: neither can the 1038B or the first two datasets in cp synthase be solved. All of these datasets were then run through *NaCelleS* two further times iteratively. This is a significant issue discussed in a later section (Section 3.9) that suggests that this method complements rather than replaces other approaches.

3.6 Two-sphere approach

3.6.1 Initial Trials

The two-sphere approach in *NaCelleS* succeeded in extracting the NCS parameters with both original test cases. The search sphere used was created on the first in the list of anomalous scatterers, and this is shown for the 1N0E case (Figure 3-11). This selection was made on the assumption that the anomalous scatterer would be within the protein electron density region. The graph shown in this figure was obtained from searches with the same sphere over sixteen different generated test maps, each containing a single sphere cut out around a different anomalous scatterer site. The log-likelihood gain (LLG) and translation function Z-score (TFZ) were recorded in each case.

The results show that the self-peak has a higher TFZ score than all the other peaks. There are then results obtained using target spheres based on the seven sites ranked in order of TFZ values, all of which have a similar high TFZ score. These are clearly grouped together, and are also clearly different from the eight remaining searches around the other sites which have a far lower TFZ score. This is what might be expected, as eight sites are linked by non-crystallographic symmetry, whereas the other sites are located at different parts of the protein.



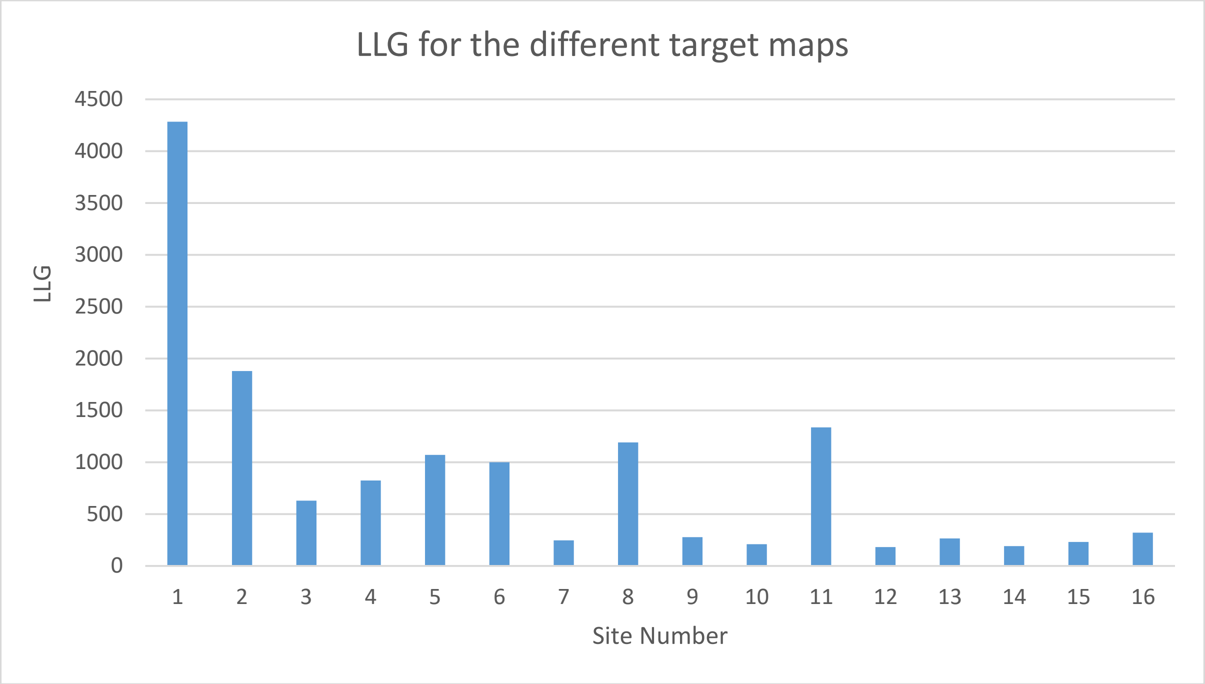
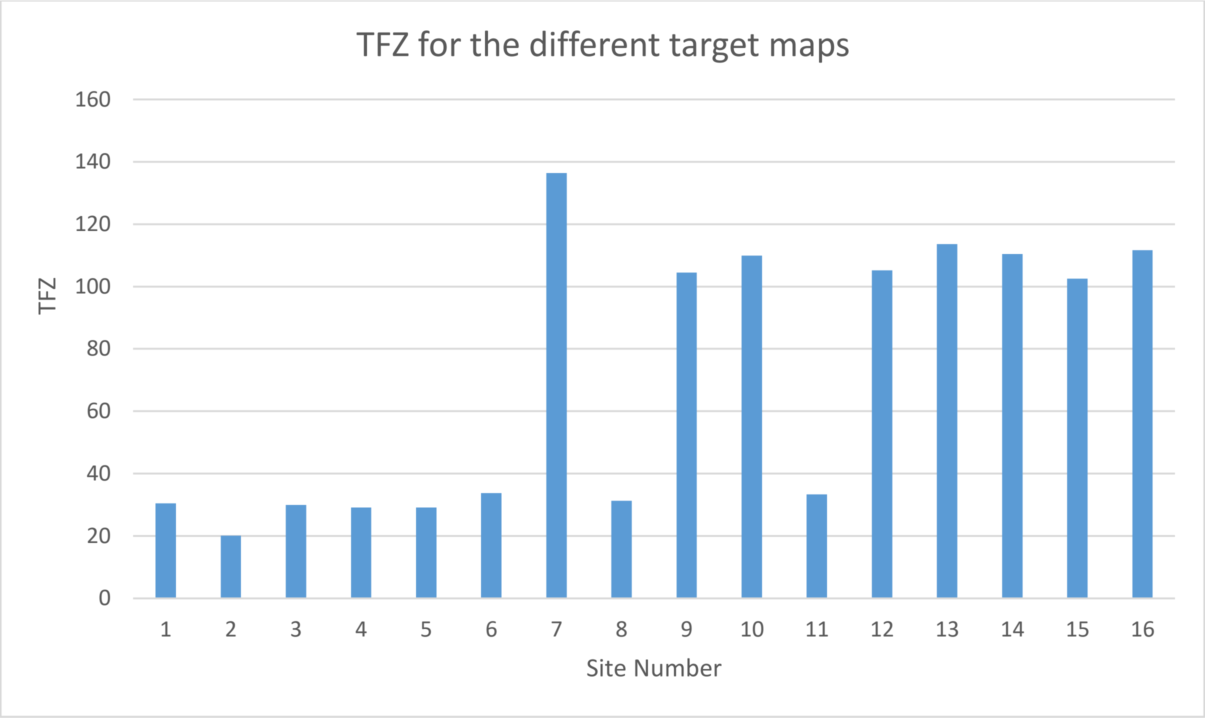


Figure 3-11: TFZ and LLG scores from 1029B two-sphere NaCelleS test case with Sphere 1 as the search sphere

A sphere of electron density was cut from the list of anomalous scatterer sites for the 1N0E test case, centred around the first site. This search model was then used to search a series of second maps, each containing a single sphere of electron density corresponding to the particular anomalous scatterer site, indicated in the label underneath the histograms. The TFZ (top) and the LLG (bottom) was recorded for each of the sixteen maps searched (charts generated in *Microsoft Excel*).



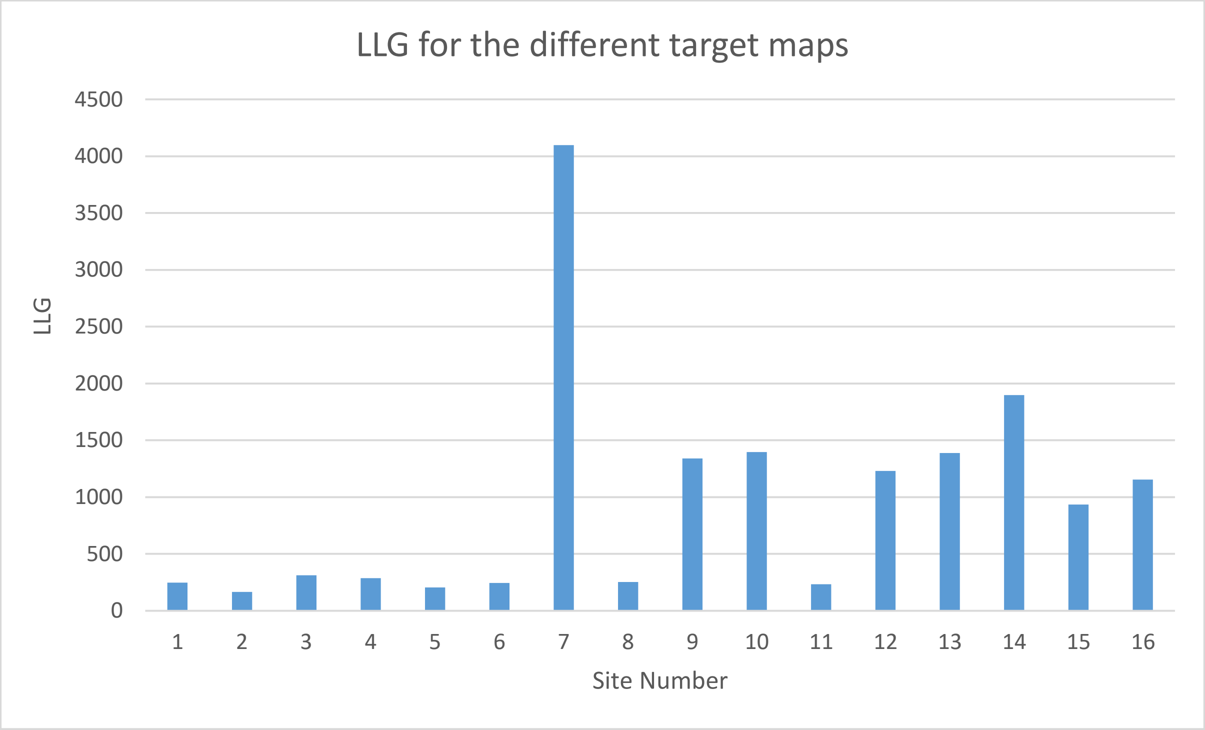
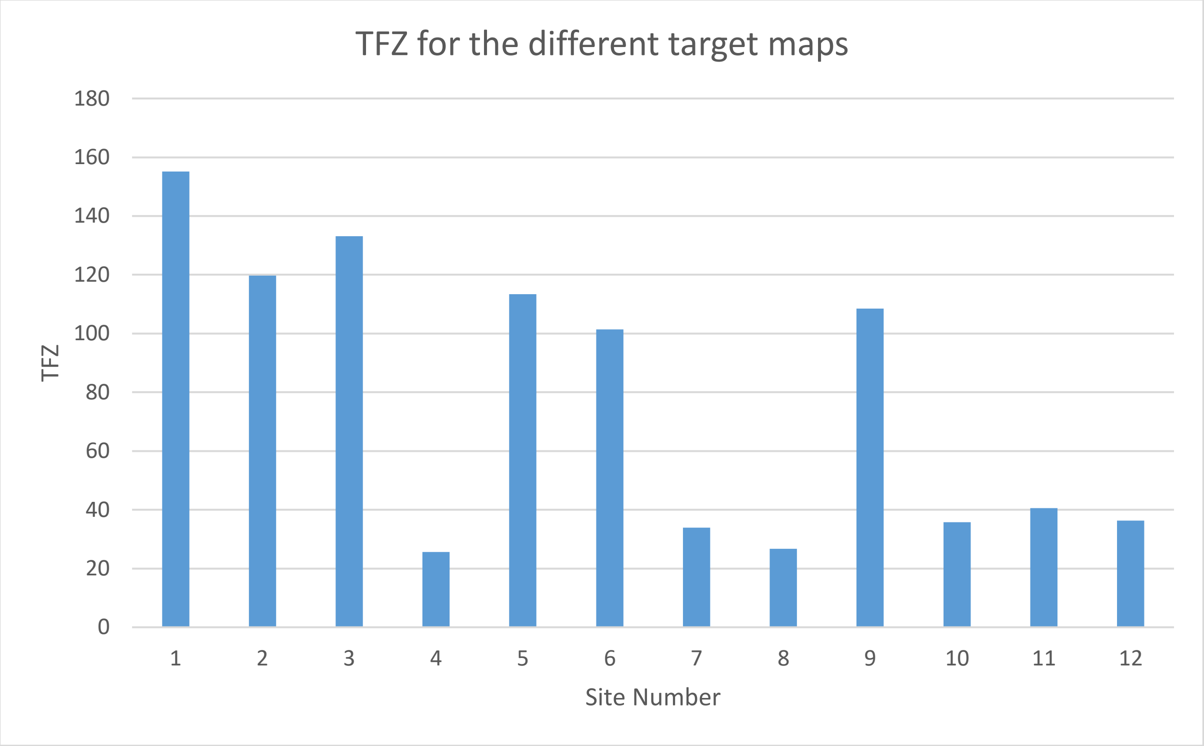


Figure 3-12: TFZ and LLG scores from 1029B two-sphere NaCelleS test case with Sphere 7 as the search sphere

A sphere of electron density was cut from the list of anomalous scatterer sites for the 1N0E test case, centred around the seventh site. This search model was then used to search a series of second maps, each containing a single sphere of electron density corresponding to the particular anomalous scatterer site, indicated in the label underneath the histograms. The TFZ (top) and the LLG (bottom) was recorded for each of the sixteen maps searched (charts generated in *Microsoft Excel*).



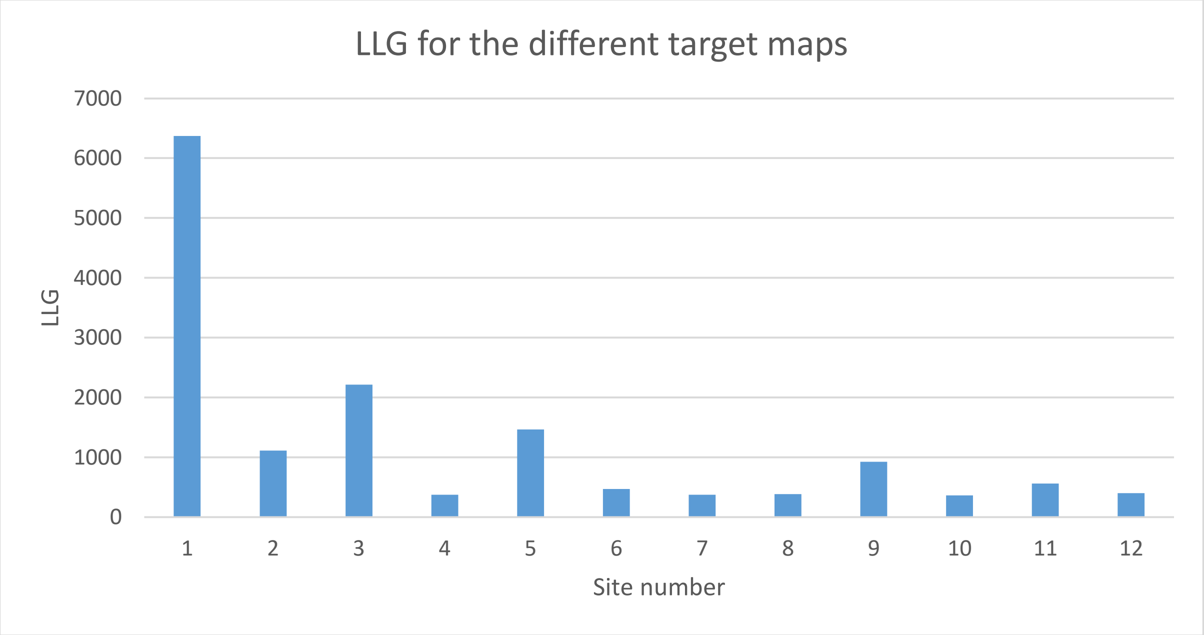


Figure 3-13: TFZ and LLG scores from 1FSE two-sphere NaCelleS test case with Sphere 1 as the search sphere

A sphere of electron density was cut from the list of anomalous scatterer sites for the 1N0E test case, centred around the first site. This search model was then used to search a series of second maps, each containing a single sphere of electron density corresponding to the particular anomalous scatterer site, indicated in the label underneath the histograms. The TFZ (top) and the LLG (bottom) was recorded for each of the sixteen maps searched (charts generated in *Microsoft Excel*).

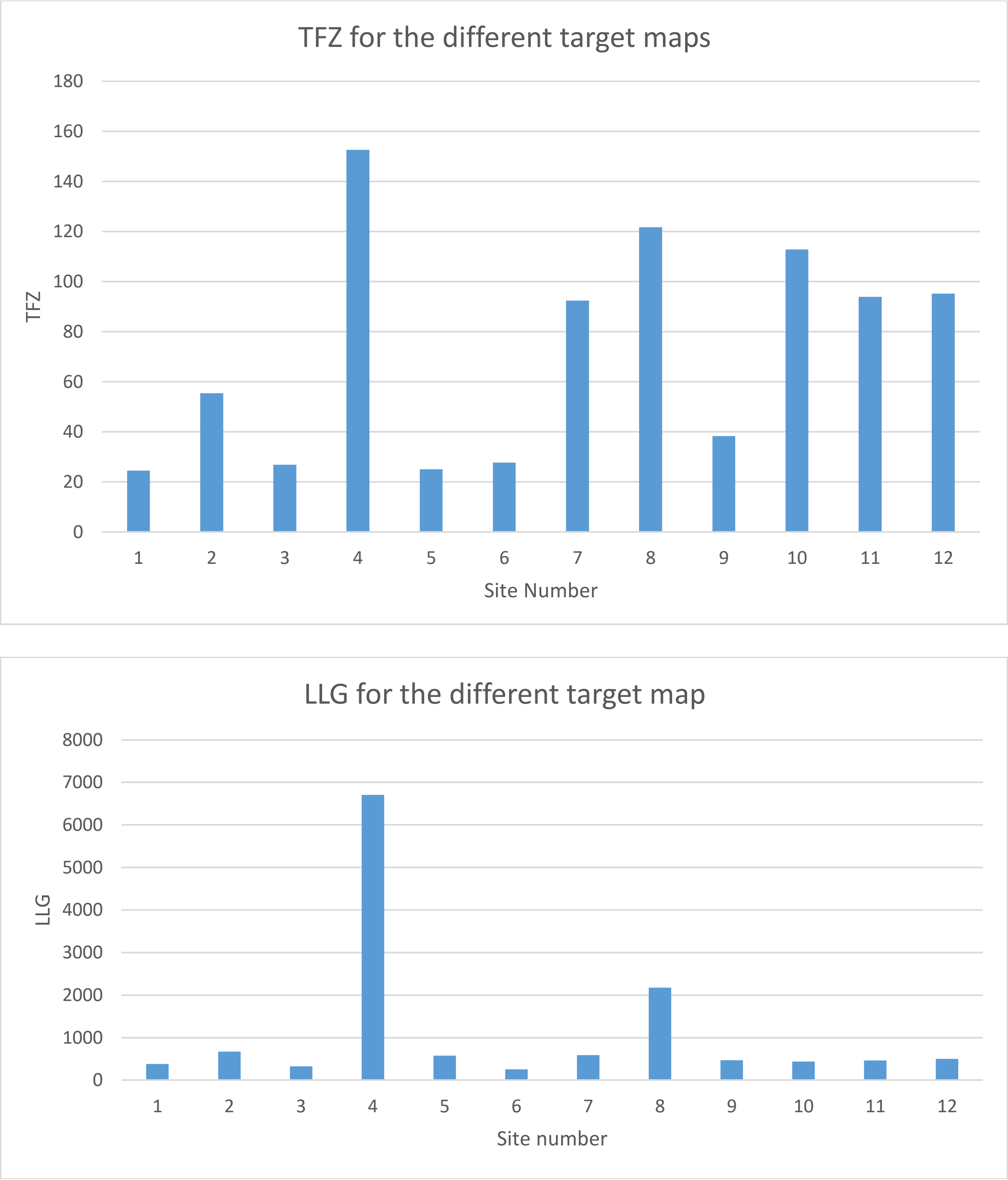


Figure 3-14: TFZ and LLG scores from 1FSE two-sphere NaCelleS test case with Sphere 4 as the search sphere

A sphere of electron density was cut from the list of anomalous scatterer sites for the 1N0E test case, centred around the fourth site. This search model was then used to search a series of second maps, each containing a single sphere of electron density corresponding to the particular anomalous scatterer site, indicated in the label underneath the histograms. The TFZ (top) and the LLG (bottom) was recorded for each of the sixteen maps searched (charts generated in *Microsoft Excel*).

This calculation was then repeated, but this time taking one of the sites that was not linked by NCS (here the seventh site) as the centre of the search sphere of electron density. The search was repeated over all sixteen maps, and the LLG and TFZ scores were recorded (Figure 3-12).

The results show that again the self-peak had the greatest TFZ score (as expected), but the eight peaks that were next highest in this TFZ score now corresponded to the low TFZ scoring sites obtained in the set of results using the first search sphere. This is as expected, as there are two anomalous scatterer sites in the protein, and searching with one of these sites will find the equivalent sites in other copies of the protein, but will be a poor match to the other scatterer locations.

The other test case, 1FSE, yielded similar results. In this case there are six sites that are linked together (Figure 3-13) in a way that is highly similar way to the first test case. By taking one of the sites that does not have a high TFZ score when searched with the initial sphere, we find that the other six sites score highly (Figure 3-14). This again shows clearly that in this case there are six NCS copies, with two anomalous scatterer sites in each monomer.

However, in both cases the LLG scores are less consistent than the TFZ scores, leading to increased difficulty in identifying linked sites using this scoring method alone. The reason for this is that the input maps already have phase information, which the LLG calculations ignore. However, the TFZ calculation takes into account the phases contained in the map, hence it obtains a better signal-to-noise ratio and distinctive results.

It can be seen that the absolute values for TFZ and LLG are very high, far higher than normally expected for a molecular replacement solution. These high scores occur because the search is undertaken with a sphere of electron density in a map that only contains a target sphere of electron density, with the remaining space being empty. Accordingly, a sphere of electron density will always be found, as it is the only feature in the maps, and the value returned indicates the quality of the match between the test and target spheres. If the spheres are different, a high value will still be returned, but it will be much higher if a match is detected. For LLG, the number of structure factors for a small sphere will be relatively small, and a high value will be obtained when a match is found. For TFZ, the random set of translations and rotations will have a very small RMS deviation, as it is mainly uniform density, leading to the very high scores.

3.6.2 Checking uniqueness of NCS operators

It was also necessary to ensure that the NCS operators calculated are all unique, i.e. that the NCS operators found are not the same NCS operator shared between different copies linked by crystal symmetry. To assess this, each NCS operator was compared to every other operator and the TFZ and LLG scores were calculated and compared.

For each of the two NCS operators being compared, a record was made of the translation and rotation of the sphere of electron density. If the sphere of electron density from the two NCS operators is rotated and translated to exactly the same place and orientation, or an identical location based on crystallographic symmetry, then the two NCS operators must be identical, and so only one should be kept in the solution. The crystal symmetry was then applied to both the translated and rotated spheres, and the distance and angle between all possible positions of the electron density were recorded. If any distance was below 2 Å, and the rotations differed by less than an angle (θ) specified as follows

$$\theta = arcsin\left(\frac{\frac{d}{4}}{\frac{radius_{sphere}}{4}}\right)$$

Equation 3-1

where *d* is the resolution of the data, then the two operators are considered identical, and only one is retained. This permitted the identification of unique operators to allow accurate reporting.

This equation is derived from two heuristic considerations. First, there is the concept of the extent to which an electron density can vary before it exceeds an acceptable value. If it moves by an amount of $\frac{d}{4}$, then the electron density will move from being in phase to making no contribution to the beam in question, and values higher than this are considered as not acceptable.

The denominator was then considered. As we are using 10 Å spheres, it was felt that taking the edge of the spheres to determine identity in electron density was not ideal. Taking electron density somewhat closer to the centre is a better representation of the electron density likely to be conserved, and so a number of $\frac{radius_{sphere}}{4}$ was taken.

3.6.3 Two-sphere approach over a wider test dataset

The dataset was now widened to involve a large number (approximately fifty) of datasets. Comparisons were made between the *phenix.find_ncs*[226] approach (Table 3-3) and the *NaCelleS* two-sphere approach, recording the various TFZ scores (Table 3-4). The results across all datasets are summarized as follows. Firstly, as expected, the greatest TFZ scoring peak was always the self-peak i.e. finding itself. In most cases this was followed in rank by a group of similar sites with a TFZ score greater than or around 75 % of the self-peak TFZ score. These were all the solution peaks, likely to be linked by NCS to each other. The remaining peaks, with TFZ scores lower than 75 %, followed. These indicated two points not linked by NCS (Figure 3-15).

This approach clearly indicated the sites that are linked by NCS and those that are not, and so the number of operators does not need to be known beforehand. This freedom is a significant advantage of the two-sphere approach. Secondly, the few borderline cases when the TFZ is around 75 % of the self-peak can be resolved by analysis of the match in both directions i.e. looking to see the score from site A to site B and the reverse. In most cases, if it is not a true site, the TFZ will be far lower in one direction. For example, the highest non-solution peak in the second dataset of 1FSE has a TFZ score of 28.1 (74 %) when site zero is used to search for site three, but in reverse it scores only 18.0 (48 %). Peaks that are true solution peaks are more symmetric, such as a search sphere based on site 0 finding a target sphere at site 10 with a TFZ score of 34.0 (90 %). When looked at in reverse (a search sphere based on site 10 finding a target sphere at site 0) a TFZ score of 33.8 (89 %) can be seen.

	No. to find	f'	f"	Number of copies *phenix.find_ncs_from _density* found	Correlation coefficient from *phenix.find_ncs_from _density*
1029B	8	-3	3.8	4	0.71
1038B	10	-3	3.31	4	0.83
1071B	3	-3	2.18	2	0.82
cp_synthase_1	2	-6	3.86	2	0.82
cp_synthase_2	2	-5.05	3.4	2	0.85
cp_synthase_3	2	-3.74	3.21	2	0.9
s-hydrolase_1	2	-4.65	1.12	2	0.91
s-hydrolase_2	2	-3.84	2.28	2	0.89
s-hydrolase_3	2	-2.28	1.33	2	0.89
ut_synthase	2	-4	1.84	2	0.88
gere_1	6	-5.22	1.42	6	0.71
gere_2	6	-5.22	1.42	6	0.69
gere_3	6	-5.22	1.42	6	0.69
gere_4	6	-5.22	1.42	6	0.66
aep_transaminase_1	6	-8.99	6.26	6	0.88
aep_transaminase_2	6	-9	5.06	6	0.88
aep_transaminase_3	6	-6	4.24	6	0.89
rab3a_1_Se	1	-6.16	0.42	0	0
rab3a_2_Se	1	-7.59	2	0	0
rab3a_3_Se	1	-7.5	2.76	0	0
rab3a_1_Zn	1	-7.65	1.47	0	0
rab3a_2_Zn	1	-7.31	1.48	0	0
rab3a_3_Zn	1	-7.5	1.4	0	0
p32_1	3	-5.36	2.92	3	0.84
p32_2	3	-3	5.7	3	0.87
p32_3	3	0.75	4.2	3	0.87
cyanase_1	10	-5.34	5.04	8	0.63
cyanase_2	10	-4	6.11	8	0.65
cyanase_3	10	-0.05	5.28	6	0.66
cyanase_4	10	-0.05	5.28	6	0.81
fusion-complex_1	3	-6	2.05	0	0
fusion-complex_2	3	-6	1.95	2	0.62
gpatase_1	2	-2.38	0.91	2	0.9
gpatase_2	2	-2.3	1.48	2	0.88
gpatase_3	2	-1.9	1	2	0.88
pdz_1	2	-4	6.18	2	0.33
pdz_2	2	-5.16	3.92	2	0.76
pdz_3	2	-0.17	4.95	2	0.76
pdz_3	2	-0.17	4.95	0	0
1167B	4	-3	4.09	2	0.87
synapsin_1	2	-9.22	0.53	2	0.84
synapsin_2	2	-9	0.84	2	0.78
synapsin_3	2	-8.67	0.61	2	0.83

	No. to find	f'	f''	Number of copies *phenix.find_ncs_from _density* found	Correlation coefficient from *phenix.find_ncs_from _density*
ic_lyase_1	2	-5.94	2.69	2	0.89
ic_lyase_2	2	-4	4.07	2	0.9
ic_lyase_3	2	-1.62	2.9	2	0.89
mpb_1	2	-31.2	12.3	2	0.85
mpb_2	2	-25	30.1	2	0.85
mpb_3	2	-10.8	13.1	2	0.85

Table 3-3 Table showing the results using phenix.find_ncs_from_density with the complete dataset

The table shows the f' and f'' for each dataset used, with some proteins having multiple datasets. The correlation of the NCS density is also reported. If the program does not predict all the NCS operators, the target is highlighted.

	Number of copies expected	Numb of sites used in two-sphere case	Number of operators found in two-sphere *NaCelleS*	TFZ of self-peak	TFZ range of solution peaks	TFZ of other peaks
1029B	8	16	8	39.6	36.2 - 28.9	18.8 - 13.2
1038B	10	50	10	40.2	36.9 - 30.7	21.3 - 10.1
1071B*	3	18	2	44	31.8	25.6 - 11.7
cp_synthase_1	2	16	2	43.4	30	23.9 - 10.3
cp_synthase_2*	2	16	2	37.6	33.1	16.9 - 10.5
cp_synthase_3	2	16	2	40.6	29.6	17.4 - 10.3
s-hydrolase_1	2	30	2	41.7	35.2	28.6 - 11.8
s-hydrolase_2*	2	30	2	39.3	35.1	28.1 - 9.1
s-hydrolase_3	2	30	2	41	34.7	27.8 - 9.5
ut_synthase	2	24	2	38.3	36.4	25.8 - 9.1
gere_1	6	12	6	39.7	36.2 - 32.6	28.4 - 14.6
gere_2	6	12	6	37.8	35.7 - 34.0	28.1 - 16.7
gere_3	6	12	6	39.5	37.1 - 35.0	28.1 - 12.0
gere_4	6	12	6	38.5	36.0 - 34.3	29.5 - 14.7
aep_transaminase_1	6	66	6	38	36.5 - 30.8	24.5 - 9.3
aep_transaminase_2	6	66	6	38.3	36.8 - 31.6	23.0 - 8.7
aep_transaminase_3	6	66	6	38.7	36.9 - 31.7	23.8 - 8.4
rab3a_1_Se	1	10	1	38.9	N/A	28.6 - 9.8
rab3a_2_Se	1	10	1	38.5	N/A	28.7 - 10.1
rab3a_3_Se	1	10	1	36.4	N/A	21.7 - 9.1
rab3a_1_Zn	1	2	1	40.6	N/A	16.2
rab3a_2_Zn	1	2	1	38.8	N/A	16.4
rab3a_3_Zn	1	2	1	40.2	N/A	13.9
p32_1*	3	9	3	40.2	34.5 - 34.2	24.5 - 13.0
p32_2*	3	9	3	40.4	35.5 - 35.5	24.4 - 12.5
p32_3*	3	9	3	40	35.1 - 34.8	24.3 - 13.5
cyanase_1	10	39	10	40.3	39.7 - 33.6	25.0 - 8.9
cyanase_2	10	39	10	40	39.1 - 33.3	24.0 - 8.0
cyanase_3	10	39	10	41.1	39.6 - 32.7	26.3 - 9.0
cyanase_4	10	39	10	41	39.4 - 32.8	26.5 - 8.5
fusion-complex_1*	3	15	3	38.8	36.0-32.4	23.7 - 12.4
fusion-complex_2 *	3	15	3	39.5	36.0-32.5	22.5 - 12.6
gpatase_1	2	20	2	40.3	35.3	22.3 - 8.3
gpatase_2	2	20	2	40	34.3	19.7 - 9.8
gpatase_3	2	20	2	39.7	34	23.0 - 10.8
pdz_1*	2	12	2	39.7	34	23.0 - 10.8
pdz_2*	2	12	2	40.5	35.9	21.0 - 8.0
pdz_3*	2	12	2	39.6	29.7	14.9 - 10.8
pdz_4*	2	12	2	41	30.4	14.9 - 9.5
1167B*	4	8	4	35.3	34.2 - 30.0	13.0 - 9.4
synapsin_1*	2	20	2	40	35.3	23.6 - 10.5
synapsin_2*	2	20	2	42.2	35.6	20.4 - 9.4
synapsin_3*	2	20	2	38.8	35.2	19.4 - 9.8

	Number of copies expected	Numb of sites used in two-sphere case	Number of operators found in two-sphere *NaCelleS*	TFZ of self-peak	TFZ range of solution peaks	TFZ of other peaks
ic_lyase_1	2	12	2	39	35.1	18.3 - 9.8
ic_lyase_2	2	12	2	39.1	34.3	23.8 - 12.2
ic_lyase_3	2	12	2	38.3	34.6	22.1 - 11.7
mpb_1*	2	3	2	44.5	41.7	22.1
mpb_2*	2	3	2	44.4	41.7	22.1
mpb_3	2	3	2	42.8	39	11.1

Table 3-4 Table showing results using the NaCelleS two-sphere approach with the complete dataset

The statistics are all for the first site in the heavy atom list, except for the * datasets, when a second site had to be used as the search volume to obtain the results. The results show the number of sites seen, the number of operators found, and the results for the operators. The TFZ score of the self-peak is reported. The TFZ score range of all solution peaks is reported, and then the TFZ scores of all the other peaks are recorded. If *phenix.find_ncs_from_densit*y cannot find all the NCS but *NaCelleS* can, it is highlighted in green. If neither *phenix.find_ncs_from_density* nor *NaCelleS* can find all the NCS, it is highlighted in yellow.

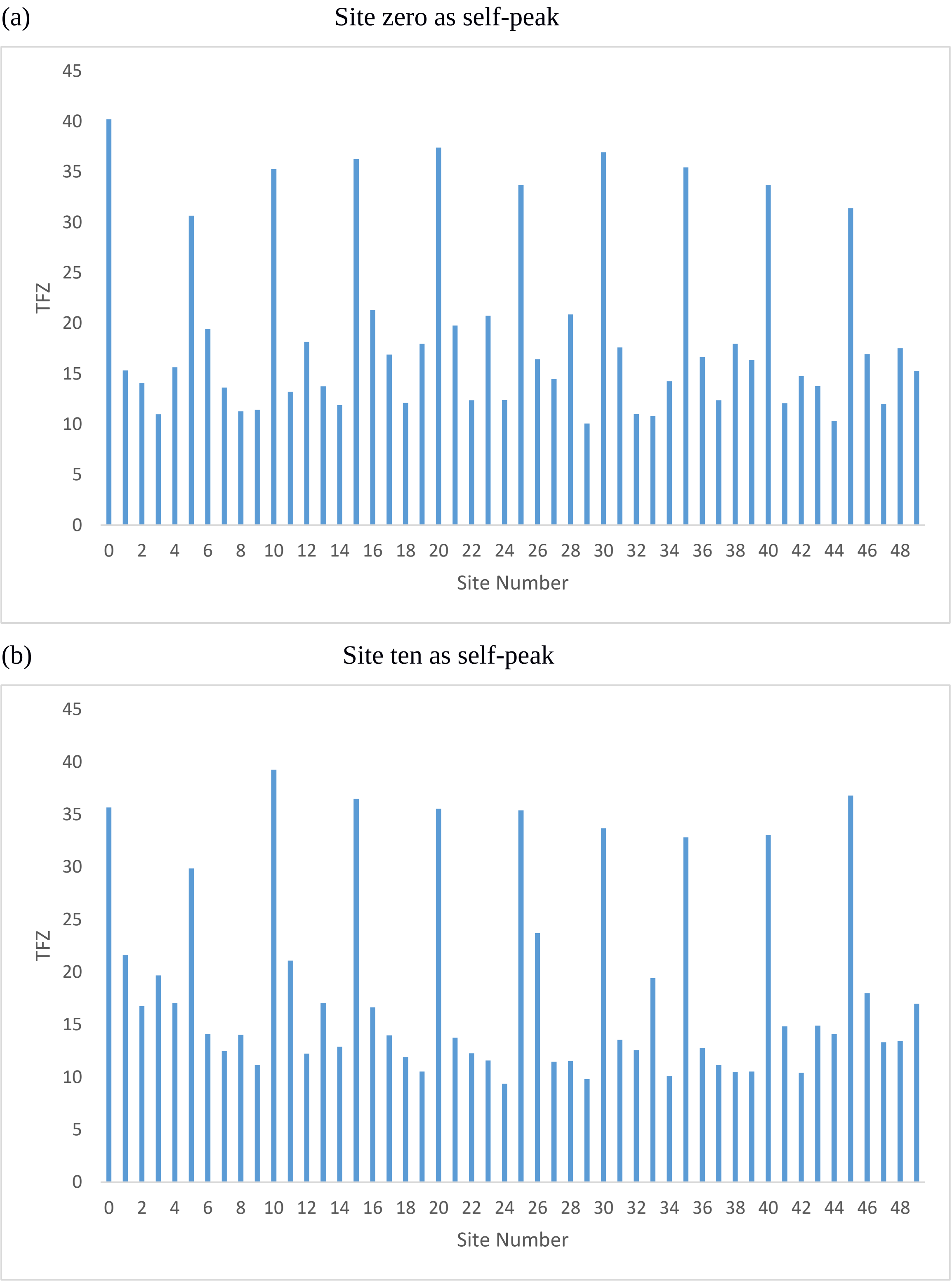


Figure 3-15 Example of one of the datasets, 1038B

The self-peak set at site zero (a) or site ten (b) are shown in this figure. The noise peaks vary greatly (charts generated in *Microsoft Excel*).

Occasionally the first site was not linked by NCS to all, or even to any other sites (examples in Table 3-4 labelled * are a good example of this). Given that the sites in the anomalous scatterers location file were in no particular order, this is not particularly unexpected. If this happens, looking at other sites can still provide the NCS operators, and can allow for the accurate determination for whether sites are linked by NCS or not.

Comparing the results from *NaCelleS* and *phenix.find_ncs_from_density*[2], it can be seen that *NaCelleS* improved the outcome in ten cases (green highlighted, Table 3-4), and performed equally well in all the other test cases. However, on several occasions not all the sites showed the NCS operators, so testing with several centres was needed. Therefore the algorithms in this program offer a useful complementary method, and allow for a comparison of anomalous scatterer sites by looking at the electron density around them.

3.7 Iterative approach

All approaches investigated so far did not identify all operators in every case. These approaches did not permit the use of the new NCS operator information that had been found through this first pass. This led to the creation of an iterative approach as the next step in this investigation. Information about the newly identified NCS operators was used to improve our ability to find additional NCS operators. This is not possible in *phenix.find_ncs_from_density*[2], where after an initial instance is completed, no further NCS operators can be generated.

3.7.1 Obtaining a single NCS copy of a protein

In order to facilitate this iterative approach, an averaged electron density map of a single monomer was made from the different monomer maps at the identified locations, using the known NCS operators in *phenix.find_ncs*[226] (NCS-averaged map). This averages suitably rotated and translated equivalent parts of the electron density map to form as large a single copy of the electron density of the monomer as possible. This single copy was then placed in an electron density map containing the same unit cell dimensions as the original map. It was then cut out and used as the search model over the entire electron density. The aim of this was to improve the signal-to-noise ratio by averaging out the noise in the map for the monomer.

When doing this iteration with the starting centre derived from *phenix.guess_molecular_centres*[2], a near-perfect NCS-averaged copy was achieved (Figure 3-16 and Figure 3-17). This implies that the NCS operators were accurate, and that the superimposition of the electron densities to achieve the average was good. This averaged map of the monomer, when used within *NaCelleS*, led to all eight copies being found in 1N0E, and all six copies being found in 1FSE, with correlation ρ-values of 0.64 and 0.61 respectively. This then produced a very similar set of NCS operators in each case to the NCS operators that had been originally found, indicating good self-consistency.

These new NCS operators were used to create an improved NCS-averaged map, which was then reinserted as the search model again for the entire map in an iterative process. In both the 1N0E and 1FSE cases, unsurprisingly all the NCS operators were again found. This indicates that the process remains self-consistent after iteration.

3.7.2 Iteration with the anomalous scatterer centres

These initial results suggested that the approach should be tested against the anomalous scatterer centres. In particular, could an iterative approach be used to find more NCS operators, starting from the partial NCS operators obtained when using the anomalous scatterer centre in *NaCelleS*? A single monomer NCS electron density map was created using the incomplete operators, and then used as a search model (Figure 3-18 and Figure 3-19). The hypothesis was that further NCS operators would then be discovered. However, this was unsuccessful for a number of reasons, as discussed in Section 3.9 in the context of the wider test set.

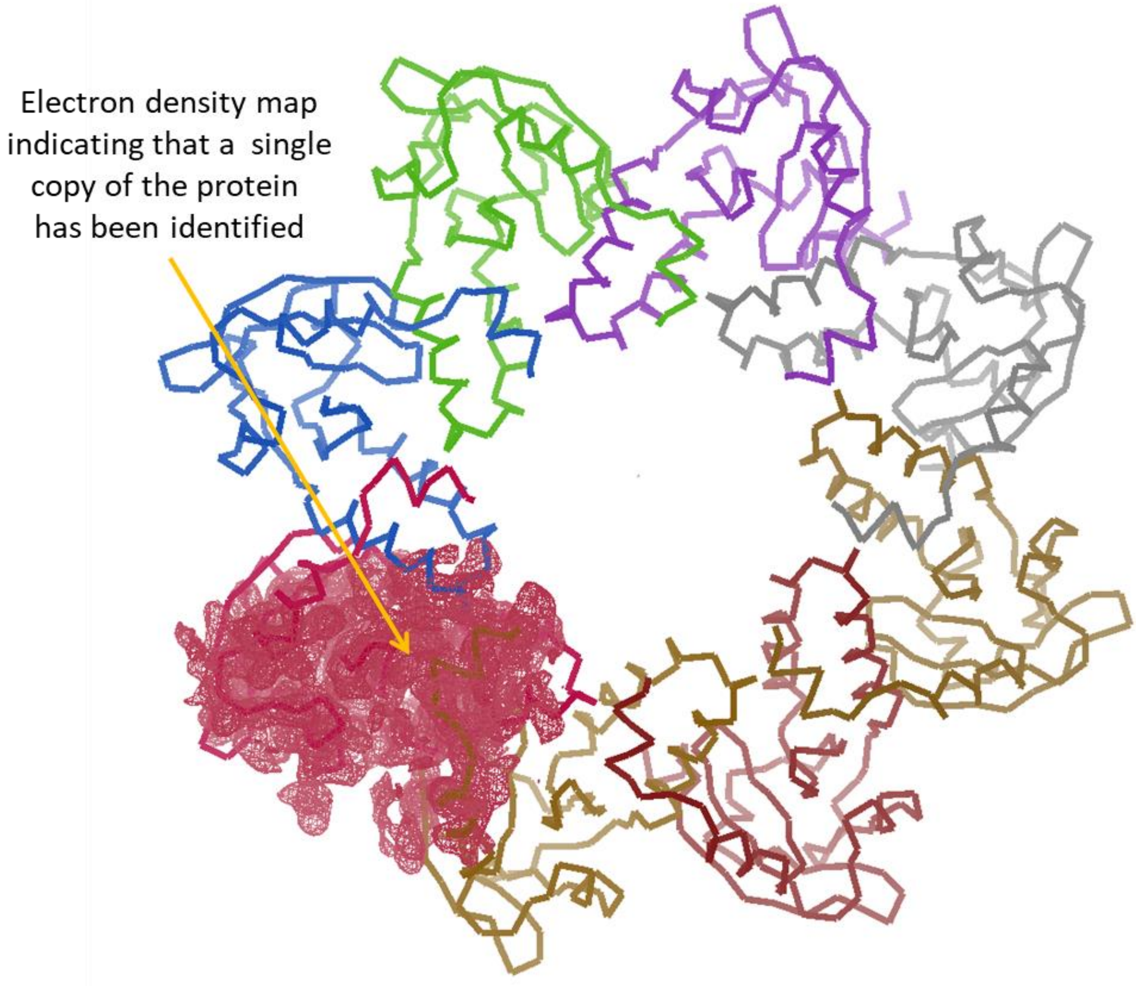

Electron density map indicating that a single copy of the protein has been identified

Figure 3-16: Electron density for single NCS-average copy of 1N0E with complete operators

This shows the 1σ contour of the electron density after creating the NCS-average map for use in the iterative approach. It shows that a single copy of the protein has been identified. The image was created using *Coot*[23], and labelled in *Microsoft PowerPoint*.

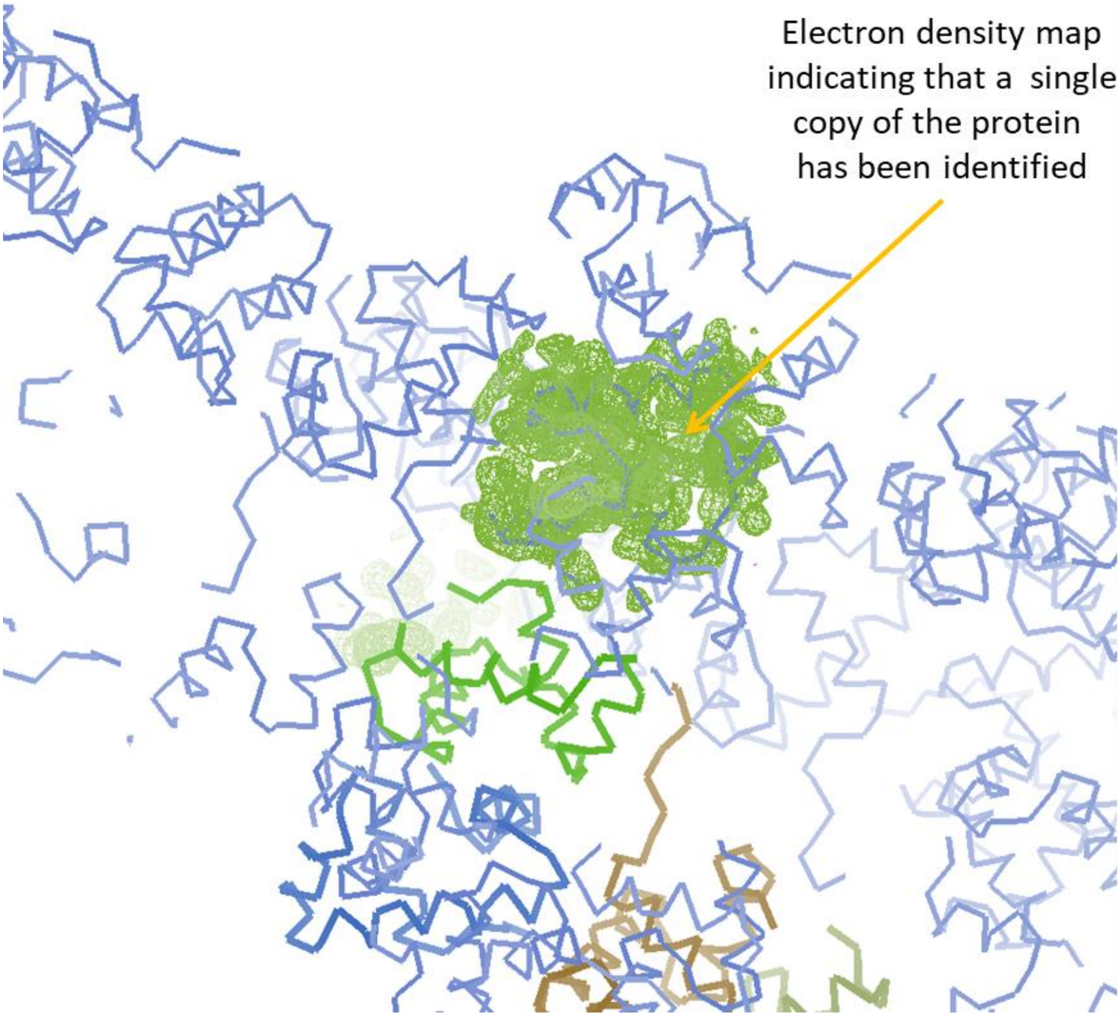

Electron density map indicating that a single copy of the protein has been identified

Figure 3-17 Electron density for single NCS-average copy of 1FSE with complete operators

This shows the 1σ contour of the electron density after creating the NCS-average map for use in the iterative approach. It shows that a single copy of the protein has been identified. The image was created using *Coot*[23], and labelled in *Microsoft PowerPoint*.

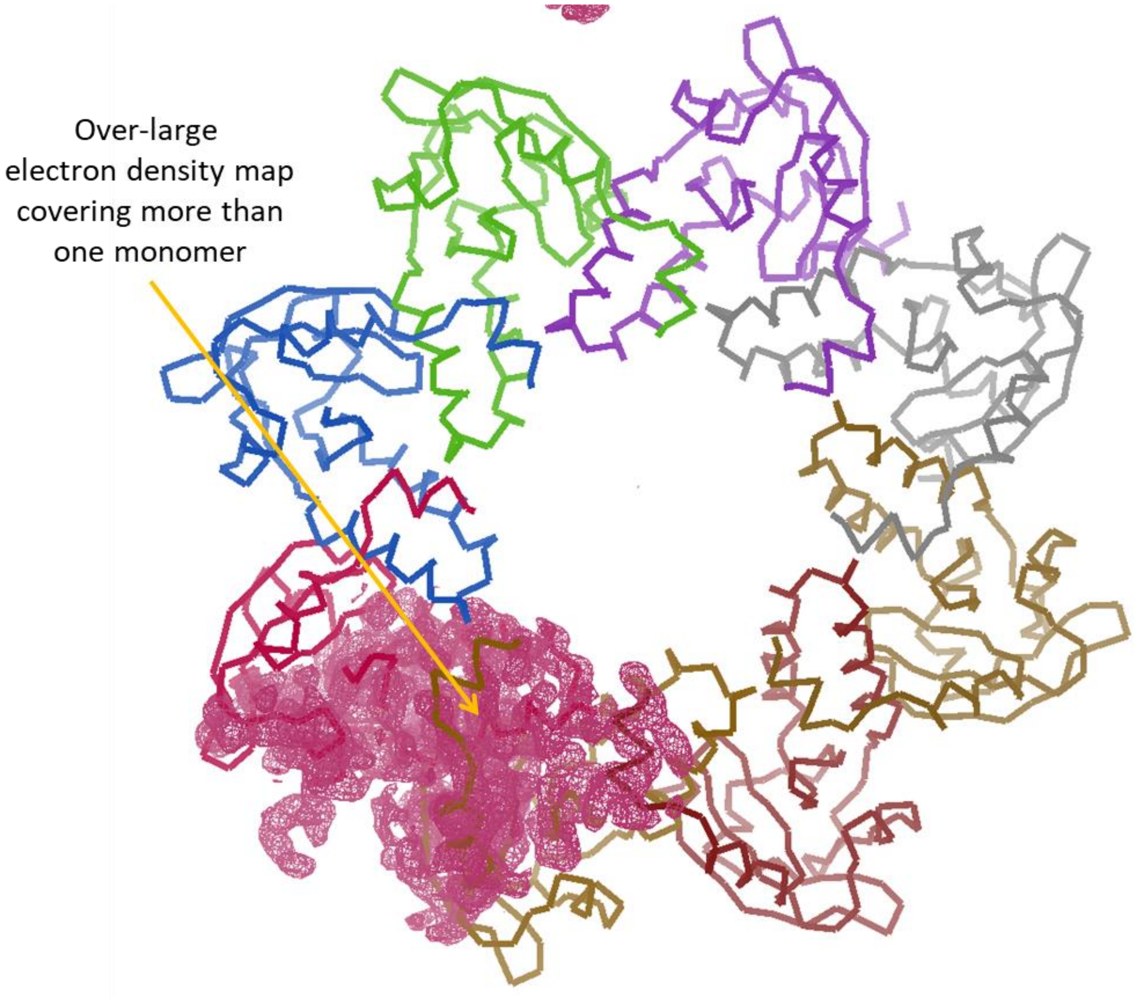


Figure 3-18: Electron density for single NCS-average copy of 1N0E with incomplete operators

This shows the 1σ contour of the electron density after creating the NCS-average map for use in the iterative approach. As can be seen, these maps cover more than one monomer due to the incomplete operators. These maps, when used for an iterative approach in *NaCelleS*, were not successful in finding more operators. The image was created using *Coot*[23], and labelled in *Microsoft PowerPoint*.

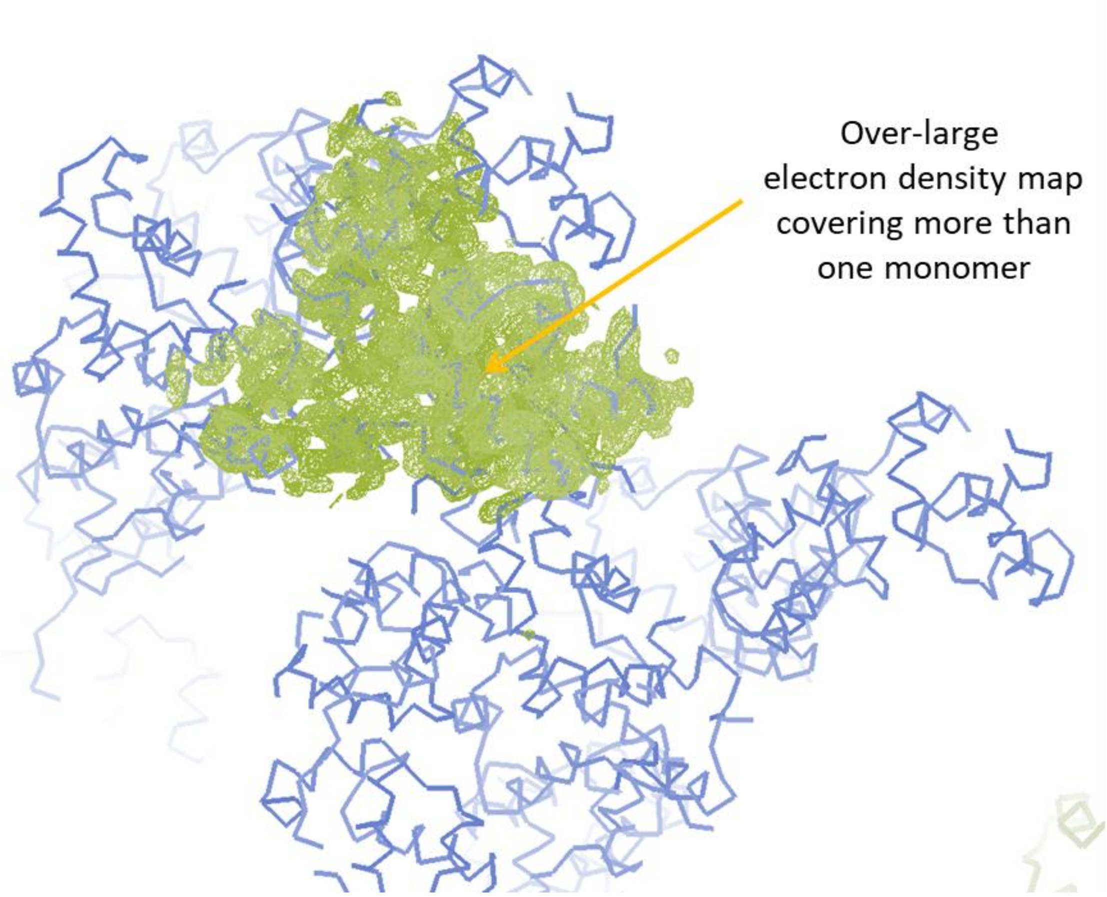

Over-large electron density map covering more than one monomer

Figure 3-19 Electron density for single NCS-average copy of 1FSE with incomplete operators

This shows the 1σ contour of the electron density after creating the NCS-average map for use in the iterative approach. As can be seen, these maps cover more than one monomer due to the incomplete operators. These maps, when used for an iterative approach in *NaCelleS*, were not successful in finding more operators. The image was created using *Coot*[23], and labelled in *Microsoft PowerPoint*.

3.8 Use of the iterative approach with the extended dataset

Applying the iterative approach to the larger dataset led to varied results (Table 3-5). Two complete iterations were undertaken, and the results analysed. Some datasets completely deteriorated (Section 3.9), such as gpatase first and third datasets, despite being solved with the original single sphere approach. Conversely, both the first and third datasets of both p32 and cyanase had a greatly improved outcome, and have achieved far better operators than were obtained originally. In the case of cyanase, improved results were achieved when compared to those originally achieved in *phenix.find_ncs_from_density*[2] and the *NaCelleS* two-sphere approach. In summary, this process showed improved results with certain proteins and offers a viable alternative approach.

In the case of aep transaminase, which the initial single sphere approach did not solve, with low scores in ρ, TFZ and LLG, the iterative approach through *NaCelleS* seems to have had success in finding the NCS operators. This shows the iterative case can occasionally rescue a solution, and so help in finding the NCS operators.

Name	Num. of copies expected	Num. of operators found	ρ	TFZ	LLG
02_1038B	10	NRF	NRF	NRF	NRF
03_1071B	3	3	0.80	19.5	585.4
04 cp_synthase_1	2	2	0.09	58.7	458.9
0 _cp_synthase_2	2	2	<0.01	60.5	493.1
06_cp_synthase_3	2	2	0.84	76.6	2028.0
07_s-hydrolase_1	2	2	0.91	58.3	12767.7
08_s-hydrolase_2	2	2	0.89	57.8	12261.1
09_s-hydrolase_3	2	2	0.89	59.7	12398.9
10_ut_synthase	2	NRF	NRF	NRF	NRF
15_aep_transaminase_1	6	6	0.60	65.9	6464.2
18_rab3a_1_Se	1	2	0.13	60.7	1055.4
19_rab3a_2_Se	1	2	0.08	60.8	1268.4
20_rab3a_3_Se	1	2	0.11	63.7	1102.7
21_rab3a_1_Zn	1	2	0.06	7.6	275.7
22_rab3a_2_Zn	1	2	0.11	56.6	1034.1
23_rab3a_3_Zn	1	2	0.09	58.1	1172.3
24_p32_1	3	3	0.89	74.0	8045.8
25_p32_2	3	NRF	NRF	NRF	NRF
26_p32_3	3	3	0.88	71.8	7860.4
27_cyanase_1	10	10	0.93	79.7	7562.4
28_cyanase_2	10	10	0.92	78.1	7394.1
29_cyanase_3	10	10	0.85	76.1	6885.2
34_gpatase_1	2	2	<0.01	60.5	19995.8
35_gpatase_2	2	2	0.88	60.9	20195.0
36_gpatase_3	2	NRF	NRF	NRF	NRF
37_pdz_1	2	2	0.73	74.1	2781.7
44_ic_lyase_1	2	2	0.06	24.8	1969.4
45_ic_lyase_2	2	NRF	NRF	NRF	NRF
46_ic_lyase_3	2	NRF	NRF	NRF	NRF
47_mpb_1	2	2	0.86	104.8	8184.4
48_mpb_2	2	2	0.85	102.9	7508.7
50_mpb_4	2	2	0.86	94.2	7943.8

Table 3-5 Results from the iterative approach, with two further iterations to identify NCS operators

The one-sphere search approach was used, starting from the same centres as *phenix.find_ncs* (provided by *phenix.guess_molecular_centres*), but taken through two iterations of *NaCelleS*, to try and improve the operators or find further operators. The TFZ and LLG are reported from *phaser*. A single NCS-averaged map was then created using *phenix.find_ncs*. The results recorded as NRF (no reported factors) indicate that an earlier iteration was not able to generate an NCS averaged map (as ρ was less than 0.01). A ρ value of less than 0.01 indicates that no NCS average map could be built in the final iteration. A ρ value is highlighted if it is under 0.2, as it is likely that either incorrect NCS operators are found, or no NCS operators were found.

3.9 Discussion

Despite encouraging and improved results in many cases, there were several significant challenges to the algorithms within *NaCelleS*, and in particular in the operation of the iterative approach. These are now discussed, and comments provided on the underlying issues. Suggestions for future work are also provided later (Section 5.2.2).

3.9.1 Number of monomer copies to identify

Some programs require initial information on the number of monomer copies that are found in the asymmetric cell. This can be a limitation if this number is not known but several ways of estimating it exist, primarily based on the Matthews coefficient[291] and the Kantardjieff and Rupp distribution[292]. For small numbers of monomers, this distribution is narrow and predicts the correct number of monomers. However, when there are more than a few monomers in a unit cell, the distribution becomes broader and calculating the exact number in the asymmetric unit becomes more difficult.

Some programs or algorithms do not need this number. *phenix.find_ncs_from_density*[2] takes one-sphere, and then reports back as many NCS operators as it can find within the asymmetric unit which are separated by more than a peak separation parameter, typically 1.5 times the radius of the search sphere. In the *NaCelleS* two-sphere approach, there is also no necessity to know the number of copies initially.

In the *NaCelleS* one-sphere approach, *phaser*[4] is used to find the places to situate the one-sphere of electron density and *phaser*[4] must therefore be told to explicitly place a certain number of copies. *phaser*[4] will attempt to find this number of copies, even if this is greater than the actual NCS that exists in the cell.

The problem that incorrect operators could be included in the solutions then arises, and is discussed below with reference to the creation of an incorrect NCS-averaged search electron density map. Attempts to filter out incorrect solutions were not successful (Section 3.9.2).

If the iterative approach is taken to improve and find further NCS parameters from this initial set of parameters, then the number of NCS operators also needs to be known to allow the program to find the correct number of locations. However, if the two-sphere initial approach is then iterated, the number of NCS-copies can be calculated from the peaks in the TFZ, so this number is not required as an explicit input. This may lead to opportunities for more automated solutions.

It is important to understand the significance of knowing the number of copies. In order to estimate this, the number of copies predicted in the asymmetric unit for all the cases of the larger dataset was explored. As can be seen, in six out of eighteen cases, the number of copies calculated was incorrect (Table 3-6), indicating that this issue is of concern.

Name	**Number of copies in asymmetric unit**	**Calc number of copies in asymmetric unit**	**Relative Prob. of finding Calc number**	**Relative Prob of finding actual number of copies**
1029B	8	10	0.954	0.387
gere	6	6	1	-
1038B[226]	10	13	0.954	0.267
1071B[275]	3	3	1	-
cp-synthase[276]	2	1	1	0.058
s-hydrolase[277]	2	2	1	-
ut-synthase[278]	2	2	1	-
aep-transaminase[279]	6	7	1	0.574
rab3a[280]	1	1	1	-
p32[281]	3	3	1	-
cyanase[282]	10	11	1	0.767
fusion-complex[283]	3	3	1	-
gpatase[285]	2	2	1	-
pdz[286]	2	2	1	-
1167B[287]	4	4	1	-
synapsin[288]	2	2	1	-
ic-lyase[289]	2	3	1	0.089
mbp[290]	2	2	1	-

Table 3-6 List of all test data sets and the ability to correctly predict how many copies will be in the asymmetric unit

Using the probabilities from the Matthews Coefficient[291,292] as implemented in *phaser,* prediction of the number of copies is correct in 66 % of cases.

3.9.2 Limitations in the creation of masks

Several limitations in the creation of masks were identified as part of the iterative approach. The NCS operators that are found in *NaCelleS* are fed to *phenix.find_ncs*, and these are then used to create a single averaged copy of the monomer in the asymmetric unit. However, if only a subset of the NCS operators is found, then the likelihood is that the process of defining the monomer mask will result in a volume larger than a single monomer being identified. The process used to define the repeat region compares the regions around the two NCS locations and tries to make this as large as possible to maximise the signal. If a further copy of the protein has not been identified close to the original protein then it will be included within this definition. This will lead to an excessively large mask of electron density being cut out, and so the averaged space that is covered extends beyond a single asymmetric unit. This leads to difficulty in using this mask to search over the whole map for all the NCS operators, as *phaser*[4] legitimately rejects correct solutions because of packing constraints. Further operators are thus not found, and there is no improvement.

Attempts were made to look for internal symmetry within the NCS-averaged map, and see whether there were multiple monomers within the defined volume. Preliminary attempts using *NaCelleS* primarily, and *phenix.find_ncs_from_density*[2] to a lesser extent, were unsuccessful. The former always placed a second copy in the highest scoring position, but it was difficult to determine if the score associated with the second copy was significant enough to indicate the existence of internal symmetry. The latter never located a second copy.

A further independent problem occurs when an incorrect operator is included with correct operators in creating the NCS-averaged map. This usually leads to a poorer search model, as the average now includes contributions from areas that do not match, resulting in a loss of signal-to-noise ratio. The iteration approach to finding more and better operators can then fail as the new search model is no better than the original, and the NCS operators reported are usually worse.

3.9.3 Challenges in calculating the mass within the search region of electron density

To allow successful *phaser*[4] function, the mass of the atoms associated with the search electron density model is required. This allows the program to calculate the fraction of the scattering associated with the search volume correctly. Unlike models for molecular replacement where the mass can be calculated from the sequence, this requirement is significantly more difficult for the search volumes, as cutting out a 10 Å sphere either around an anomalous scatterer site or a centre estimated by *phenix.guess_molecular_centres*[2] presents inherent difficulties in mass estimation.

3.9.3.1 Mass estimation using Matthews coefficient and 50 % solvent

One possible assumption that could be used is that the crystal is 50 % protein, and so the molecular weight of the protein can be calculated as (assuming a specific volume of 1.21 Å^3 Da^{-1})

$$Mass = \left(\frac{1}{1.21}\right) * 0.5 * Volume$$

Equation 3-2

However, as can be seen (Figure 3-20), there can be a substantial difference in mass distribution inside the sphere, which can account for different scattering characteristics. Using this approach on its own will be a poor estimate, and is not particularly useful.

3.9.3.2 Upper limit: Matthews coefficient

There are various upper limits that can be applied to estimates of this mass. The first upper limit is that the sphere cannot be greater than 100 % protein, using a Matthews Coefficient partial specific volume[292] of 1.21 Å^3 Da^{-1}. Therefore the absolute maximum mass in the sphere is then

$$Mass = \left(\frac{1}{1.21}\right) * Volume$$

Equation 3-3

3.9.3.3 Upper limit: monomer mass

Another upper bound is that the search mass can never exceed the mass of one monomer. A monomer is being sought, so the mass of the electron density being searched must not be greater than this.

However, the assumption that the mass of the search volume cannot exceed the mass of a single monomer can cause issues. As discussed previously, the limitations in defining the mask can sometimes lead to a search volume that covers two or more monomers. This is problematic, as it means that we are incorrectly stating the fraction scattering ascribed to this volume in *phaser*[4]. This, and the fact that the search will be looking for too many copies under these conditions, will lead to significant problems in the determination of NCS.

3.9.3.4 Upper limit: extent of NCS-averaged mask

In the iterative approach, in creating an NCS-averaged map of the monomer from operators that have already been defined (the search volume in this case), the three dimensional limits for the NCS-averaged volume are reported. This then provides the limits of a box within which all the electron density is now contained, and so from this box, the volume is obtained. If the box is taken to be 100 % protein, then another limit of the mass is determined. However, this value is not particularly useful, as it is usually a far higher mass than is actually in the fragment. The mass is also usually much higher than the previously defined upper limit of having only one monomer within this volume.

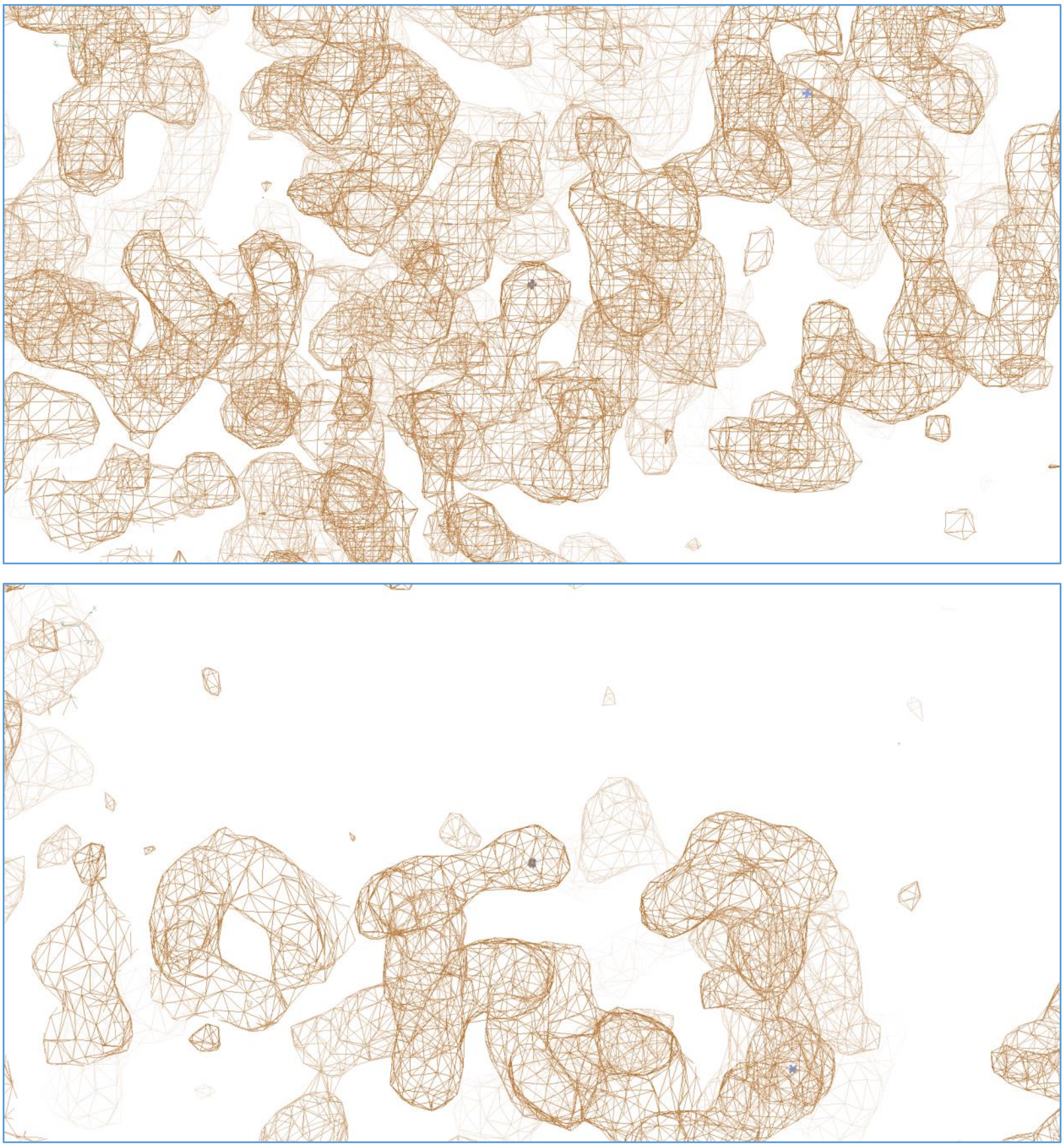


Figure 3-20 Illustration of the vastly differing amount of electron density around different anomalous scatterer sites (in 1FSE) selected at different centres

Two-spheres of 10 Å have been cut out from the electron density maps. One of the spheres clearly has significantly more electron density than the other, as one is a site deep within the protein, whereas the other is on the surface. These cuts have been taken from the 1FSE protein with a sigma level of 1σ (images created using *Coot*[23]), and the blue dot in the centre illustrates the location of the site around which a sphere is cut out..

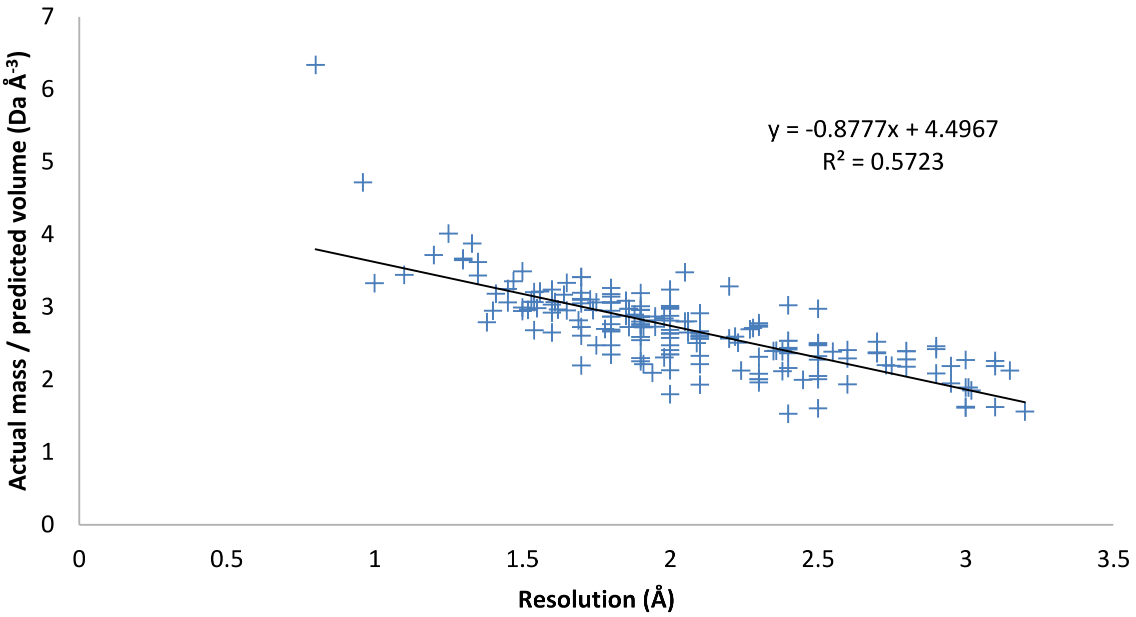
3.9.3.5 Upper limit: calculation of mass from electron density map

More generally, a different approach could be to consider the electron density map itself. The mass can then be determined by looking at the proportion of points in the search volume (sphere or NCS-averaged volume) that have an electron density greater than the background mean, and so are likely to be within the protein. All points that have an RMS higher than 0.5 σ were selected and defined as being within the electron density of the protein, and so the region of the protein could be defined. The value 0.5 σ was chosen from reviewing electron density histograms; it was found that points tended to fall under 0.2 σ or above 0.8 σ, corresponding to points outside or inside the protein respectively. A conversion factor is then required to relate this volume proportion and the mass.

Calculating this conversion factor presented some challenges. A suitable value was explored through investigating a collection of 200 random proteins, looking to see if there was a clear dependence between the measure and such a conversion factor, and searching for any resolution dependence. The test data used for the mass calculations were a random selection of 200 proteins from SQL database of X-ray targets created by Dr. Robert Oeffner in Prof. Rand Read's group, Cambridge Institute for Medical Research, University of Cambridge (unpublished data). There were a variety of resolution limits and number of residues, and the masses were calculated from the reported PDB files and the deposited experimental data.

As can been seen (Figure 3-21), there was a weak correlation between the conversion factor to the resolution of the data. There is a slighter better correlation with 1 / resolution ^{2}. However, this correlation remains relatively poor, with an R^2 value of only 0.6708.

(a)



(b)

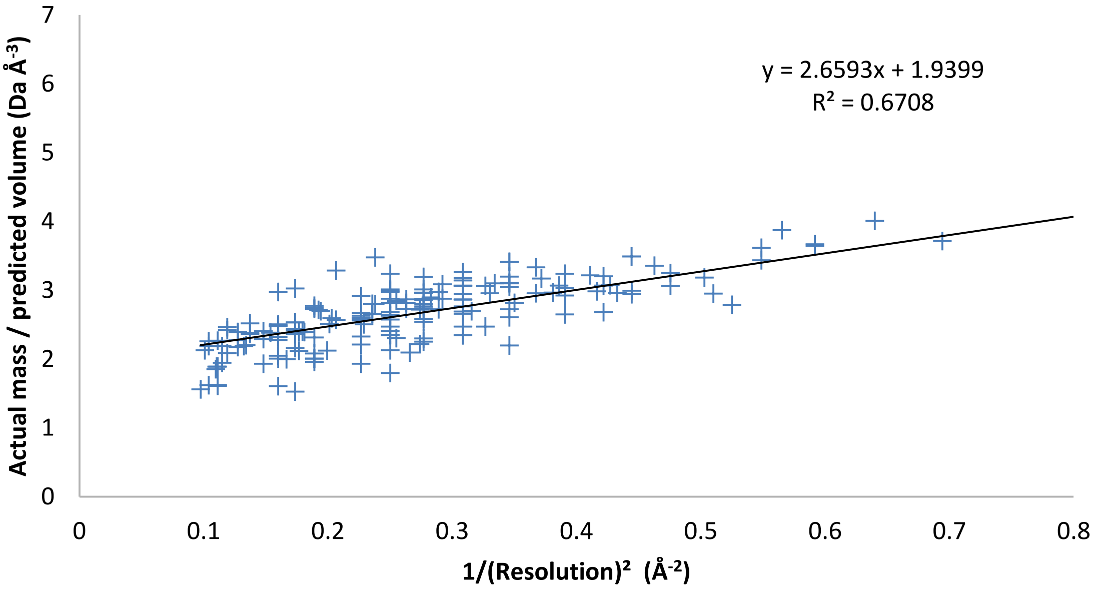


Figure 3-21 Two graphs showing the effect resolution has upon the conversion factor required to convert the predicted volume into mass

The correlation with 1/(Resolution)^2 is slightly better, and so was selected for use in calculating a mass. The graphs were created using *Microsoft Excel*.

The approach taken using this conversion factor did not predict the mass of the NCS-averaged monomer maps with any great accuracy. This lack of success is ascribed to an NCS-averaged map containing far more bulk solvent than would normally be expected. The sigma units are therefore underestimated, and so the RMS of the electron density will be higher than in a normal map. In order to predict correct values, it can be shown that the threshold value, 0.5, needs to be multiplied by $n^{0.5}$, where n is the number of NCS operators used in averaging, as explained below (Figure 3-22). Looking at each grid point in the electron density, a decision needs to be made if it represents protein or bulk solvent. A suitable threshold to make this determination is:

$$\mu(\rho) + 0.5 * \sqrt{n} * \sigma(\rho)$$

Equation 3-4

where the mean (μ) and the standard deviation (σ) is of the whole map being used as the search model.

The square root factor has been introduced to correct for the characteristics of an NCS-averaged map. In this case there is a large amount of empty space (Figure 3-22) in the search volume, which changes the standard deviation significantly.

This modified equation was derived by using the following model. If we assume *RMS(solvent)* = 0 and for the n monomers a value of *RMS(monomer)* is used in calculating the RMS of a map, then it can be stated that the overall RMS is just

$$RMS\big(topmap\ in\ \text{Figure 3-22}\big) = \sqrt{n(RMS(monomer))^2}$$

Equation 3-5

This then allows for the RMS of the NCS-averaged map used in the iterative approach (Figure 3-22) to be approximated as

$$RMS(bottomNCSaveragedmap\ in\ \text{Figure 3-22}) = \sqrt{(RMS(monomer))^2}$$

Equation 3-6

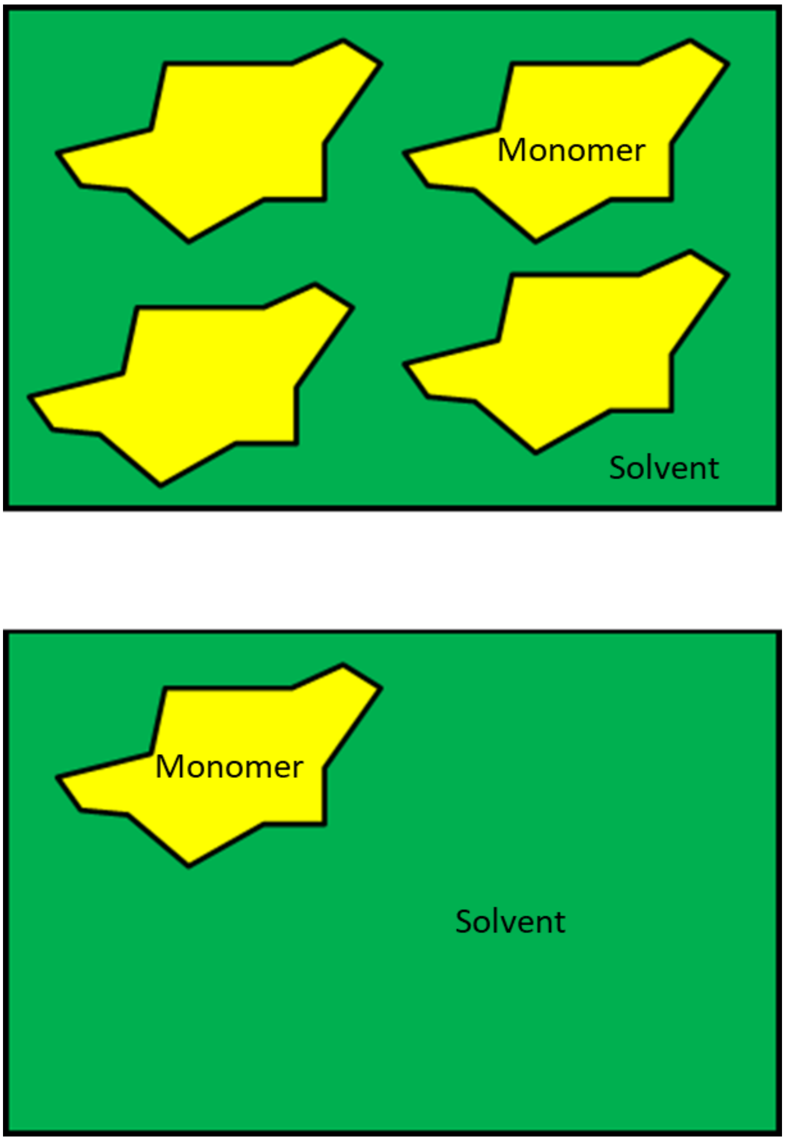


Figure 3-22 Reason for the change in σ scaling threshold

The lighter yellow part indicates the protein, whilst the darker green part shows the solvent within an area representing a unit cell. This diagram shows how much more protein there is in the original map (upper image) compared to the NCS averaged map (lower image). Therefore, with solvent having little variation in mass and so no standard deviation, the overall standard deviation of the second map is much lower than the first. The images were created in *Microsoft PowerPoint.*

It is therefore clear that in order to contain the same volume, the RMS for the NCS-averaged monomer would need to be adjusted by the extra √*n* parameter.

In summary, all of these upper limits were applied in *NaCelleS*, and the lowest predicted mass used. This leads to what seems to be the most consistent way of calculating mass.

3.9.4 Lack of phase information in the rotation function

NaCelleS efficiency is also limited by *phaser*[4] ignoring phase information from the input electron density map in the rotation search function. Unlike the phased translation search which takes existing phase information in the input map, the rotation search does not take advantage of these data, except in determining the site around which the electron density is cut. This rejection of available information makes the rotation search much less successful, and less likely to contain the correct rotation values. This is likely to be one of the reasons why the iteration approach had variable success in the different cases tested.

3.9.5 Time

Processing time for results can be a limiting factor in some of the *NaCelleS* functions. One of the key advantages that *phenix.find_ncs_from_density*[2] offers is that it maintains its speed at producing results, usually taking a couple of minutes to generate its output. The one-sphere approach within *NaCelleS* can be as quick as *phenix.find_ncs_from_density*[2] if it is a case which is readily solvable. However, *phaser*[4] can take longer to complete its processing, and still only find as many results as *phenix.find_ncs_from_density*[2]. *NaCelleS* can require more time when using the two-sphere approach, typically taking around an hour. The iterative approach can then be quite slow, especially if the averaging has included an incorrect operator.

Name	Two-sphere run-time (min)	One-sphere using *phenix.find_ncs_from_density* centre time (min)	One-sphere using *phenix.find_ncs_from_density* centre after two repeats through *NaCelleS* time (min)
02_1038B	164	34	N/A
03_1071B	37	11	10
04_cp_synthase_1	20	50	77
05_cp_synthase_2	19	8	207
06_cp_synthase_3	20	2	4
07_s-hydrolase_1	82	8	7
08_s-hydrolase_2	82	5	5
09_s-hydrolase_3	86	7	7
10_ut_synthase	67	261	N/A
15_aep_transaminase_1	224	18	39
18_rab3a_1_Se	13	2	10
19_rab3a_2_Se	12	10	10
20_rab3a_3_Se	12	18	13
21_rab3a_1_Zn	2	13	11
22_rab3a_2_Zn	2	23	10
23_rab3a_3_Zn	2	37	10
24_p32_1	9	23	6
25_p32_2	11	529	N/A
26_p32_3	11	49	6
27_cyanase_1	94	31	85
28_cyanase_2	75	30	64
29_cyanase_3	94	293	48
34_gpatase_1	132	20	37
35_gpatase_2	147	12	18
36_gpatase_3	137	34	N/A
37_pdz_1	20	3	14
44_ic_lyase_1	47	178	35
45_ic_lyase_2	66	293	N/A
46_ic_lyase_3	84	203	N/A
47_mpb_1	4	2	3
48_mpb_2	4	2	3
50_mpb_4	5	2	3

Table 3-7 Table showing the run-time taken for each algorithmic approach within NaCelleS

The two-sphere run-time represents the duration for each site, so this figure needs to be multiplied by the number of sites to reflect the processing time to take full advantage of this approach. N/A means that the run was not possible, as an iterative approach cannot be taken if *phenix.find_ncs* has not been able to make a NCS-averaged map.

3.9.6 Limitations at the anomalous scatterer centres

Processing times were looked at in detail with the larger dataset (Table 3-7). In all the cases examined, *phenix.find_ncs_from_density*[2] runs in under ten minutes. In cases where *NaCelleS* finds a strong solution, with a high ρ value (e.g. mpb), it can take a similar time or very slightly longer when using a single-sphere algorithm. In cases where it did not find a solution, it would spend much longer (e.g. 261 minutes with ut synthase). When using the two-sphere approach, the time recorded (Table 3-7) is for each individual site and therefore total time depends on the number of identified sites. The two-sphere approach can be seen to be significantly slower, but retains its advantage that it can identify NCS operators that are not found in the other programs.

Anomalous scatterer sites were occasionally located on the interface between protein and solvent, leading to lower signal-to-noise ratio, as half of the search sphere is just bulk solvent. In other cases the centres were found on interfaces between two monomers. If the protein has point group symmetry, this is not a problem. If it does not have this symmetry, issues can arise as the search includes a specific interface of two monomers, a configuration which might not exist between other monomers. Some anomalous scatterer sites might not be fully occupied and furthermore, in some anomalous scatterer soaks, the anomalous scatterer might not bind in the same place on different copies of the monomer. This makes it difficult for the two-sphere algorithm in *NaCelleS* to find NCS.

3.10 Conclusions

NaCelleS delivers similar performance to *phenix.find_ncs_from_density*[2] when using the one-sphere approach. In certain instances it can perform better (see earlier discussion), so it might be useful to be incorporate this algorithm as an option within the *phenix*[226] suite of programs.

In the two-sphere approach, *NaCelleS* provides additional functionality which is not available in other programs – it can tell whether the electron density around two anomalous scatterer sites are linked by NCS, and can calculate the relevant operators from this determination. This offers advantages in predicting such NCS operators, and is a useful addition to the armoury of programs within the macromolecular crystallographic community.

The iterative approach currently exhibits limitations, but can still be useful for refining operators in certain situations. Occasionally, it seems to be able to recover NCS operators that other methods do not find, and to improve them significantly. However, it is not reliably an improvement on *phenix.find_ncs_from_density*[2], so, without further work, it should be an option rather than a default for this program.

The two sphere approach offers additional options that would be beneficial to other scientists in finding NCS, so discussions are ongoing about incorporating it into *phenix*. The one-sphere approach is a different implementation of an existing method, and was extended to include an iterative approach, which requires additional validation effort before it can be incorporated into the released version of *phenix*.

Chapter 4 Effects of Data Pathologies on Molecular Replacement Solution

4.1 Rationale

Any X-ray crystallography experiment should attempt to obtain the best possible data. However, less than ideal data sets are frequently obtained, which can cause problems. It is important to understand the relative significance of different reflections in the diffraction pattern to ensure that the most appropriate experimental conditions are used. Key examples of such conditions are the size of the beam stops and the dynamic range and positioning of the detector. The former leads to the loss of low resolution data if the stops are too large, and overloading of the central zone of the detector due to overexposure with X-rays if the stops are too small. The latter requires an understanding of the relevant properties of the detector, and the balance between sensitivity (signal-to-noise in the measurement) and the higher resolution reflections that can be captured. Modern detectors are less susceptible to these effects, but they should still be kept in mind. Both these effects can lead to the greatest intensity reflections being poorly measured or being removed entirely from data sets. An important question is the impact of this loss on the attainment of a good solution using the molecular replacement technique in protein crystallography.

For over twenty years it has been assumed that low resolution data are very important for molecular replacement[293]. This assumption seems to have stemmed from a talk given by Gideon Davies[3] at a CCP4 study weekend, written up in the proceedings. In it he showed the effect of leaving out the most intense reflections on solving a molecular replacement problem. It was shown that, while originally the correct solution was easily found, removing just one intense reflection made it harder to find the desired solution. The removal of most of the most intense spots leads to the solution not being achieved.

In most datasets the most intense spots are also the reflections in the low resolution region. Unfortunately, the result has been remembered as the lowest resolution spots being the most important for molecular replacement studies. This seems to be intuitive as molecular replacement models are likely to be similar to the structure being sought in the general (low resolution) outline as opposed to the fine detail. Therefore it has been reasoned that the low resolution will pick out the general shape, whilst the high resolution picks out the fine detail. Consequently the high resolution spots will presumably be far less important in molecular replacement analysis, as the molecular replacement structures are likely to differ more in the fine detail. This hypothesis is tested in this work, and initial conclusions drawn.

Two cases were considered, a first to simulate loss of experimental data relating to low resolution information, or to incorrect collection of very intense reflections due to detector limits or similar factors, and a second to investigate the sensitivity of computational methods to low resolution or greatest intensity reflections in the collected information. In the case of the lower resolution data, the two are equivalent, as the relevant reflections can be removed from the data set. For the removal of a percentage of the more intense spots, the data normalisation function in the program comes into play, and is discussed further in Section 4.5.

An extension of the second case was also considered: whether the loss of information in the most intense reflections or the removal of the lowest resolution information affects programs that use maximum likelihood methods (such as *phaser*[4]) and programs that use older Patterson-based methods in different ways.

4.2 Factors to consider in selecting the beam stop in a diffraction experiment

Stops are used to absorb the X-rays that do not interact with the atoms in the test crystal (only approximately 3 % of the incident radiation will interact with the crystal[294]) and prevent this radiation from reaching and affecting the detector. As the incoming beam is not perfectly collimated and has a finite size, but the low resolution reflections are close to the incident axis, a choice has to be made in determining the size of the beam stop, as it can intercept these reflections and data can be lost. There is an associated problem of aligning the stop with the optical axis to ensure it performs its function, which is easier to achieve with a larger stop.

With modern synchrotron-based X-ray sources, beam collimation is good. This collimation is indeed so good that a single, optimised stop can be used for the majority of experiments, which are now typically run remotely. However, some experiments are conducted with a local, conventional X-ray source, where these considerations are more important. Accordingly, it is important to understand the relative importance of the low resolution reflections near the optic axis in assessing this trade-off.

4.3 X-ray detectors

X-ray detectors that can be used to capture the reflections in molecular crystallography have evolved since the technique was first used. Early experiments were with photographic film, which suffered from severe limitations. These detectors developed into online imaging plates, and thence to the use of charge coupled devices (CCD), flat panel detectors, and the modern pixel array detectors. Imaging plates provide a luminescent readout of diffraction patterns using scintillation material. They are commonly used because of their many other strengths: good linear response, high quantum efficiency, wide dynamic range, high spatial resolution, and low price tag. CCD detectors[295] convert the diffraction pattern into a digital image, offering good linear response, as well as low noise and high sensitivity. However, their response can be saturated by very intense X-ray diffraction, leading to a need to repeat an experiment (two-pass collection) to obtain sufficient dynamic range in the measurement of the intensity of the reflections. The pixel array detector[222,296,297] (used in this work) is especially useful for data collected using synchrotron X-ray sources. These detectors comprise an integrated circuit coupled to a diode detector. Their main advantages over CCD detectors are speed (high sensitivity) and a higher signal-to-noise ratio. This type of detector offers better spatial resolution than imaging plates due to the direct collection of charge rather than the use of a scintillation event, which can suffer from scattering of the emitted light.

A short summary of their main characteristics is included (Table 4-1) below.

Detector type	Advantages	Disadvantages
Off-line imaging plates	Large dynamic range, large area Inexpensive	Slow read-out (10-300 secs per image) Relatively broad point spread function (PSF) Cumbersome and no immediate feedback
On-line imaging plates	Large dynamic range, large area possible Relatively inexpensive	Slow read-out (10-300 secs per image) Relatively broad PSF
Tapered fibre optic CCD, tiled	Fast readout (<1-50 sec)	16-bit dynamic range typical Expensive for large solid angle work
Lens coupled CCD	Large active area (300mm) Relatively Inexpensive	Large and heavy detector Dynamic range limitations
Flat panel detectors	Large active area (400mm+) Fast readout (~seconds) Light and inexpensive	Noisy detectors
Pixel array detectors	Very good PSF Very fast readout	Difficult to tile to cover large area

Table 4-1: Summary of main characteristics of X-ray detectors suitable for macromolecular crystallography

The earlier detectors exhibited relatively broad point spread functions (PSF), which meant that they were less spatially selective, but they were also prone to overload when the incoming signal was very intense. This limitation in dynamic range made the choice of beam stop more important, and signals could saturate the detector and lead to challenges in interpretation. Overloaded spots can be identified and rejected or value limited, but they can be corrected as their expected profile is known. This profile can be used to allow their intensity to be estimated, but an approach based on two runs whereby overload is avoided by using a lower intensity beam to collect information on these spots used to be preferable.

The limitations in size mean that the positioning of the detector is important, as a detector placed further from the crystal is less prone to overload due to imperfect beam collimation, but would capture a smaller solid angle, and closer placement might lead to problems with resolving reflections and a larger influence of the main beam.

4.4 Obtaining complete data

The collection of data from as much of the reciprocal space as the resolution of the crystal allows, as required for the resolution of molecular structure, can be achieved by acquiring a large number of diffraction images at different angles of incidence between the primary beam of X-rays and the crystal. This is done by using a goniometer stage to support the crystal in the beam. One of the more common types of X-ray diffraction goniometers used today is known as a kappa goniometer, which characteristically permits four circular independent motions, and provides three angles of rotation for the crystal sample.

For each reflection that is recorded, a profile needs to be created, in order to accurately measure the intensity. This can be done in two different ways. Two-dimensional profile fitting builds profiles by looking at a single image. This is used particularly in coarse-angle search strategies, where most reflections are not spread out over multiple images, but appear on one image only. The other way is three-dimensional profile fitting, which builds up a profile across several adjacent images. This is particularly of use in fine slicing search strategies, where one reflection is spread out over multiple images.

4.5 Normalisation of data

The simple model of diffraction assumes that the elastic scattering sources (atoms) are simple points. In reality, the atoms have finite size comparable with the incoming X-ray wavelength, so the spread of their electron density will lead to a difference in path length depending on the 2θ angle. Considering both the X-ray and electron as particles, then the interaction can be considered to occur when the electron is closest to the source (shortest overall path), or perhaps when the electron is furthest from the source (longest path). This leads to a difference in the overall path to the detector, or a phase difference resulting from this path length difference (Figure 4-1).

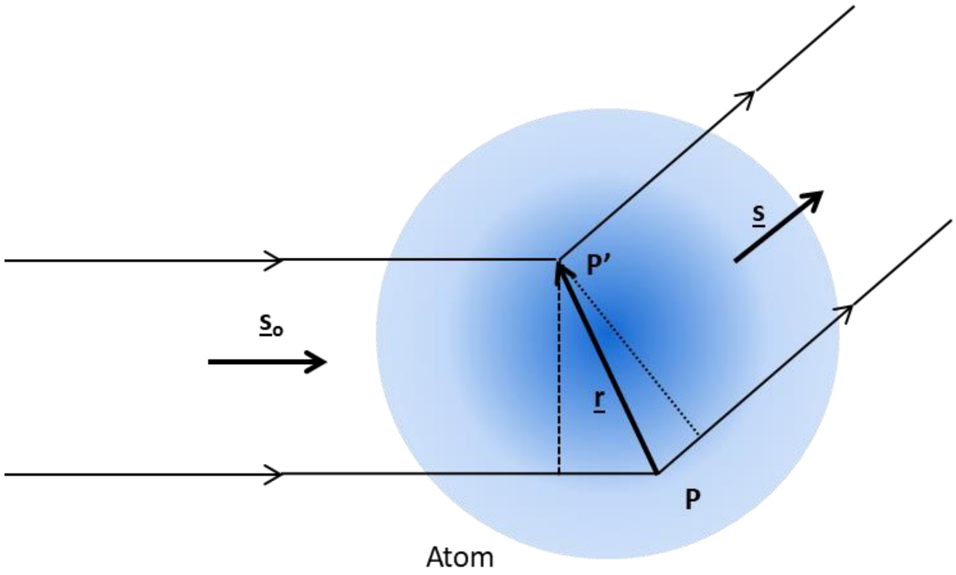


Figure 4-1: Path differences when atoms are not considered as point sources

The image was created in *Microsoft PowerPoint*.

This path difference (and resulting phase difference) is greater for larger 2θ angles, and larger phase differences lead to increases in destructive interference, with a resultant reduction in the intensity of the collected reflection. If all interactions across the individual atom are equally likely, this would result in a cosine squared distribution of intensity for atoms of equal size (same atomic number). In practice there are a large number of atoms of different atomic number in different chemical conditions (different bonds), so there is a complex electron distribution that will depend on the orientation of bonds and other factors. The reduction in intensity with increase in angle will continue to occur, but following a different distribution. Mosaicity within the crystal will introduce additional variation in the phase contribution from each of the crystallites within the overall crystal, contributing to angular variation of intensity among other effects. This angular reduction of intensity is important to this work as the measured intensity for each reflection does not represent the actual intensity if we assume a point reflector (the ideal representation of elastic scattering). Accordingly, some form of normalisation is required to ensure that the information is weighted correctly to allow comparison between different structure factors.

A second, smaller effect is due to the difference in path length to the detector. This means that reflections at higher angles will vary in intensity at a fixed detector pixel size following the inverse square law of radiation propagation.

These effects need to be taken into account in comparing the different reflections. This could be done by applying a correction factor to each reflection, knowing the deflection angle θ. This correction could be derived from theoretical considerations, or by a modified value based on experimental results, but this rapidly becomes complex for real molecules.

A pragmatic approach is taken in *phaser*[4] to allow for these effects, where the reflections are divided into a series of spherical shells corresponding to different resolutions, centred on the nominal *F(0,0,0)* position. For each shell, the mean intensity of the reflections contained therein is calculated, and the value of each reflection is normalised by dividing by this mean. The maximum number of shells is determined by the need to have sufficient reflections in a shell to generate an acceptable mean and range of normalised values, and the minimum by the need to allow for the drop off in intensity with angle to be well modelled. The normalised amplitudes, which are derived from the normalised intensities, are then referenced by the notation *E(h,k,l)* rather than *F(h,k,l).*

In the case of the lower resolution investigations, normalisation has no effect, as *phaser*[4] will select the shells it uses to perform this function with the data it is given. Accordingly, the normalisation function will still work correctly if the data below a given resolution are removed, as this is independent of the other data, which is allocated to different shells. Therefore both the experimental (data lost in the collection) and computational case (investigation into the relative importance of these data) are being investigated simultaneously, as the cases are equivalent.

In the case of the removal of the most intense points in the diffraction pattern, the situation is quite different, as this can change the distribution of data points within every shell. Removal of the data for the most intense reflections before normalisation would be the closest approximation to the experimental case where these data points are not collected (first case). For the computational investigations, it would be better to remove the reflections after normalisation, as this would then not change the other normalised values and generate a better like-for-like comparison. Under these conditions, the information lost by removing these most intense reflections can be assessed.

4.6 Experimental Results

A molecular replacement test set was extracted from the large suite of test data that has been kindly created and curated by Dr Robert Oeffner in Professor Randy Read's group within the Cambridge Institute for Medical Research. The rules of selection to create a useful set of molecular replacement cases relating to the effects of data pathologies were the following: the dataset must have originally been solved with a TFZ0 score between 7 and 10, and the molecule must be over 400 residues in size. The TFZ0 score is the initial Z-score for the translation search. The resultant set used contained 217 molecular replacement cases, with the data high resolution limit ranging from 0.97 Å to 4 Å.

These criteria resulted in the selection of test cases with solutions that had marginal TFZ scores, so the loss of a few reflections caused either by removing low resolution data or the most intense overall reflections should lead to a number becoming unsolvable. This would therefore allow an investigation of the extent to which losing the low resolution data or the greatest intensity spots would affect the ability to find the solution.

Extraction of these test cases was achieved using specially written *python* code, calling on standard libraries in *phenix*[226] where necessary. SVN version 5646 of *phaser*[4] (also referred to as *phaser*[4] 2013) was then used to calculate the LLG, TFZ and report the solutions that were found for each case. This version could use either maximum likelihood methods or Patterson-based methods (not available in the version on general release; the Patterson-based algorithms have not been maintained in later versions). These results could be compared with a known solution which had been generated in earlier work, and so permitted the identification of successful molecular replacement trials, and cases where solutions were not found. The exclusion of specific reflections was implemented within *phaser*[4], both when this was done to test the effect of removing reflections at a lower resolution than a specified limit (already available and documented), or the removal of the greatest intensity reflections (test feature only, not released). Version 2.8.0 (also referred to as *phaser*[4] 2016) was used in experiments that excluded data before normalisation.

The stated hypothesis that low resolution data were uniquely important was then tested by removing some of the reflections and determining if a full solution could still be found. This was done either by removing the low resolution data, with resolution cut-off limits of 15 Å (used as a test to ensure that the program was performing as expected, as the impact at this level should be minimal), 8 Å and 6 Å, or by removing the 5 % most intense reflections from the data sets.

Python scripts were developed to calculate the number of cases where the best match solution was the same as the known solution, the mean TFZ, the mean LLG and the average number of solutions, for each case. The data were then handled initially through spreadsheets in *Libre Office*, and then through *R* and *Microsoft Excel 2010*.

For this process to generate a fair comparison, a check had to be done to ensure that the number of reflections being eliminated was similar in both cases i.e. for the different resolution limits when taking out reflections, and for the removal of a selected percentage of the most intense reflections. Initial trials with the data revealed that an elimination of 5 % of the most intense spots resulted in a significant number of failed solutions, so this figure was selected for the tests. The average fraction of reflections remaining was calculated for both the greatest intensity limit and the low resolution limit (Figure 4-2).

As can be seen, the 6 Å resolution limit removes more reflections (94 % remaining) than the case when 5 % of the most intense reflections are removed (95 % remaining). By definition, when no resolution filters are applied and no intensity filter is applied to the reflections then 100 % of the reflections remain. The most reflections are removed when both a 6 Å resolution limit is imposed and 5 % most intense reflections are removed.

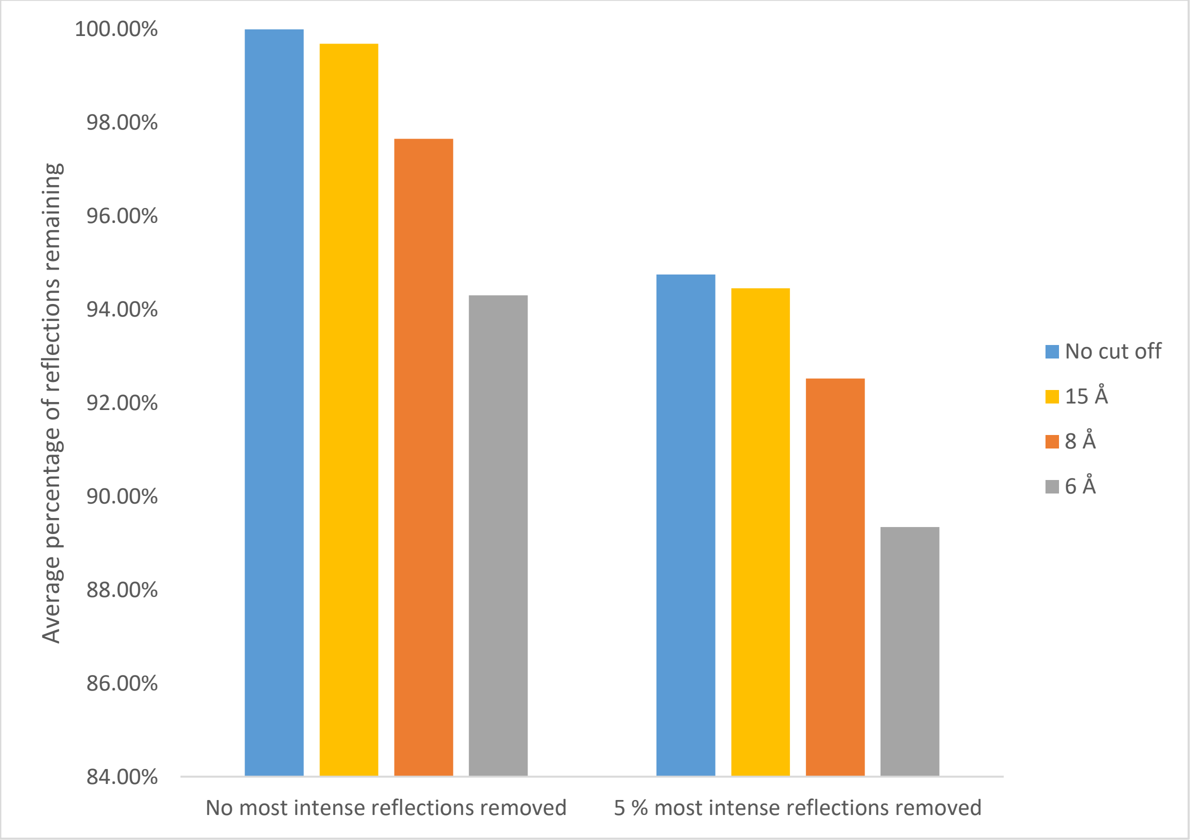


Figure 4-2 ***A****verage percentage of the reflections remaining with different limits for removing data applied*

This shows the average percentage of the reflections remaining with different applied resolution limits (none, 15 Å, 8 Å and 6 Å), with and without the 5 % most intense reflections being removed. The charts were prepared in *Microsoft Excel*.

4.6.1 Measuring the success of molecular replacement solution

There are a number of different measures available to determine if a molecular replacement solution is correct under the different test conditions. The simplest is to examine how many cases matched the known solution for the first item in the returned solution list. This gives a rapid initial method of determining how the exclusion of different reflections in a pattern leads to a change in the ease of finding a solution.

A second good measure of success is to use the Translation Function Z-score (TFZ), and to gain a single value across all test sets, the TFZ of the top solution from each test set is taken and averaged together to give an average TFZ. This average TFZ value is a good indication of how well structures solve on the whole. Low averages in TFZ indicate that few structures could be correctly placed.

A third measure that can be used is the average number of solutions reported. The more uncertain *phaser*[4] is in evaluating the solutions for a particular molecular replacement case (i.e. there are several solutions of generally similar likelihoods), the more solutions it tends to report. On the whole this measure supports the information provided by the average TFZ and the number of correct solutions. However, being a more indirect measure, once the mean rises to a number around 50, the values reported are noisy and can fluctuate randomly. Hence, in cases where most molecular replacement test cases are seen to fail, it is limited as a measure for determining the difference under the conditions tested.

The log-likelihood gain (LLG) was the final measure that was used. LLG is the probability that the data would have been measured, given the model. This particular measure is very good at determining how good the solution is and how likely it is to be a correct solution, with higher scores giving more certainty[298]. Unfortunately, problems occurred with the implementation of this measure in *phaser*[4] under certain circumstances. A number of tests using *phaser*[4] 2013 which had the ability to use Patterson-based methods revealed that there were significant limitations in the refinement of solutions where parts of the data had been left out. This led to LLG scores not being accurately recorded. This has been corrected in the latest version, but this no longer supports Patterson-based methods. Therefore, apart from providing a qualitative assessment of the difference between Patterson-based methods and Maximum Likelihood Methods, this score had limited use for this analysis.

4.6.2 Use of different versions of phaser

Two different versions of *phaser*[4] were used in obtaining these results. A 2013 version (SVN 5646) allowed the most intense reflections to be removed within the program after normalisation (relevant for the computational case), whereas the 2016 version (Version 2.8.0) does not. This was performed using the unpublished environmental variable “PHASER_CULL_BRIGHT 0.05”. Removal of data before normalisation was performed by *SFTOOLS*, controlled by dedicated *C-shell* or *Python* code as required (relevant for the experimental case).

The earlier version had significant weaknesses in the handling of uncertainty and errors in the data, especially in the handling of phase errors, and code errors in some relevant sections which meant that removing reflections before normalisation was problematic. This led to data sets with less low resolution information proving more difficult to solve. The position was greatly improved in the later release, which deals with data difficulties much more comprehensively, especially in the case of missing data points.

In *phaser*[4] 2013, Patterson-based methods were invoked by setting the unpublished keywords “TARG ROT CROWTHER” and “TARG TRA CORRELATION”. The low resolution limits were set by the keywords “RESOLUTION LOW x” and “RESOLUTION AUTO LOW x”, where x is the desired low resolution limits.

4.6.3 Results achieved using maximum likelihood methods

phaser[4] has been optimised to use maximum likelihood methods, so these methods were initially applied to test the hypothesis. Experiments using different levels of removal of the low resolution data are reported in Section 4.6.3.1, and those where the most intense reflections are removed in Section 4.6.3.2. A combination of both data removals is then investigated in Section 4.6.3.3, and the results examined and compared in Section 4.6.3.4.

4.6.3.1 Removing low resolution reflections from the datasets

The effect of placing an artificial limit on the low resolution data was explored using three different limits, 15 Å, 8 Å and 6 Å. Any reflections of a lower resolution than the selected value were discarded. Normalisation has no effect on the decision criteria, or on the overall elimination of reflections. Results were taken in both cases using different versions of *phaser*[4]. Results using *phaser*[4] 2013 are considered first.

	2013 *phaser*				2016 *phaser*
	Mean TFZ	Mean LLG	Mean size of solution list	Number of matched solutions	Number of matched solutions
No cut off	8.4	73.7	3.3	191	185
15 Å	8.5	87.8	3.3	189	NRF
8 Å	7.2	65.3	25.3	168	175
6 Å	6.9	69.3	33.8	156	173

Table 4-2: Summary results when low resolution reflections are removed versus original data for maximum likelihood methods. NRF is No Reported Factors, as the 15 Å resolution limit was only a check, and not repeated when using the 2016 phaser.

If we now look at the results in more detail, the number of matched solutions with all the data (191 and 185 in the 2013 and 2016 versions of *phaser* respectively) is less than the number of test cases (217). In these test cases, there were a cluster of similar solutions at the top of the solutions list, and so the correct solution was actually second or third in the set, and so not counted in the number of matched solutions. The main measure of interest is therefore the relative change, and not the absolute number.

The effect of the resolution filter can clearly be seen (Table 4-2, Figure 4-3 and Figure 4-4). As the filter is made harsher (i.e. more lower resolution data eliminated), the results deteriorated significantly. The 15 Å case removed relatively few data points, and was there as a check that the elimination process was operating as expected. Equally, setting the filter at 5 Å or less resulted in large numbers of reflections being removed, and the results proved difficult to interpret.

As harsher resolution limits are applied, the molecular replacement process is no longer reporting the same most likely solution as the known solution. When compared to the 191 initially successful solutions, the 15 Å case lost only two solutions, showing that the process of removing this small number of reflections had little effect. With the 8 Å and 6 Å limits, this figure reduced to 168 and 156 solutions respectively, indicating some effect, but still relatively small.

The mean number of solutions reported did not change when the 15 Å filter was imposed, indicating that the process was insensitive to these lowest resolution reflections. There were significant changes to this measure for the 8 Å and 6 Å experiments, with the mean increasing to 25.3 and 33.8 respectively, compared to the original value of 3.3. This indicated that the algorithms within *phaser*[4] are not being able to discriminate as effectively between potential solutions with the removal of the low resolution information.

Under the harsher resolution conditions, the TFZ score decreases, indicating lower signal and therefore success in the molecular replacement process is reducing. The slight increase in the TFZ in the 15 Å case is not fully understood, but is not considered to be significant. The guidelines for interpreting the TFZ scores suggest that values above 8 indicate a successful solution, values between 7-8 indicate a probably correct solution, values between 6-7 a possibly correct solution and lower values are unlikely to be true representations. The full data average result for these tests is 8.4, indicating a good solution. For the 8 Å and 6 Å case, the value decreases to 7.2 and 6.9, indicating less certainty as expected.

In the set of results obtained using *phaser*[4] 2016, it is clear that there have been significant improvements in the software. On applying the lower resolution limits, very little percentage difference was seen in the number of solutions where the highest scoring result matched the known solutions, following an initial loss of around ten structures. The drop in the number solved as harsher resolution limits were applied was significantly less for the newer version, though the test case where the original data sets were rerun returned more missed solutions.

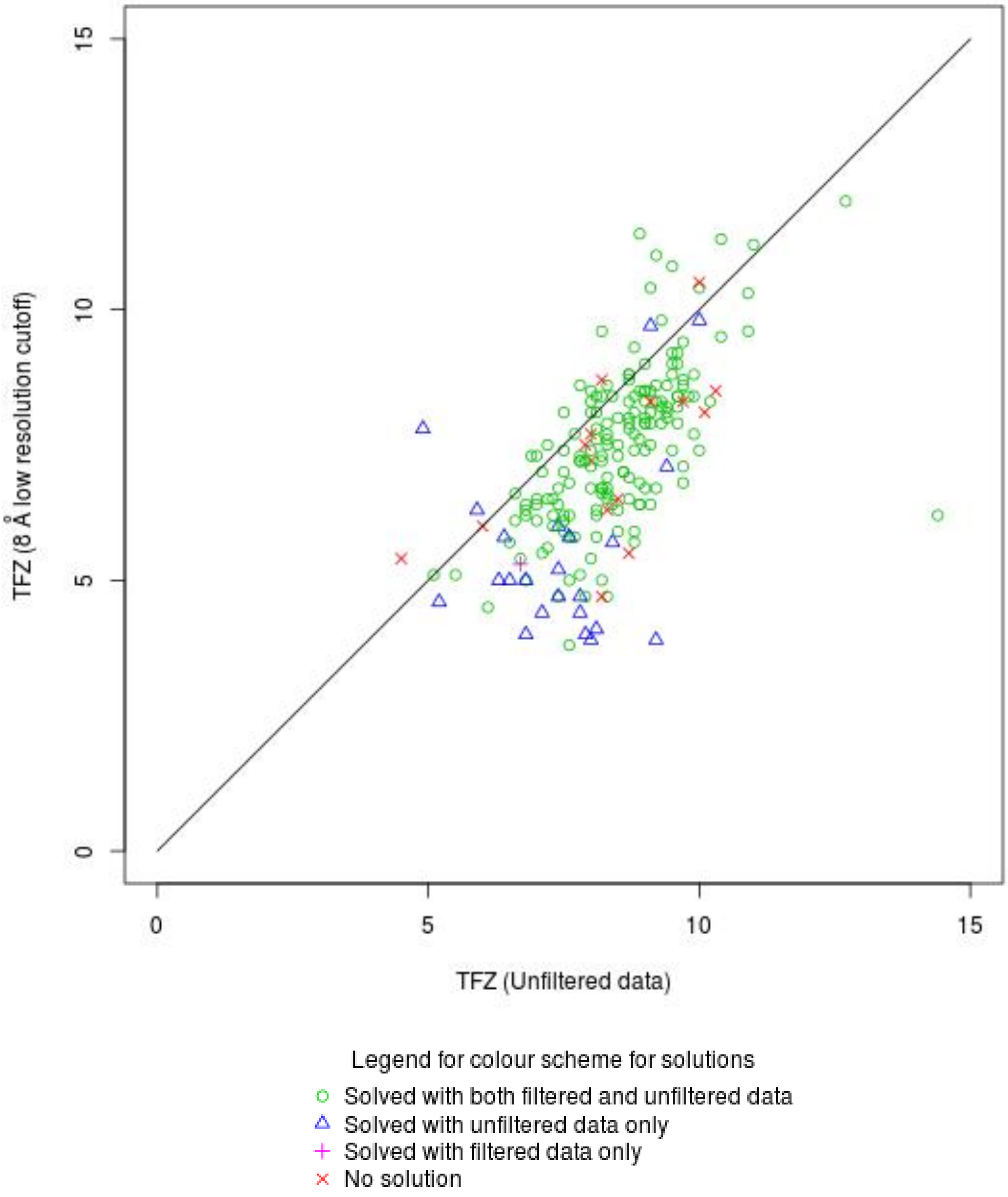


Figure 4-3: TFZ scores for 8 Å low resolution filtered data versus unfiltered data for maximum likelihood methods

The clustering of points below the line indicates that removal of the low dimension data reduces the quality of the results. A number of test cases are not solved (blue triangles), but had been in the unfiltered data. The plot was created in *R*.

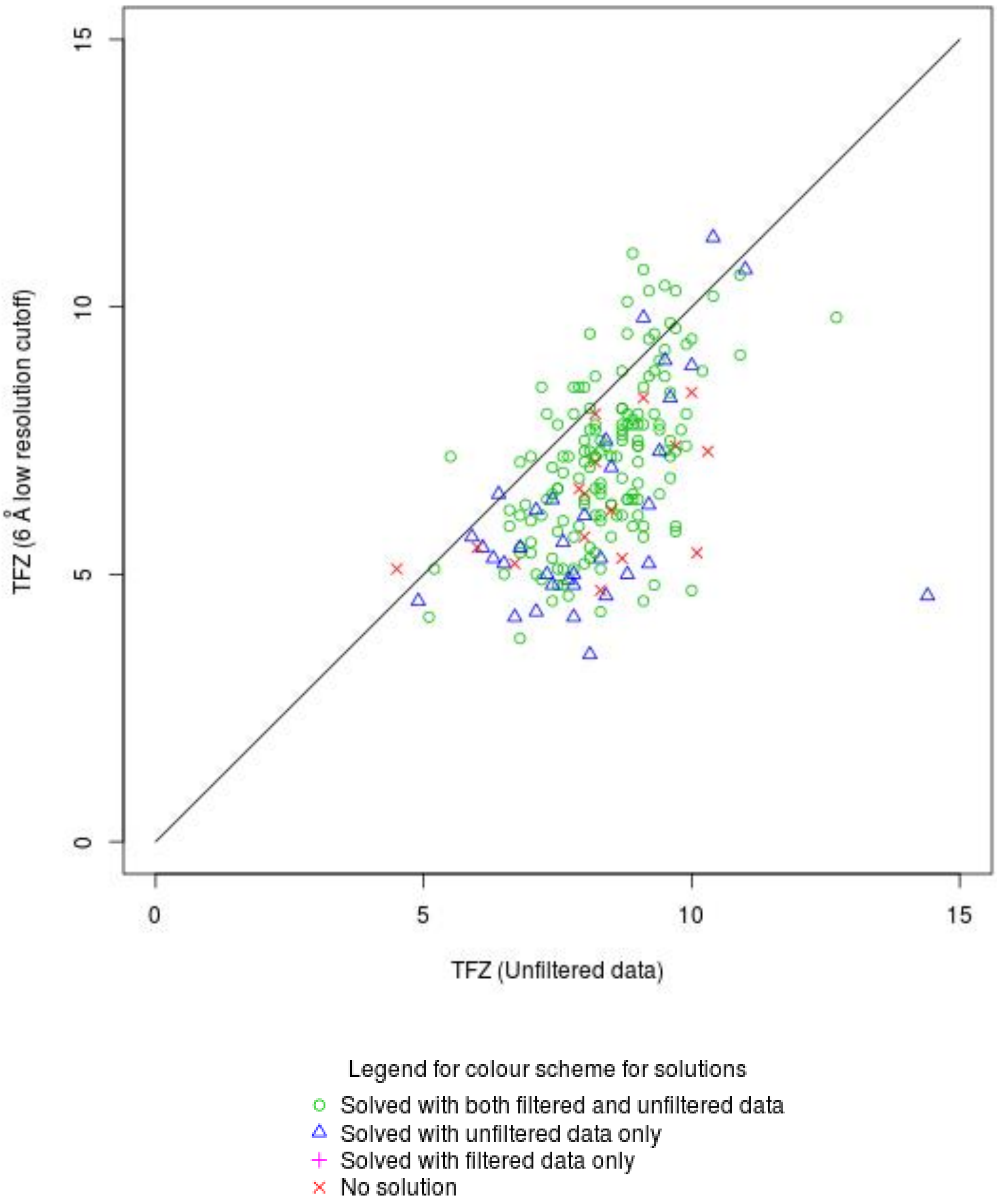


Figure 4-4: TFZ scores for 6 Å low resolution filtered data versus unfiltered data for maximum likelihood methods

The clustering of points has moved further below the line, indicating that the more aggressive removal of the low dimension data have had a more significant impact on the quality of the results. A larger number of test cases are also now not solved (blue triangles). The plot was created in *R*.

The results achieved in the 2016 version of *phaser*[4] were compared with the results from the earlier version, but the only fully valid measure is the number of solved solutions. Other measures changed significantly between versions as they depend on the implementation of the maximum likelihood functions, and were not considered to be a useful comparison.

4.6.3.2 Removing 5 % of the most intense reflections from the datasets

Removal of the most intense reflections in the diffraction pattern from the dataset can be done before or after normalisation, as required for the case under consideration. Removal before normalisation would be the closest approximation to the experimental case, whereas removal after normalisation is more suited to the investigation of the impact of this data on the ability to find solutions (computational case). This is discussed in Section 4.5.

With the most intense 5 % of reflections removed after normalisation (*phaser*[4] 2013 version), clear degradation in the computational results could be seen (Table 4-3 and Figure 4-5). The number of solved structures reduces significantly from 191 to 135, showing a large degradation in the effective information content. Considering the mean size of the solution list, it has risen from 3.3 to 10.2. This loose measure shows that *phaser*[4] is struggling under these conditions, though this is an inaccurate measure, as complete failures do not influence this value. For TFZ scores, the mean has dropped from 8.4 (a value reflecting definite solutions) to 5.7 (rating degraded to only possible solutions).

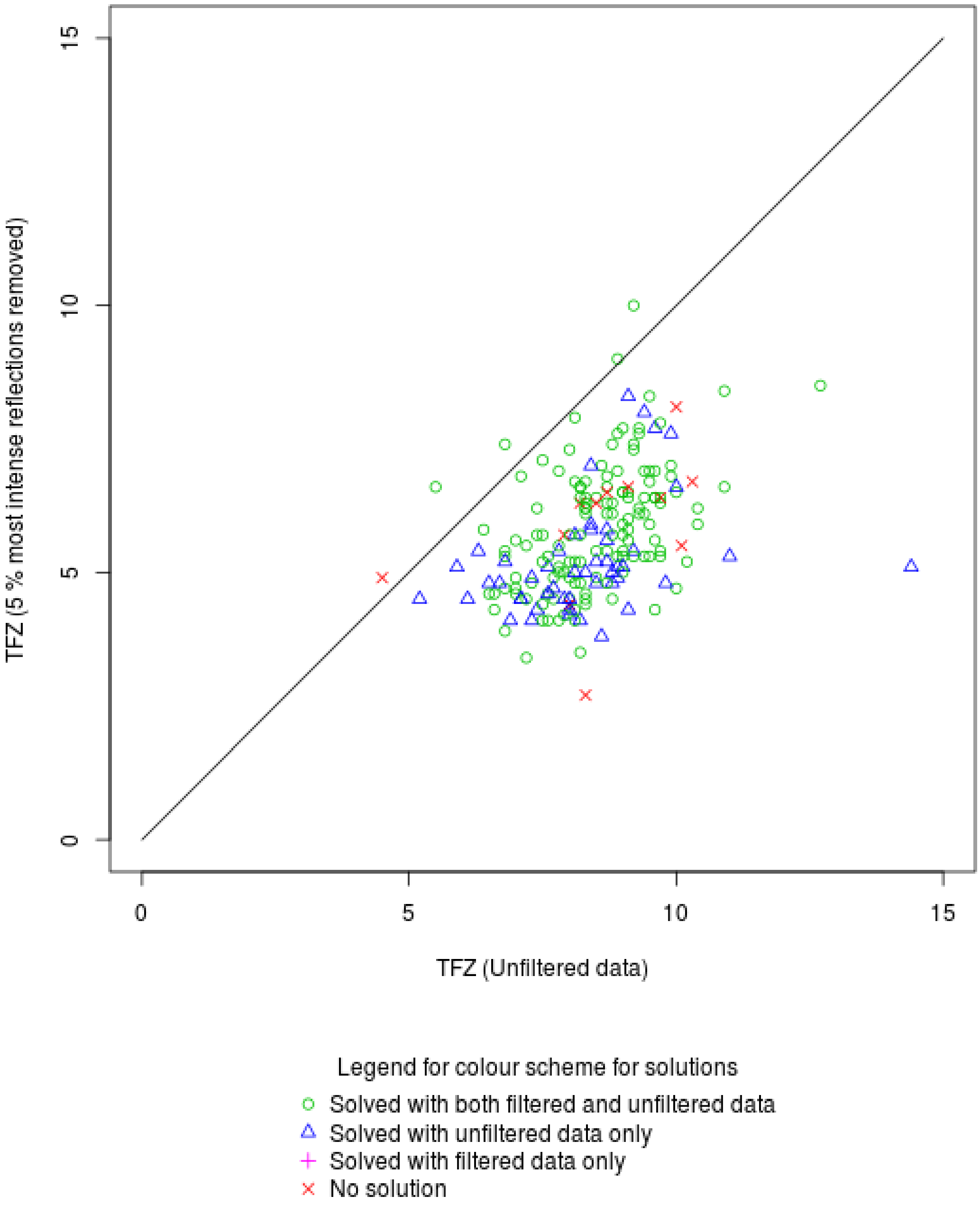


Figure 4-5: TFZ scores for 5 % most intense removed data versus unfiltered data for maximum likelihood methods

The clustering of points is now well below the line, indicating that removal of the data associated with the most intense reflections greatly reduces the quality of the results. A large number of test cases are not solved (blue triangles) under these conditions. The plot was created in *R*.

	Original Data	5 % most intense reflections removed
phaser 2013 Mean TFZ	8.4	5.7
phaser 2013 Mean LLG	73.7	156.8
phaser 2013 Mean size of solution list	3.3	10.2
phaser 2013 Number of matched solutions	191	135
phaser 2016 Number of matched solutions	185	151

Table 4-3: Summary results when 5 % most intense reflections are removed versus original data for maximum likelihood methods

When the same percentage of most intense reflections is removed before normalisation (*phaser*[4] 2016 version), the reduction in the number of solutions is less. This indicates that this version of the software is much more resistant to a systematic reduction in the input data.

4.6.3.3 Removing 5 % most intense reflections and applying low resolution cut-offs at the same time

Experiments were run to investigate the effect of removing the 5 % most intense reflections and applying a low resolution cut-off simultaneously. The degradation in the ease of molecular replacement solution was seen to be cumulative (Table 4-4). The strong effect seen earlier when the 5 % most intense reflections are removed (Section 4.6.3.2) is added to the weak effect of the low resolution cut-off (Section 4.6.3.1).

Data removed	2013 *phaser*				2016 *phaser*
	Mean TFZ	Mean LLG	Mean size of solution list	Number of matched solutions	Number of matched solutions
5 % + 15 Å	5.7	115.8	9.9	133	
5 % + 8 Å	5.3	105.8	27.6	103	175
5 % + 6 Å	5.3	108.3	34.6	95	173

Table 4-4: Summary results when 5 % most intense reflections and low resolution reflections are removed for maximum likelihood methods

The strongest measure supporting the statement above is the number of structures where the most significant result matched the known solution. Considering the work with *phaser*[4] 2016 first, where the reflections were removed before normalisation (modelling the experimental case), adding the resolution limits to the removal of the 5 % most intense reflections, results in the number of solutions dropping from 151 to 143, a similar relative decrease to that seen without this initial removal of reflections.

The results achieved with *phaser*[4] 2013 also showed the cumulative effect of the two data deletions. The large initial loss of solutions caused by the removal of 5 % of the most intense reflections was increased by smaller numbers as the more harsh resolution criteria were applied. As has been seen previously, the 15 Å resolution cut-off has little effect.

The other measures do not give clear indications of the cumulative effect. The mean TFZ stays approximately the same, with a slight reduction for higher resolution limits. The mean number of solutions with the 5 % most intense reflections removed did increase as the resolution cut-off was increased, again supporting the cumulative effect of these data eliminations.

4.6.3.4 Comparison of the effects of removing the low resolution data and the most intense spots

The most appropriate comparison is between the 6 Å resolution limit and the 5 % most intense reflection removal, as this has approximately the same percentage of reflections remaining (94 % versus 95 %). The results were similar for both versions of *phaser*[4], but are discussed in more detail below.

Results obtained from the 2013 version of *phaser*[4] are more suitable for considering case 2, the removal of data after normalisation. The mean number of solutions with the maximum likelihood score matching the known solution differed significantly for the two data eliminations. It was higher (156 cases) in the datasets with the selected 6 Å low resolution data cut off, than in the datasets where 5 % most intense reflections were removed (135 cases). The mean TFZ in the data was also very different – 6.9 in the low resolution case, and 5.7 in the case where the most intense resolutions were removed. Both of these suggest that the removal of the 5 % most intense reflections was having more of an effect than the removal of the low resolution data.

However, the mean size of the solution list, showed the opposite effect. Here the mean size when removing the most intense reflections was 10.2, but when removing the low resolution data it was 33.8.

In the *phaser*[4] 2016 version, even with the superior error management, removing the most intense reflections also had most impact on the number of successful solutions, though the program was generally more successful and differences less noticeable. Here the number of most likely solutions that matched the known solutions for the low resolution cut off was 173, but for the most intense reflections removed was 151.

4.6.4 Results achieved using Patterson-based methods

Patterson-based methods were developed before maximum likelihood methods emerged, and discussions indicated a perception that low resolution data were important when this methodology was used. Patterson-based methods were also incorporated into *phaser*[4], but are undocumented and not as completely exercised. However, they were available and appeared to offer adequate functionality to test the hypothesis. Note that normalisation is not a consideration when using Patterson-based methods, as the inverse Fourier transform is taken of the intensities as a whole. Experiments using this method with different levels of removal of the low resolution data (Section 4.6.4.1) and of most intense reflections (Section 4.6.4.2) are reported. A combination of both data removals is then investigated (Section 4.6.4.3), and the results examined and compared (Section 4.6.4.4).

4.6.4.1 Removing low resolution reflections from the datasets

The same three limits, 15 Å, 8 Å and 6 Å, were used to explore the placing of an artificial limit on the low resolution data. Any reflections of a lower resolution than the selected value were discarded.

Looking at the results in detail (Table 4-5, Figure 4-6 and Figure 4-7), the effect of the resolution filter can clearly be seen. As the filter is made harsher (i.e. additional lower resolution data are eliminated), the results deteriorate significantly. The 15 Å case again removed relatively few data points, and was primarily used as a check that the elimination process was operating as expected with the Patterson-based codes.

	Mean TFZ	Mean LLG	Mean size of solution list	Number of matched solutions
No cut off	7.6	68.8	15.5	180
15 Å	7.5	81.7	19.8	174
8 Å	6.6	59.1	36.8	150
6 Å	6.4	66.3	59.8	136

Table 4-5: Summary results when low resolution reflections are removed versus original data for Patterson-based methods

LLG scores are calculated in *phaser* for Patterson-based methods as they are for maximum likelihood methods, but they are not used for solution determination. They are on the same scale so are directly comparable.

As harsher resolution limits are applied, the molecular replacement process no longer reports the same most likely solution as the known solution. When compared to the 180 initially successful solutions, the 15 Å case lost only six solutions, showing that the process of removing this small number of solutions had a minimal effect. With the 8 Å and 6 Å limits, this figure reduced to 150 and 136 solutions respectively, indicating that these methods had increasing difficulty with the increased loss of low resolution data, but it was by no means catastrophic.

The mean number of solutions reported is a less useful measure in this implementation of the Patterson-based methods, because the filter to find poor solutions was not implemented to an equivalent degree as for maximum likelihood methods. However, the same increasing trend with loss of low resolution data in the 15 Å, 8 Å and 6 Å cases was seen. For 15 Å, as small increase from 15.5 to 19.8 was observed, but at 8 Å this increased to 36.8, and at 6 Å to 59.8. This indicated that the Patterson-based methods within *phaser*[4] are not able to discriminate between potential solutions as well when the low resolution information is removed.

The TFZ score decreases under the harsher resolution conditions, showing that the success achieved in performing the molecular replacement process is reducing. The full data average result for these tests is 7.6, indicating a probable solution. The 15 Å case shows little difference, with a figure of 7.5 being reported. For the 8 Å and 6 Å case, the value decreases to 6.6 and 6.4, indicating less certainty, as expected.

4.6.4.2 Removing 5 % of the most intense reflections from the datasets

With the most intense 5 % of reflections removed, clear degradation could be seen in the computational results (Table 4-6 and Figure 4-8). The number of solved structures reduces significantly from 180 to 31, showing a very large degradation in the effective information content. Considering the mean size of the solution list, it has risen from 15.5 to 75.4. These high values indicate that the implementation of Patterson-based methods in *phaser*[4] is struggling in these tests. Additionally, this remains an inaccurate measure as complete failures are not included in this value. For TFZ scores, the mean has dropped from 7.6 (a value reflecting a probable solution) to 4.5 (rating degraded to unlikely solutions).

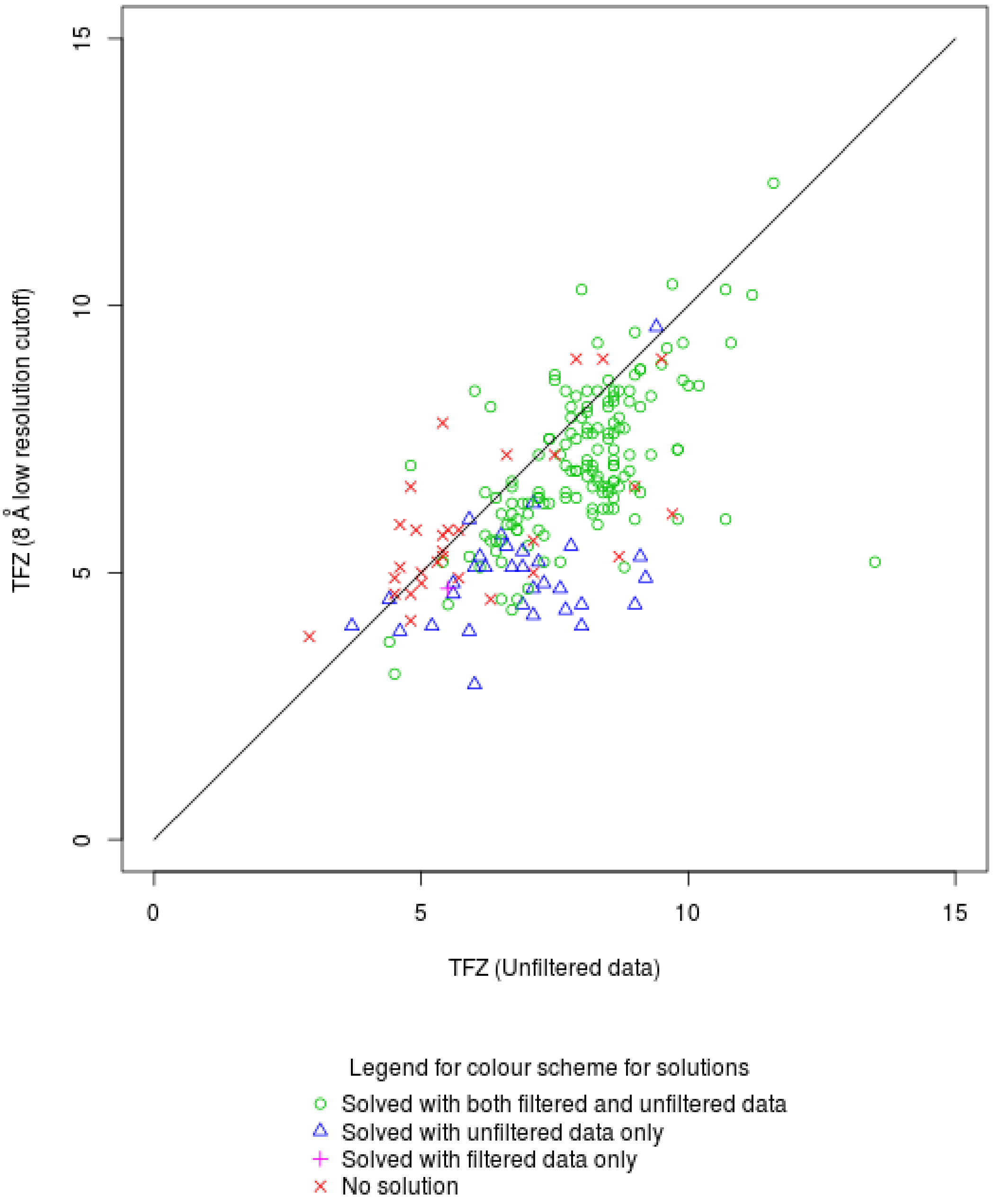


Figure 4-6: TFZ scores for 8 Å low resolution filtered data versus unfiltered data for Patterson-based methods

The clustering of points below the line indicates that removal of the low dimension data reduces the quality of the results when using Patterson-based methods. A significant number of test cases are not solved (blue triangles) with the filtered data. The plot was created in *R*.

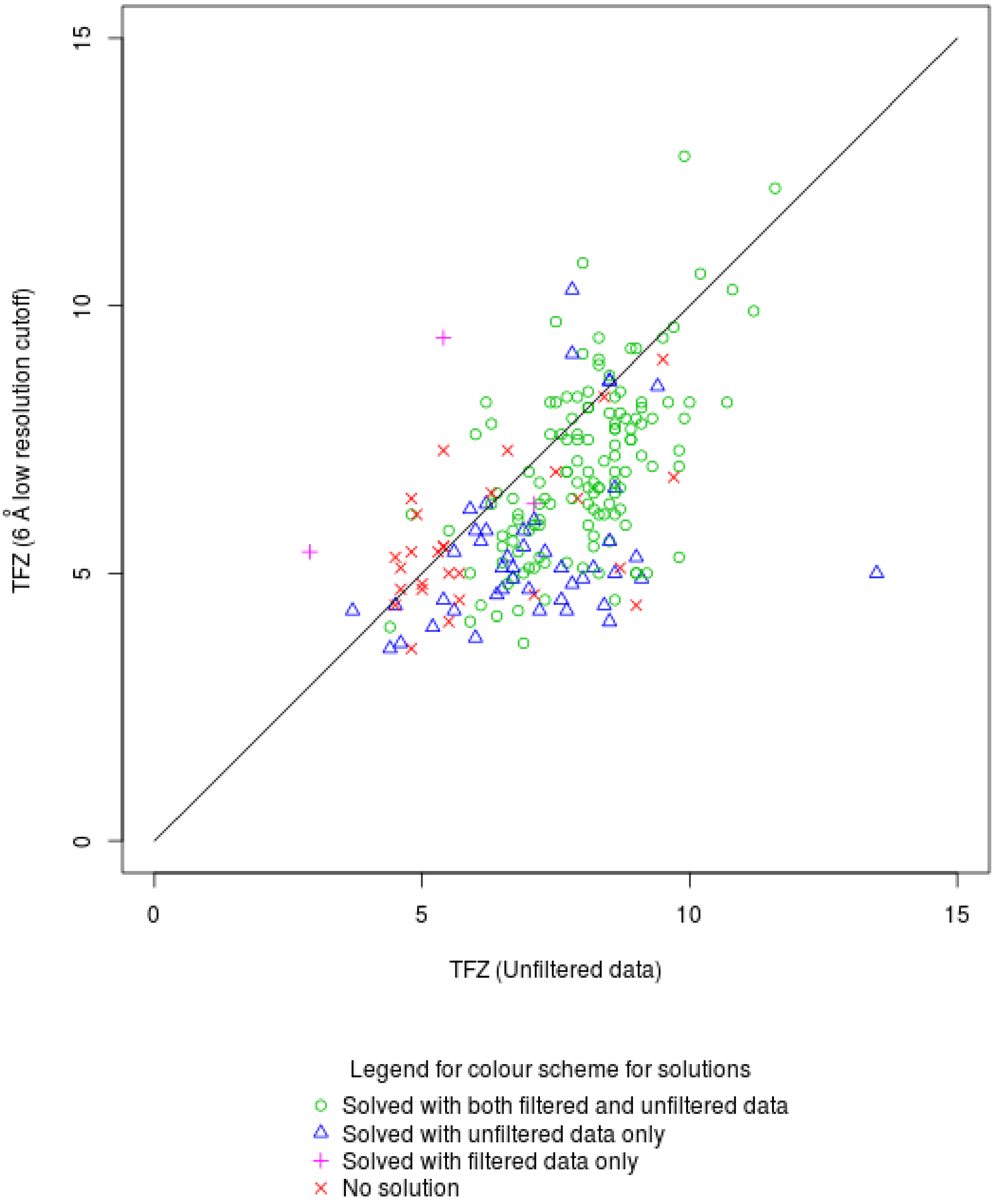


Figure 4-7: TFZ scores for 6 Å low resolution filtered data versus unfiltered data for Patterson-based methods

The data have been subjected to a more severe low resolution filter here, and the clustering of points is somewhat further below the line, indicating further reduction in the quality of the results. Again, a more significant number of test cases are not solved (blue triangles) with the filtered data. The plot was created in *R*.

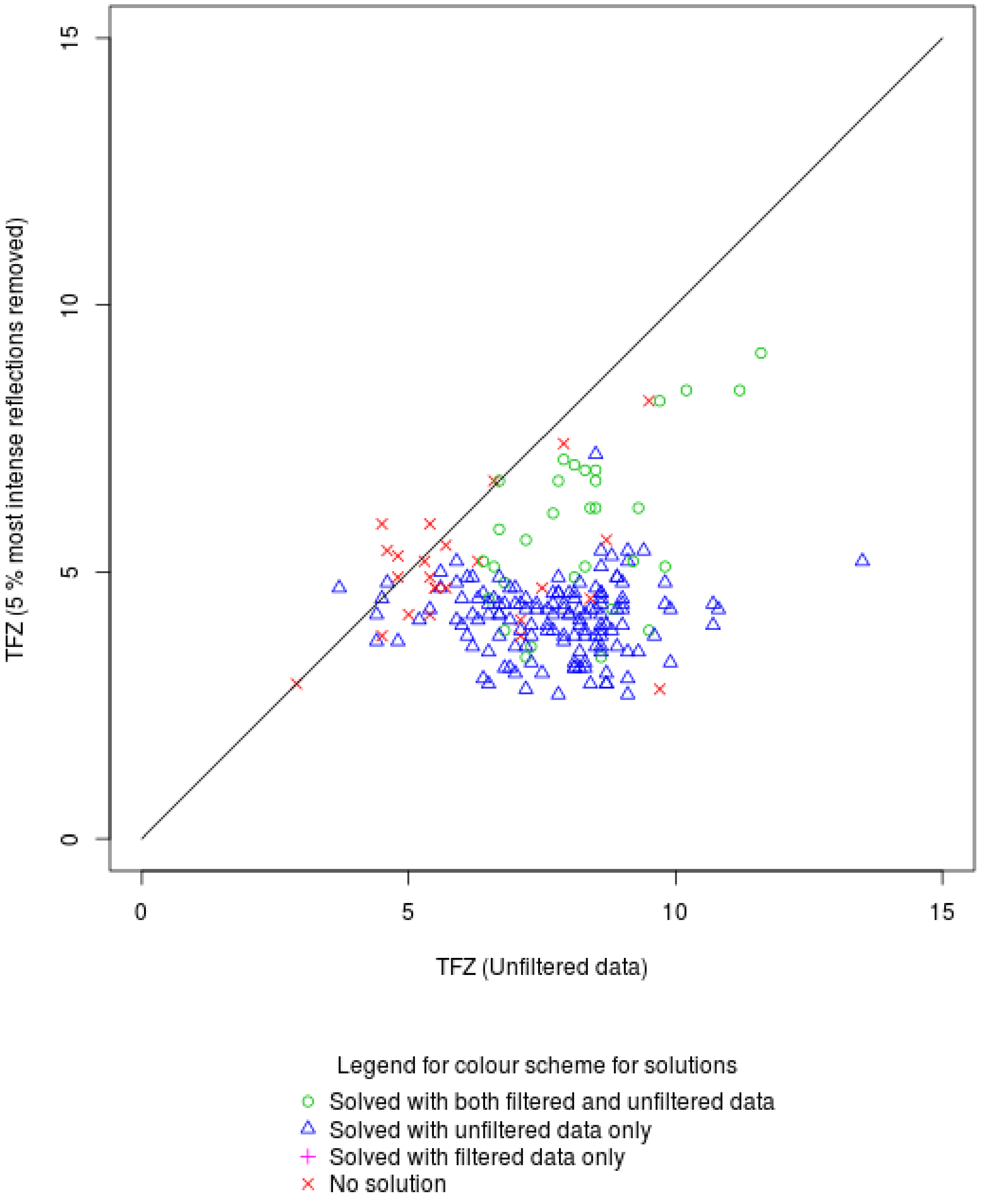


Figure 4-8: TFZ scores for 5 % most intense removed data versus unfiltered data for Patterson-based methods

The clustering of points is well below the line in this case, indicating that removal of the data associated with the most intense 5 % of the spots has greatly reduced the quality of the results when using Patterson-based methods. The majority of test cases are now not solved (blue triangles) with the filtered data. The plot was created in *R*.

	Original Data	5 % most intense reflections removed
Mean TFZ	7.6	4.5
Mean LLG	68.8	132
Mean size of solution list	15.5	75.4
Number of matched solutions	180	31

Table 4-6: Summary results when 5 % most intense reflections are removed versus original data for Patterson-based methods

4.6.4.3 Removing 5 % most intense reflections and applying low resolution cut-offs at the same time

As for the maximum likelihood investigations, experiments were run to investigate the effect of removing the 5 % most intense reflections and applying a low resolution cut-off simultaneously. The degradation in the ease of molecular replacement solution was also seen to be cumulative (Table 4-7), but the strong deleterious effect of the 5 % intensity filter (Section 4.6.4.2) masked the weak effect of the low resolution cut-off (Section 4.6.4.1).

Data Removed	Mean TFZ	Mean LLG	Mean size of solution list	Number of matched solutions
5 % + 15 Å	4.5	92.6	79.5	31
5 % + 8 Å	4.6	90.1	117.4	29
5 % + 6 Å	4.6	91.1	68.9	24

Table 4-7: Summary results when 5 % most intense reflections and low resolution reflections are removed for Patterson-based methods

The strongest measure supporting the statement above is the number of structures where the most significant result matched the known solution. The large initial loss of solutions caused by the removal of 5 % of the most intense reflections (180 to 31 solutions) was increased by smaller numbers as the more harsh resolution criteria were applied (29 for 8 Å and 24 for 6 Å). As has been seen previously, the 15 Å resolution cut-off had no effect.

The other measures do not give clear indications of the cumulative effect. The LLG, as discussed, is not helpful when data are excluded in this particular *phaser*[4] version, but showed small trends in the expected direction. The mean TFZ stays approximately the same, with a slight increase (0.1) for higher resolution limits. The mean number of solutions with the 5 % most intense reflections removed did increase as the resolution cut-off was increased until the 6 Å case. At this point, the number of cases that failed completely and did not produce any solutions skewed this number.

4.6.4.4 Comparison of the effects of removing the low resolution data and the most intense spots

The most appropriate comparison remains the 6 Å resolution limit and the 5 % most intense reflection removal, as this has approximately the same percentage of reflections remaining (94 % versus 95 %).

The mean number of solutions with the maximum likelihood score matching the known solution differed significantly for the two data eliminations. In the datasets with the selected 6 Å low resolution data cut off, the mean number was higher (136 cases), than in the datasets where 5 % most intense reflections were removed (31 cases). The mean TFZ in the data showed similar large differences: 6.4 in the low resolution case, and 4.5 in the case where the most intense resolutions were removed. Both of these indicate strongly that the removal of the 5 % most intense reflections has a much more significant effect than the removal of the low resolution data.

The mean size of the solution list also supported this contention. Here the mean size when removing the most intense reflections was 75.4, but when removing the low resolution data it was 59.8.

4.7 Conclusions

4.7.1 Comparison of different measures of success

The two most effective measures of success in finding solutions were the number of top solutions measuring the known solution and the TFZ score. These showed consistent results and permitted conclusions to be drawn from the various experiments performed. The first measure, the number of top solution matches, showed consistent trends, and sensitivity to small changes in the experimental conditions. This measure was also insensitive to complete failures to find solutions, whereas the TFZ and other scores, being averages, were affected when this occurred.

The average number of solutions proved to be only a rough guideline to the successful solutions. As *phaser*[4] encountered more difficulties in determining solutions, this number increased, but under extreme conditions, the variation in this number become large and obscures the trend.

The LLG was not particularly useful because of issues with the implementation in the version of *phaser*[4] that was available at the time of these experiments.

4.7.2 Validity of the test set

The test set used was very useful for the 2013 version of *phaser*[4], as it showed a full range of success and failure under the experimental conditions. However, the 2016 version has exhibited significantly improved performance, and reduced the sensitivity of the test sets. As *phaser*[4] is continuously improved, a new test set will be required for future work.

The 15 Å low resolution cut-off limit proved to be a good check of the method used to remove reflections. The close correlation of these results with the original data indicated that the method was not introducing artefacts.

When this data set was used to interpret the Patterson-based methods as implemented in *phaser*[4], the 5 % most intense cut-off filter reduced the reported successes so much that it made the interpretation of the additional impact of eliminating low resolution reflections more difficult. Future work could consider a different data set to investigate this combined effect more fully.

4.7.3 Low resolution vs. greatest intensity spots for maximum likelihood methods in phaser

Some important trends were evident in the results. Firstly, the removal of low resolution data did impact upon the success of molecular replacement trials. Very low resolution data (15 Å) was important in very few cases, but increasingly harsh resolution limits led to a gradual reduction in the number of solved solutions. The harsher resolution limits also led to a noticeable decrease in the average TFZ. They also caused an increase in the number of solutions reported, leading to potentially more candidates to search through if the correct answer was not known, with no guarantee that the correct solution was among them.

The removal of the 5 % most intense intensities had a very significant effect on the ability of *phaser*[4] to succeed in the molecular replacement test cases. A far greater effect was seen when compared to the low resolution limits on all measures that indicate the success rate. This is despite the number of reflections removed in the 6 Å case (94 % reflections remaining) being more than the number removed for the 5 % greatest intensity (95 % reflections remaining). The effect was apparent in various measures of the success of molecular replacement test cases – there were fewer correct solutions as the first solutions, lower mean TFZs (Figure 4-9 and Figure 4-10) and a longer list of potential solutions. This therefore shows that removing the greatest intensity spots has a greater effect than removing the low resolution data.

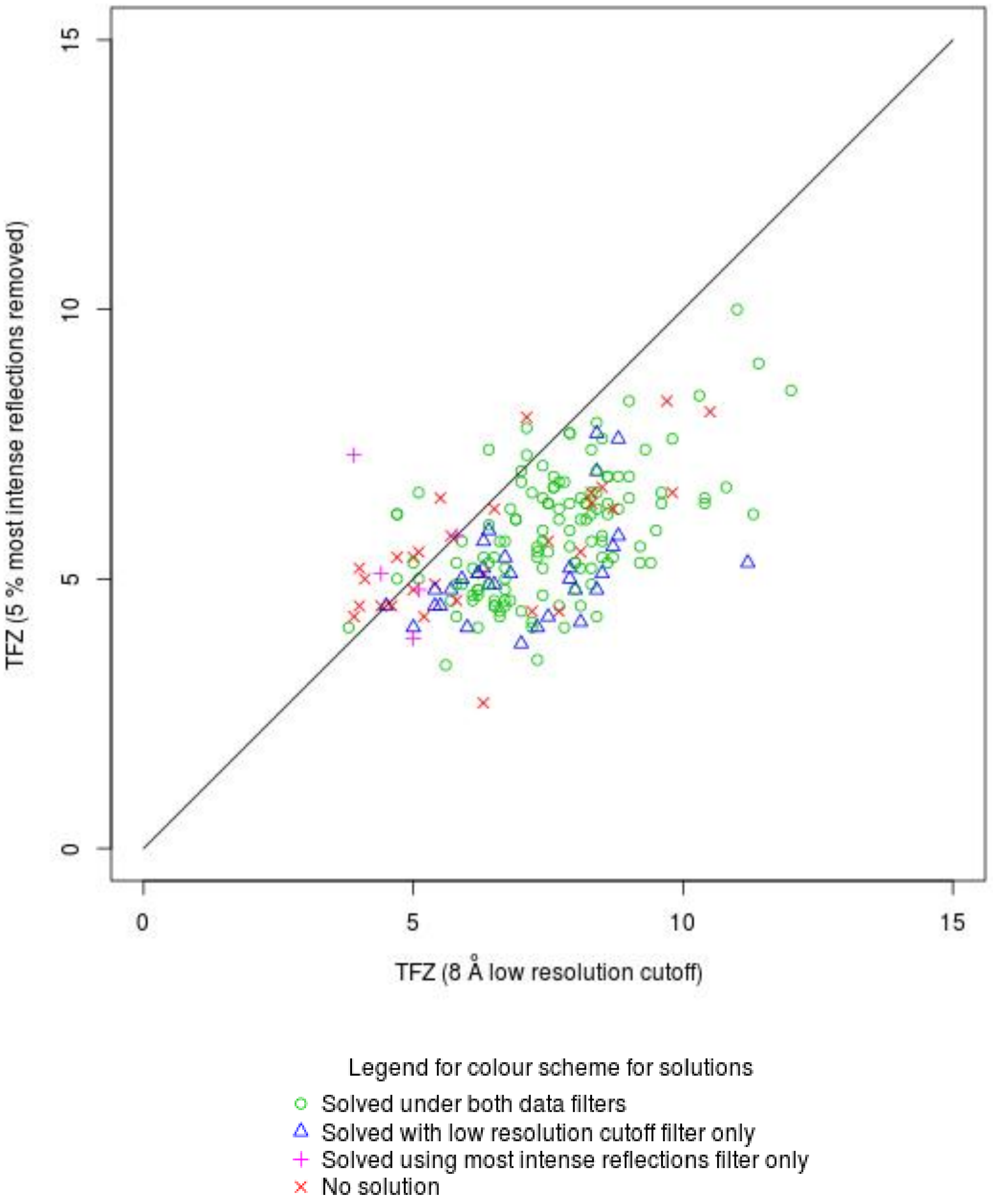


Figure 4-9: TFZ scores for 5 % most intense removed data versus 8 Å low resolution filtered data for maximum likelihood methods

The clustering of points below the line indicates that removal of the low resolution data have significantly less effect than removal of the most intense 5 % of the spots. Filtering the low resolution data removes fewer spots, so this result indicates a possible trend. A significant number of test cases are solved (blue triangles) only with the low resolution cut-off, but not with the high intensity data removal. The plot was created in *R*.

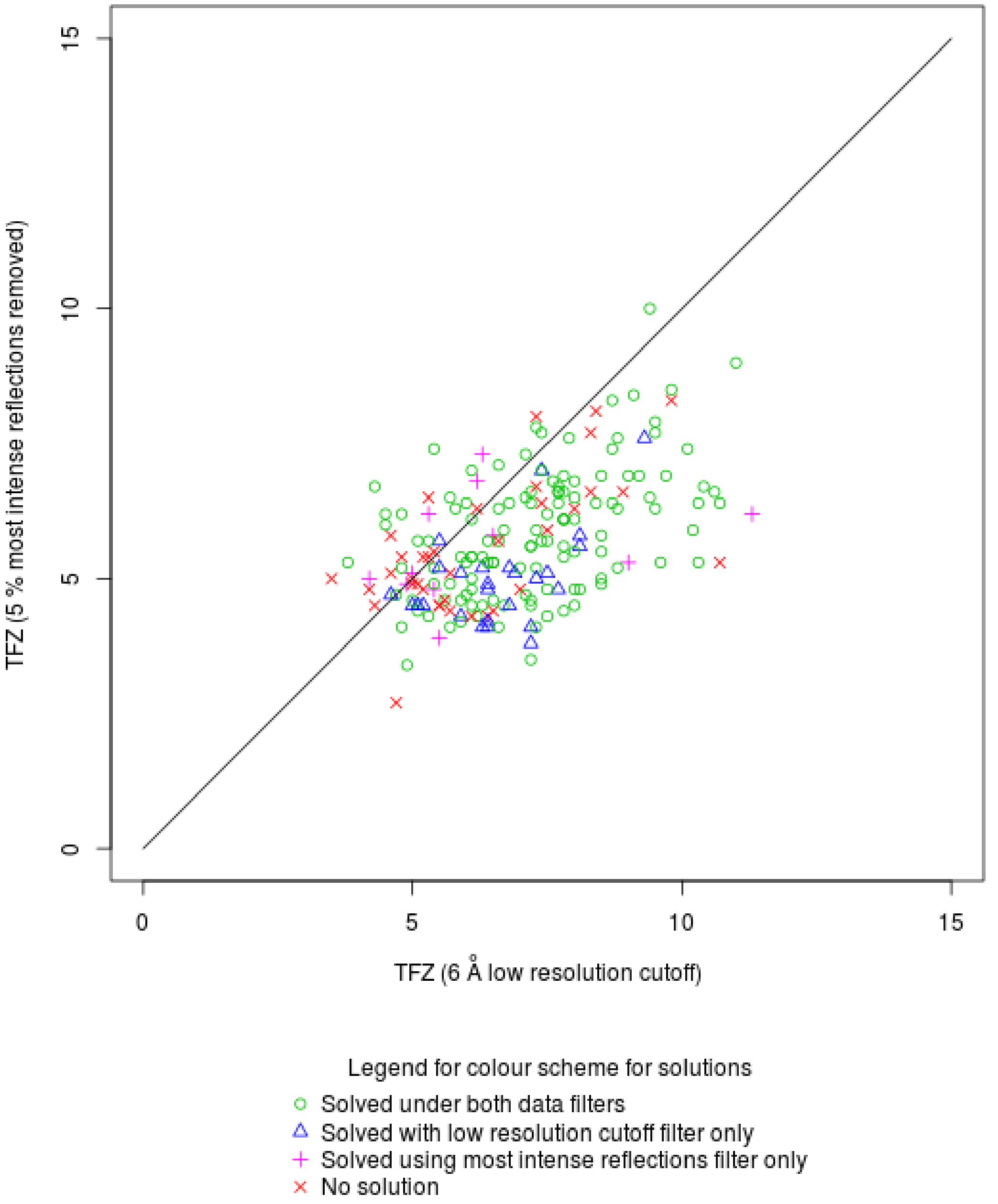


Figure 4-10: TFZ scores for 5 % most intense removed data versus 6 Å low resolution filtered data for maximum likelihood methods

The clustering of points clearly below the line indicates that removal of the low resolution data have significantly less effect than removal of the most intense 5 % of the spots. Filtering the low resolution data removes an equivalent number of spots in this case, so this result shows that the low resolution data are less significant than the most intense spots in finding solutions. A number of test cases are solved (blue triangles) only with the low resolution cut-off, but not with the high intensity data removal. The plot was created in *R*.

The confusion that has arisen over the past few years[299], that the lowest resolution data are uniquely important for molecular replacement and must be collected, has been shown to be less valid. However, as the low resolution data tend to have the greatest intensity spots, this could suggest a source of this interpretation.

The results obtained also indicate that the size of the beam-stop is of less importance, though it remains a significant factor. Having a large beam-stop will mean that the very low resolution data cannot be recorded, which has been shown to lead to a lower success rate. However, there are other factors which will affect the success significantly more, such as recording the 5 % most intense reflections accurately.

To summarize, it is very important to collect the most intense reflections accurately and that overloading the detector at these points is detrimental to solving the structure by molecular replacement. This supports the results found by Gideon Davies[3]. Modern instrumentation, such as the Pilatus[222,297], has been developed and it is now becoming notably harder to overload the detector. This is a significant improvement on the older detectors and should therefore lead to more accurate structure determination.

Finally, the effects of the degradation of the data by missing out the lowest resolution reflections and missing out the 5 % greatest intensity were shown to be cumulative. Deleting both sets of reflections led to an even lower success rate in the molecular replacement test cases than deleting either set of reflections separately.

4.7.4 Low resolution vs. greatest intensity reflections for Patterson-based methods in phaser

Having seen a difference in the effect of elimination of the most intense reflections and the lowest resolution data in diffraction patterns when using maximum likelihood techniques, the question arose as to whether low resolution data were uniquely important when using Patterson-based methods, or whether a similar trend would be seen. This assessment was restricted to the implementation of such methods in *phaser*[4], and was not an attempt to measure the relative success of the two molecular replacement methodologies in general.

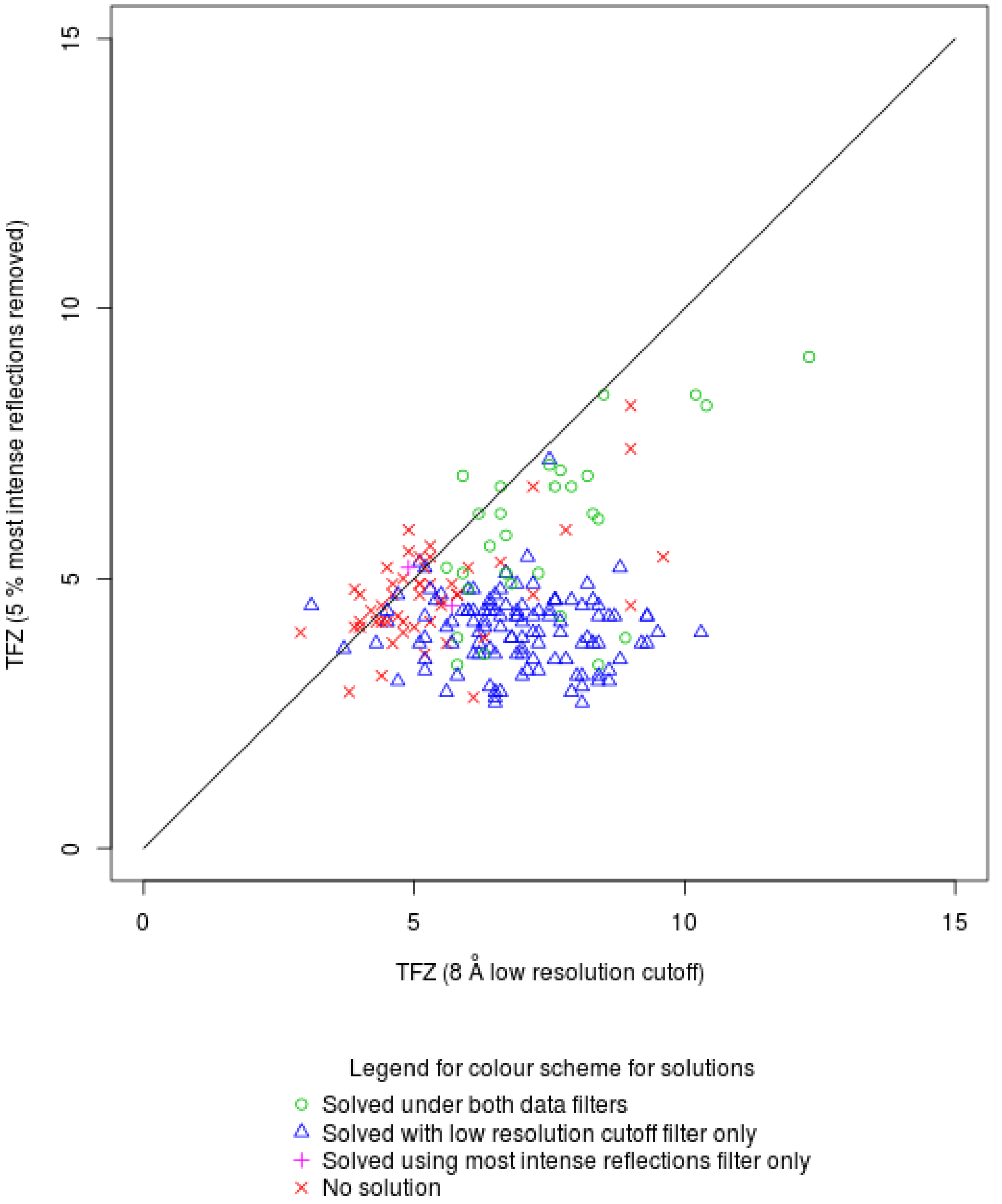


Figure 4-11: TFZ scores for 5 % most intense removed data versus 8 Å low resolution filtered data for Patterson-based methods

The stronger clustering of points below the line indicates that removal of the low resolution data have significantly less effect than removal of the most intense 5 % of the spots for Patterson-based methods as implemented in *phaser*. Filtering the low resolution data to 8 Å removes fewer spots, so this result indicates a possible trend. A majority of test cases are solved (blue triangles) only with the low resolution cut-off, but not with the high intensity data removal. The plot was created in *R*.

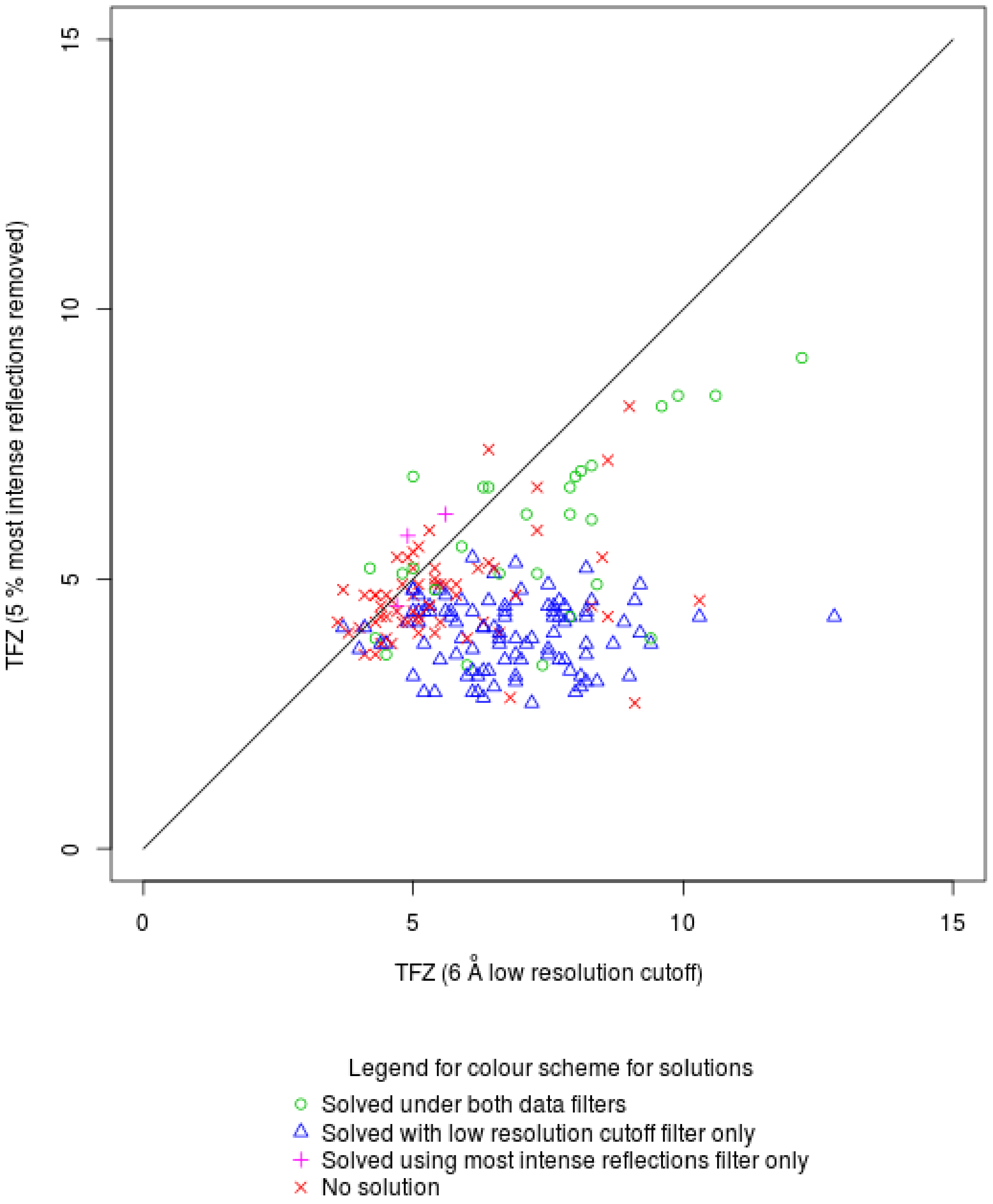


Figure 4-12: TFZ scores for 5 % most intense removed data versus 6 Å low resolution filtered data for maximum likelihood methods

The similar strong clustering of points below the line indicates that removal of the low resolution data have significantly less effect than removal of the most intense 5 % of the spots for Patterson-based methods as implemented in *phaser*. Filtering the low resolution data to 6 Å removes a very similar percentage of the spots, so this result shows that the low resolution data are less significant than the most intense spots in finding solutions. A majority of test cases are solved (blue triangles) only with the low resolution data removed. The plot was created in *R*.

Patterson methods were developed earlier than maximum likelihood methods, and are still in routine use successfully. Many of the earlier results were obtained using these methods. Accordingly, understanding the relative importance of the low resolution data and the greatest intensity information in assessing results using this method will help in experimental design and interpretation of results.

This work shows that, in the measures of success that have been used, the same trends can be seen as for maximum likelihood methods (Figure 4-11 and Figure 4-12). The low resolution data have some importance, and does affect the success in obtaining results, but removal of the most intense reflections, regardless of resolution, has a more significant effect. This was seen even though removing the low resolution data eliminated more reflections in the pattern.

For these as for the maximum likelihood methods, the focus should be on collecting the most intense reflections accurately, regardless of position in the pattern, though many will be at the lower resolutions. Focus on collecting low resolution data at the expense of higher resolution data that includes some high intensity reflections may be a poorer strategy when determining experimental compromises.

4.7.5 Maximum likelihood methods vs. Patterson-based methods

This section compares the performance of the maximum likelihood and Patterson-based methods, as implemented in *phaser*[4], under these test situations. This comparison is shown for each individual test case graphically (Figure 4-13 to Figure 4-18), and combined results from all test cases are shown in a tabular form (Table 4-8) and through bar charts (Figure 4-19 to Figure 4-22).

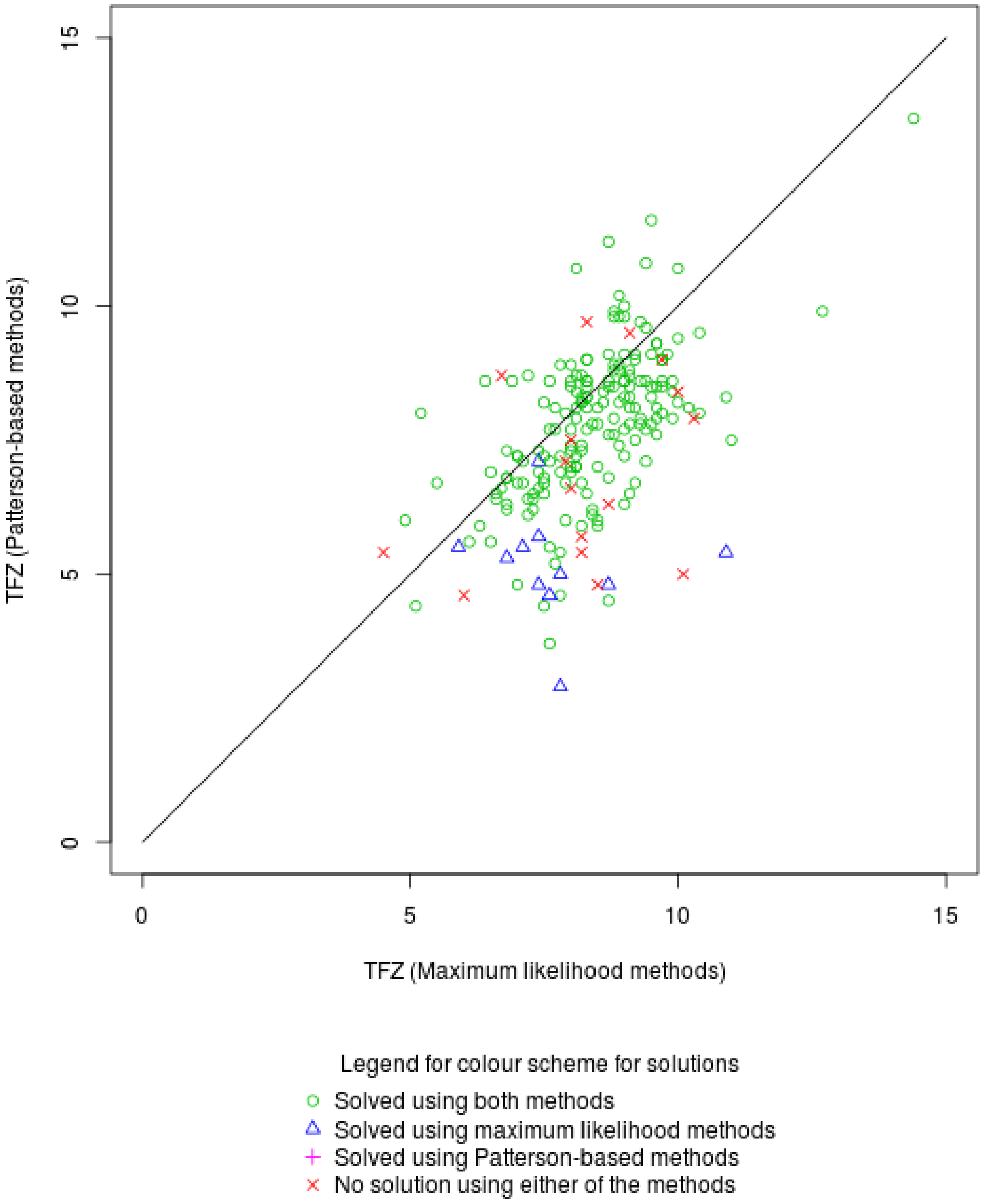
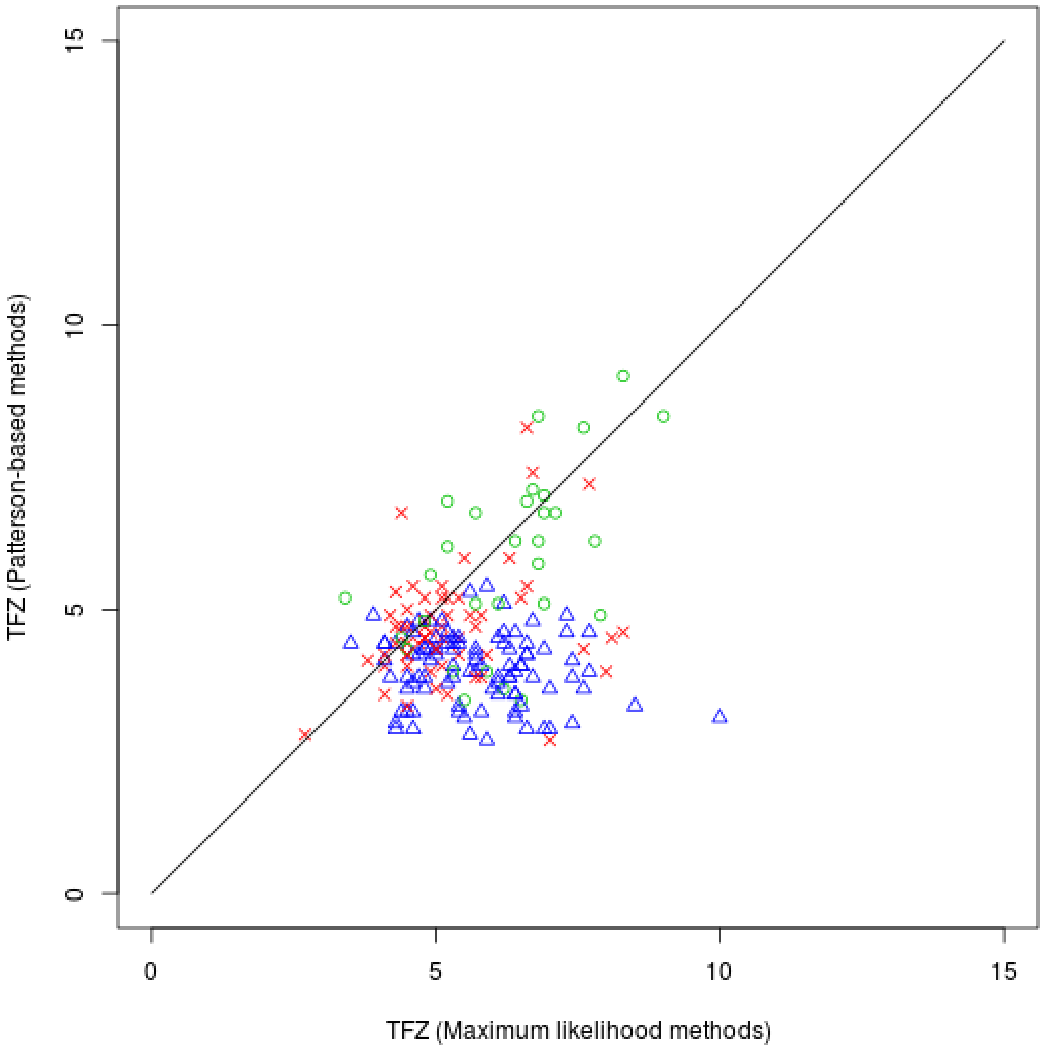


Figure 4-13: TFZ scores for Patterson-based methods versus maximum likelihood methods for the original test data

Most original test cases were solved by both methods (green circles). The centre of gravity of the points is below the line, indicating that the maximum likelihood methods yield higher TFZ scores than Patterson-based methods in the *phaser* implementation. A small number of test cases were only solved by the maximum likelihoods methods, and none were only solved by the Patterson-based methods. The plot was created in *R*.



Legend for colour scheme for solutions

- ○ Solved using both methods
- △ Solved using maximum likelihood methods
- + Solved using Patterson-based methods
- × No solution using either of the methods

Figure 4-14: TFZ scores for Patterson-based methods versus maximum likelihood methods with 5 % of the most intense reflections removed

When the most intense 5 % spots are removed, we see that the maximum likelihood methods are more successful in the *phaser* implementation. A significant number are also only solved using these methods (blue triangles), and a similarly significant number are not solved by either method (red diagonal crosses). This indicates that this removal of the most intense reflections has a significant effect on the determination of structures. The plot was created in *R*.

.

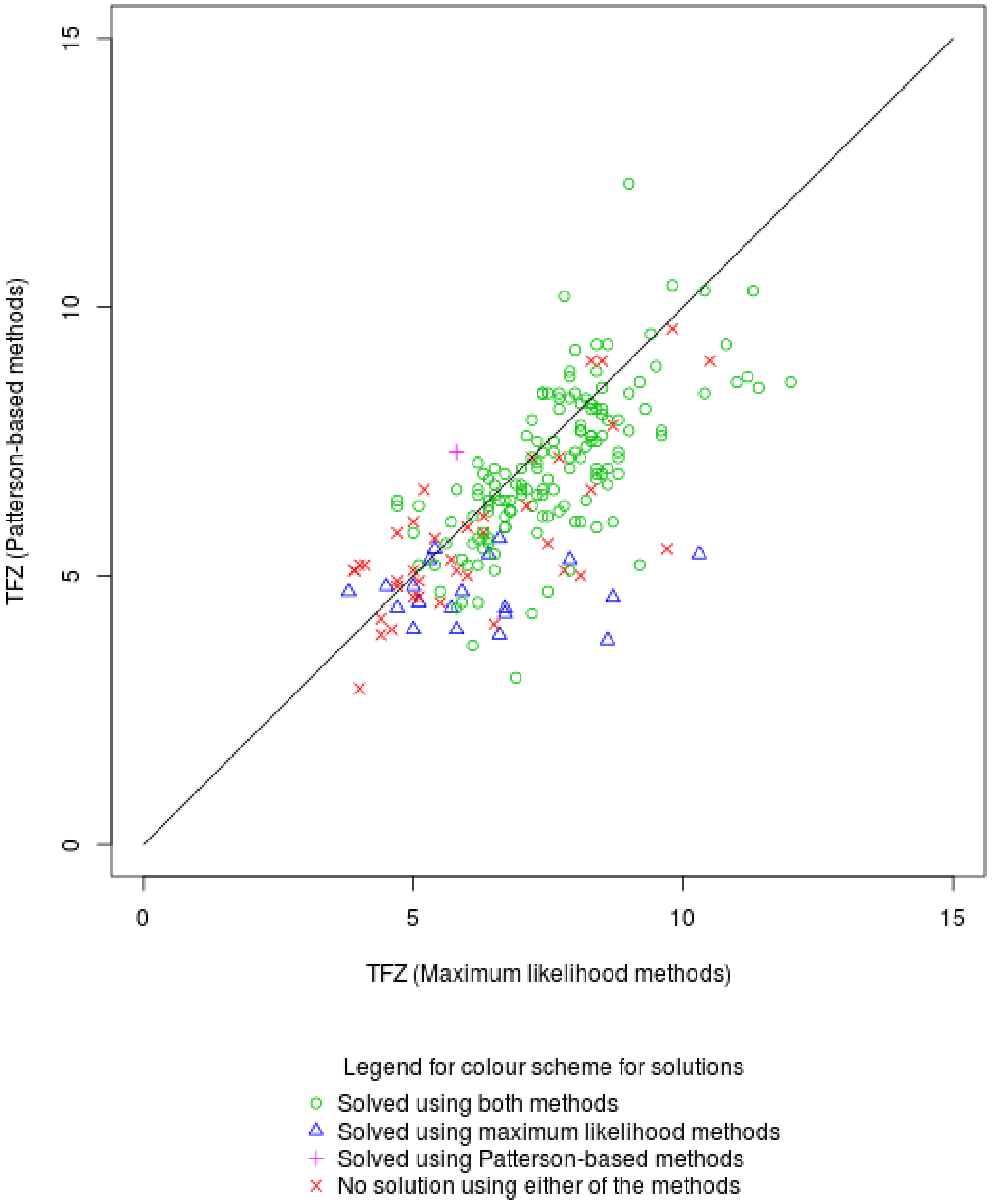


Figure 4-15: TFZ scores for Patterson-based methods versus maximum likelihood methods with an 8 Å resolution cut-off limit on the reflections

When an 8 Å resolution cut-off limit is applied to the data, we see a similar result to that observed with the original data, though the spread is somewhat larger. This indicates that this reduction of the data have not had a large impact, and that solutions are generally being found by both methods. The plot was created in *R*.

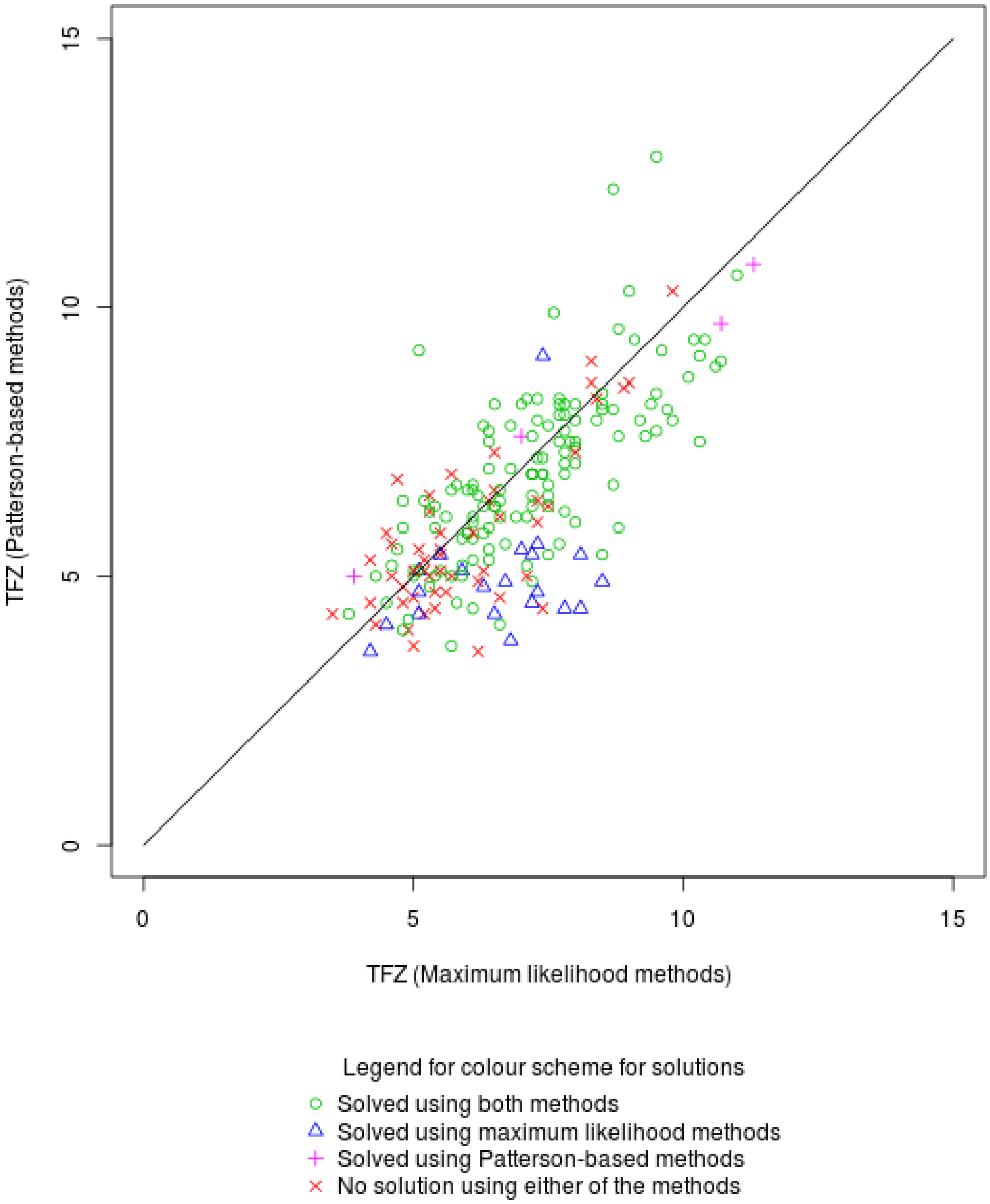


Figure 4-16: TFZ scores for Patterson-based methods versus maximum likelihood methods with a 6 Å resolution cut-off limit on the reflections

A similar spread can be seen in these results, but a larger number of red diagonal crosses are beginning to emerge. However, the general distribution is similar and few conclusions can be drawn. The plot was created in *R*.

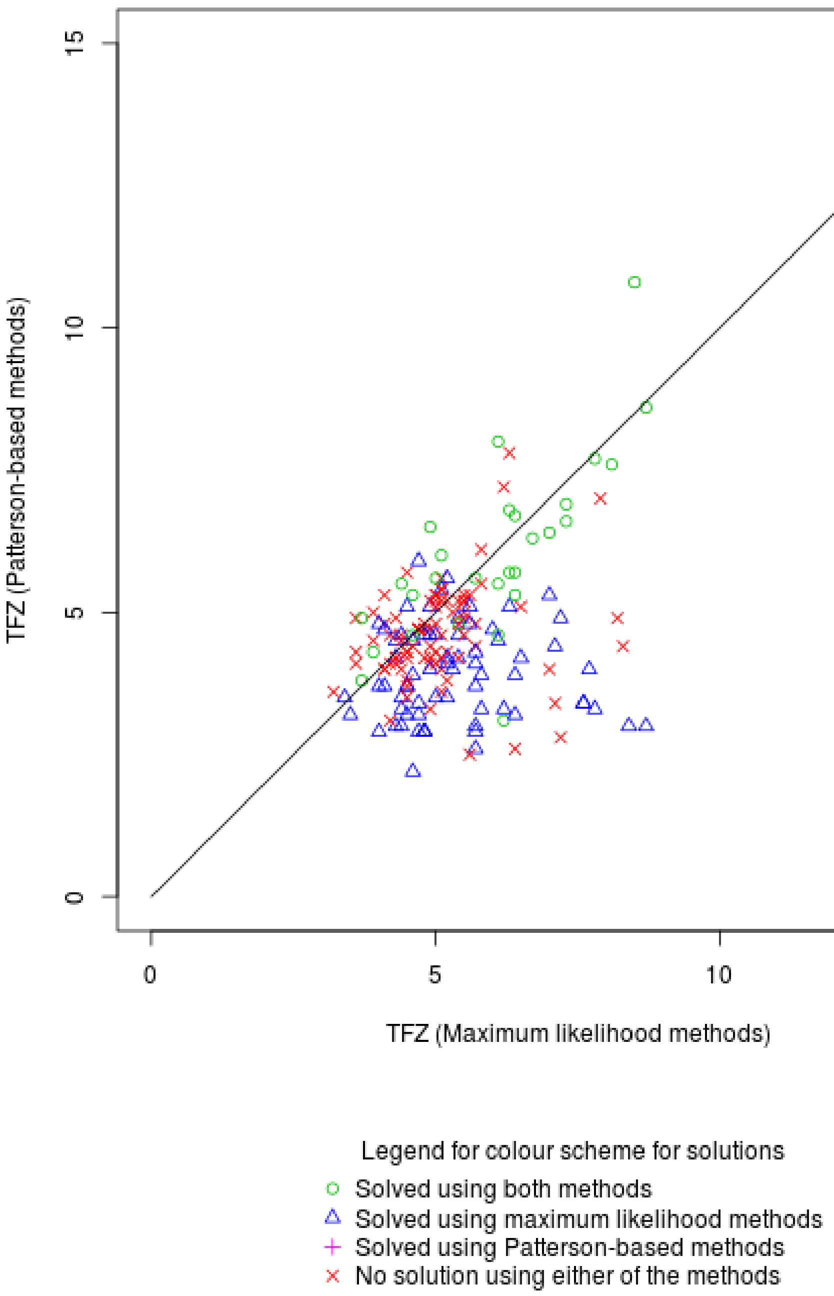


Figure 4-17: TFZ scores for Patterson-based methods versus maximum likelihood methods with an 8 Å resolution cut-off limit on the reflections and with 5 % of the most intense reflections removed

The cumulative effect of the dominant impact of removing the most intense spots and the less severe resolution limit can be seen on this graph. The number of green circles has reduced, and the grouping slightly favours the maximum likelihood methods, as expected from earlier results. The plot was created in *R*.

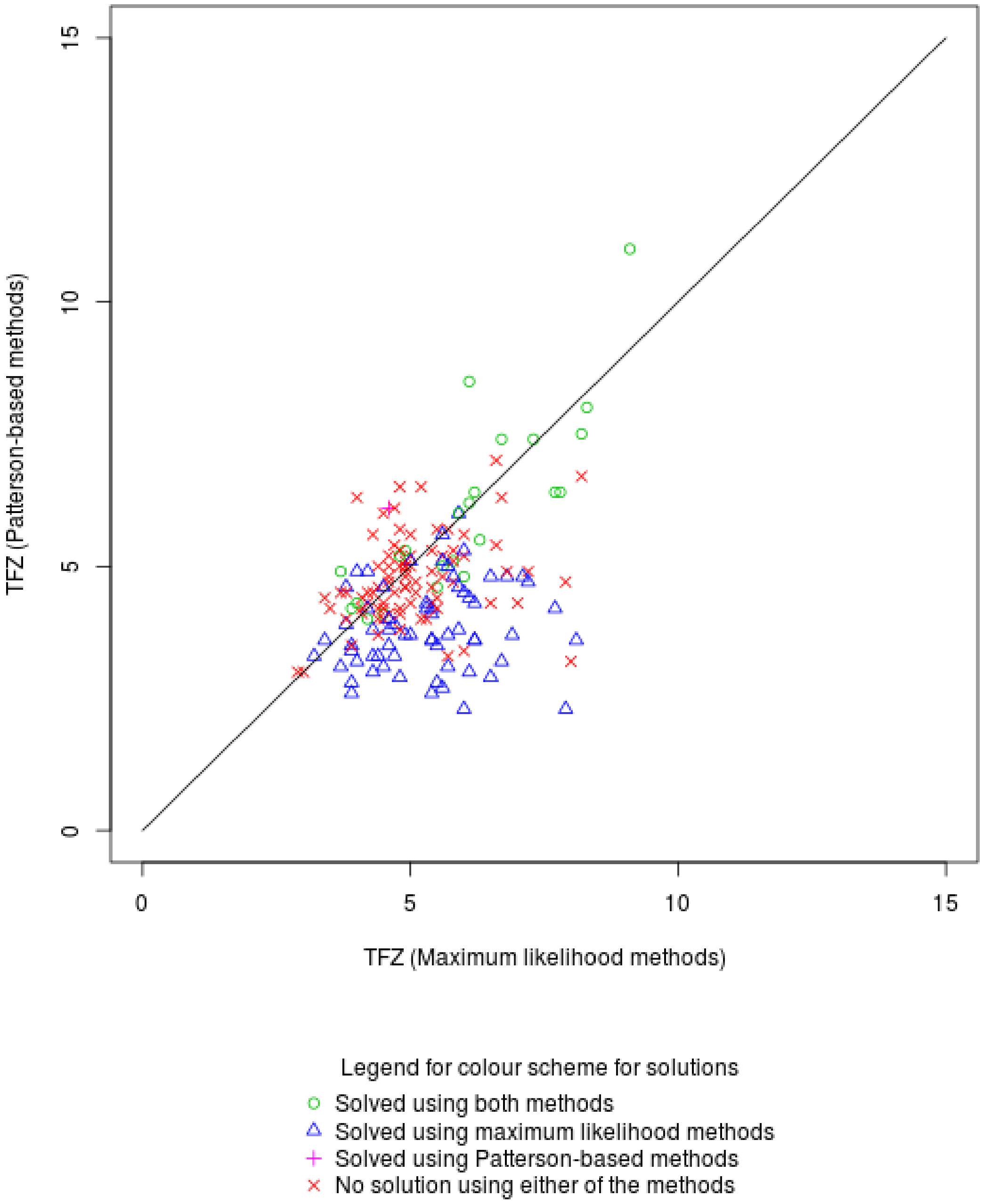


Figure 4-18: TFZ scores for Patterson-based methods versus maximum likelihood methods with a 6 Å resolution cut-off limit on the reflections and with 5 % of the most intense reflections removed

The cumulative effect of the dominant impact of removing the most intense spots and the more severe resolution limit can be seen on this graph. The number of green circles has further reduced, and the grouping slightly favours the maximum likelihood methods, as expected from earlier results. The plot was created in *R*.

(a)

Mean (LLG)	Likelihood	Likelihood (with 5 % most intense removed)	Patterson	Patterson (with 5 % most intense removed)
No cut off	73.7	156.8	68.8	132.0
15 Å	87.8	115.8	81.7	92.6
8 Å	65.3	105.8	59.1	90.1
6 Å	69.3	108.3	66.3	91.1

(b)

Mean (TFZ)	Likelihood	Likelihood (with 5 % most intense removed)	Patterson	Patterson (with 5 % most intense removed)
No cut off	8.4	5.7	7.6	4.5
15 Å	8.5	5.7	7.5	4.5
8 Å	7.2	5.3	6.6	4.6
6 Å	6.9	5.3	6.4	4.6

(c)

Mean size of solution list	Likelihood	Likelihood (with 5 % most intense removed)	Patterson	Patterson (with 5 % most intense removed)
No cut off	3.3	10.2	15.5	75.4
15 Å	3.3	9.9	19.8	79.5
8 Å	25.3	27.6	36.8	117.4
6 Å	33.8	34.6	59.8	68.9

(d)

Number of solved structures	Likelihood	Likelihood (with 5 % most intense removed)	Patterson	Patterson (with 5 % most intense removed)
No cut off	191	135	180	31
15 Å	189	133	174	31
8 Å	168	103	150	29
6 Å	156	95	136	24

Table 4-8: Comparison between maximum likelihood methods and Patterson-based methods at different low resolution cut-offs with or without the 5 % most intense reflections

This shows the mean LLG (a), the mean TFZs (b), the mean number of solutions reported (c) and the number of structures where the first molecular replacement solution solved was the correct structure (d).

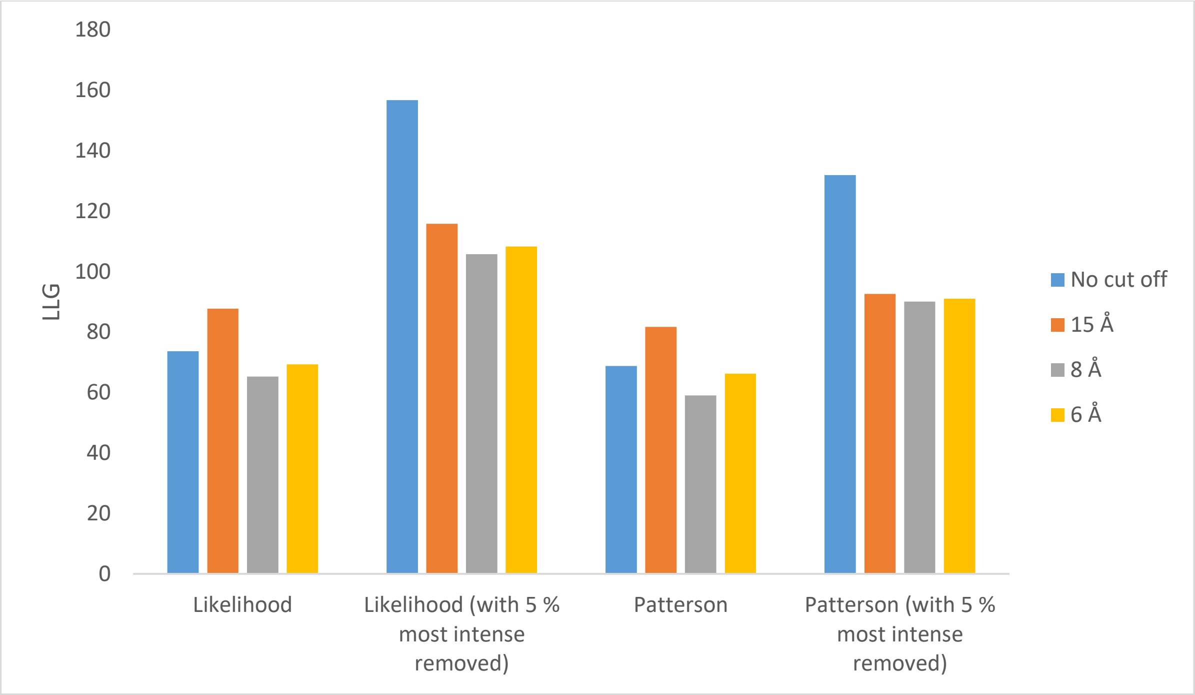


Figure 4-19: LLG scores for the two analysis methods, showing the effect of different exclusions of data

The bar graph was created using *Microsoft Excel*.

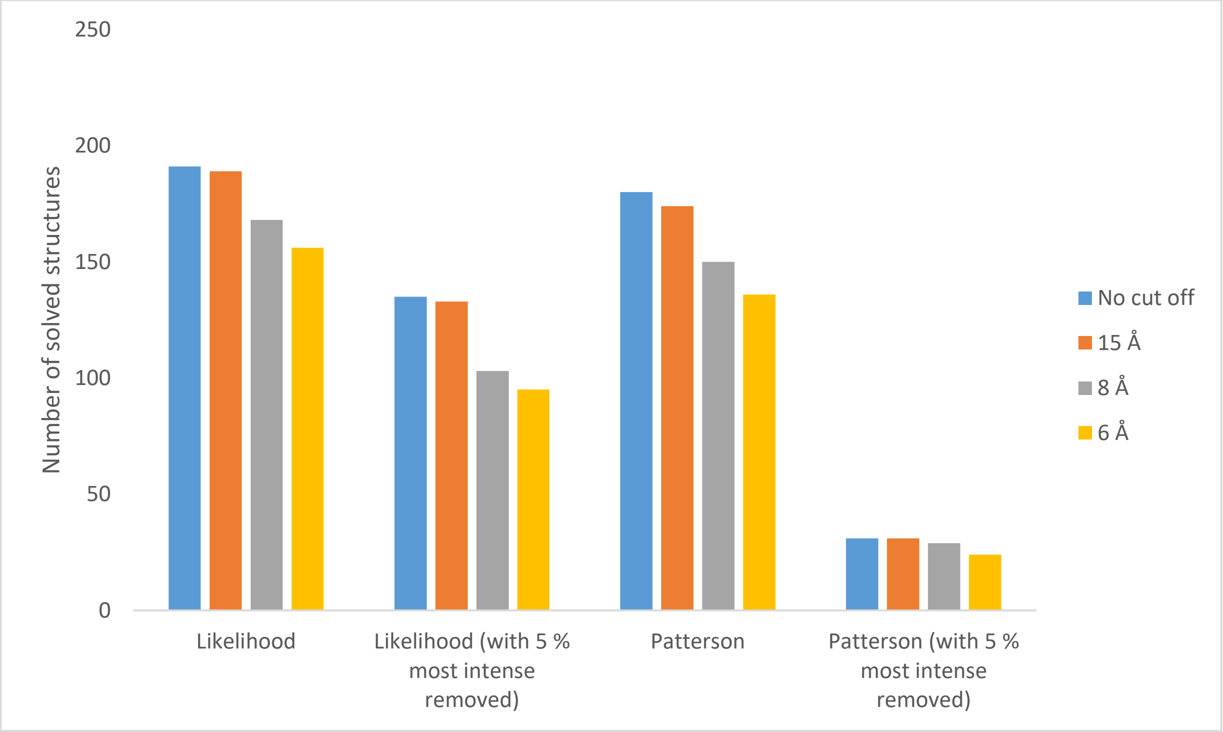


Figure 4-20: Number of correct first solutions that match the known solutions for the two analysis methods, showing the effect of different exclusions of data

The bar graph was created using *Microsoft Excel*.

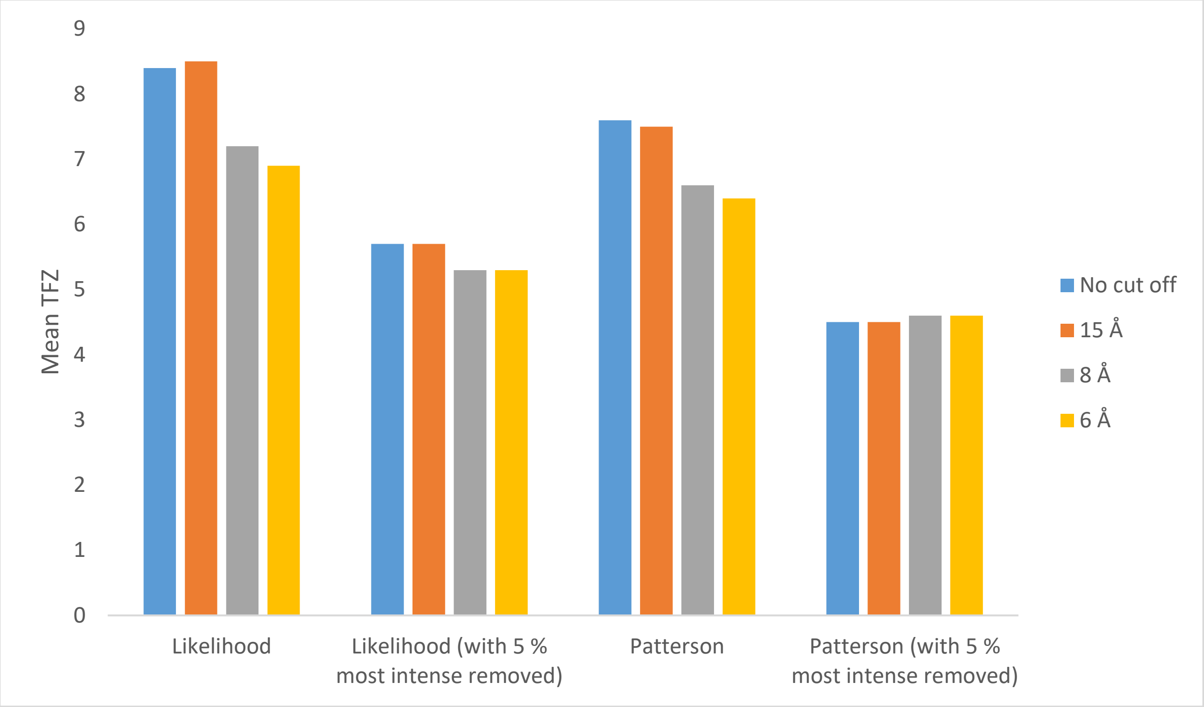


Figure 4-21: TFZ scores for the two analysis methods, showing the effect of different exclusions of data

The bar graph was created using *Microsoft Excel*.

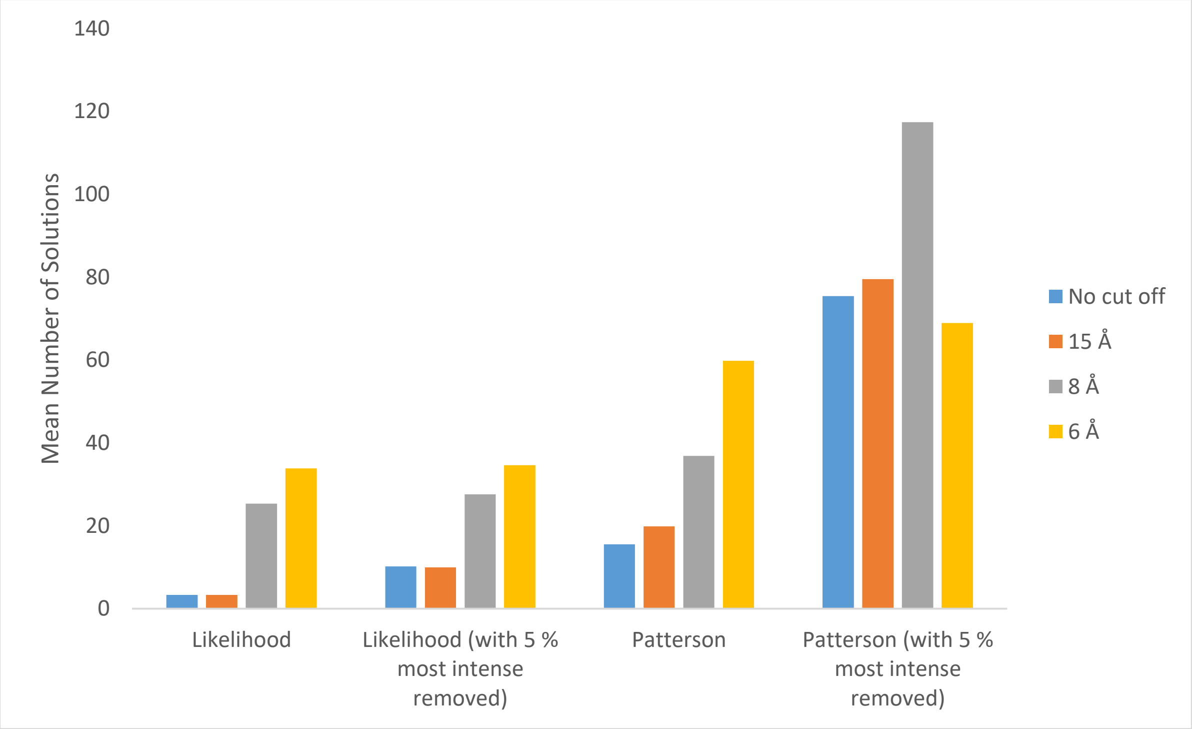


Figure 4-22: Mean number of solutions for the two analysis methods, showing the effect of different exclusions of data

The bar graph was created using *Microsoft Excel*.

In every condition tested the Maximum likelihood methods produced more successful results than the Patterson-based methods in *phaser*[4], no matter how many reflections were removed, or whether they were removed from the low resolution shells or the greatest intensity spots. This was clearly shown in all the measures recorded. Furthermore, maximum likelihood methods were better at filtering out incorrect solutions than the Patterson-based methods, as error handling is particularly well implemented for the former in *phaser*[4]. This effect was particularly seen in the trends in the mean number of potential solutions.

LLG has not proved to be a reliable measure in this implementation in *phaser*[4], so the results need to be treated with caution, and in this case do not provide much evidence. When comparing identical data exclusions, the LLG in the maximum likelihood methods is always better than in the Patterson-based methods, suggesting that the former is more successful.

The TFZ scores show a difference between the implementations of the two methods. As can be seen below, there were far more TFZ scores below the $y = x$ diagonal (Figure 4-13), which demonstrate graphically that the TFZ scores were higher with the maximum likelihood methods than the Patterson-based methods. This indicates that the former is a more successful method in *phaser*[4].

Other indicators also show differences in results for the two methods in *phaser*[4]. The number of solutions with the highest likelihood score matching the known solutions was higher with maximum likelihood methods, and it is particularly noticeable that the average size of the potential solutions list was much lower with maximum likelihood methods.

Turning now to the results of the data exclusion experiments, both methods show similar patterns in the degradation they experience under the two test conditions. On removing the low resolution data, some degradation of results obtained for the different measures can be seen, but to a lesser extent than when the 5 % most intense reflections are not present.

Looking in more detail at the results obtained at different low resolution cut-off limits, the mean TFZ, whilst remaining approximately the same in both Patterson-based methods and maximum likelihood methods for 15 Å resolution cut-off, decreases as more reflections are discarded. This confirms that losing the low resolution data does indeed make it harder to get a molecular replacement solution. The mean number of solutions also demonstrates this well. There is little increase as a 15 Å resolution cut-off is applied, but when more reflections are removed at more stringent resolution cut-offs the average size of solution lists increases greatly. Moving on to the number of correct first solutions that match the known solutions, a similar trend is again seen.

However, in these results the Patterson-based methods are more affected by the omission of low resolution data than maximum likelihood methods. Proportionally, when comparing the number of solutions at the 6 Å resolution cut-off limit with the number when no reflections are left out, Patterson-based methods have 76 % of structures originally solved still being solved, whereas maximum likelihood methods have 82 % solved. Graphically, this is also seen with the change in TFZ scores (Figure 4-15 and Figure 4-16).

The contrast between maximum likelihood methods and Patterson-based methods was particularly noticeable in their response to the removal of different sets of reflections. Patterson-based methods are far more susceptible to deficiencies in the data than maximum likelihood methods in *phaser*[4], as was seen when different sets of reflections were excluded.

This difference between the methods is even more apparent when the 5 % most intense reflections are removed. Patterson-based methods only solved 17 % of the structures that they originally solved, unlike maximum likelihood methods that still solved 71 % of the molecular replacement cases. Patterson-based methods were also able to discriminate the correct solution far less well than maximum likelihood methods, as shown in the larger mean number of solutions reported. The change in TFZ scores along with the cases in which solutions are found is also particularly noticeable (Figure 4-14).

The results from the maximum likelihood methods and Patterson-based methods when both the 5 % most intense reflections are removed and a low resolution cut-off applied allows a number of conclusions to be drawn. Firstly, it can be seen graphically (Figure 4-17 and Figure 4-18) that most of the points plotting the TFZ score of the two different methods lie below the y = x line, suggesting that the TFZ score is predominantly higher in maximum likelihood methods than in Patterson-based methods in this implementation.

The maximum likelihood methods consistently find more solutions than the Patterson-based methods in *phaser*[4] for the test database. 50 % of molecular replacement test cases that originally had a successful solution still solve with maximum likelihood methods when a 6 Å resolution cut-off and the 5 % most intense reflections are removed, as compared to 13 % with Patterson-based methods. This implies that Patterson-based methods are more affected by problems with the data, but showing similar trends to likelihood based methods.

4.7.6 Normalisation

The initial experiments using *phaser*[4] 2013 were successful when the data elimination was performed after normalisation, but reported errors and returned no solutions when performed before normalisation. The handling of errors or conditions where the intensity distribution was skewed (as is the case where the most intense reflections are removed) were poor in this version. However, when *phaser*[4] 2016 was used, data elimination before normalisation was successful, permitting this phase of experimentation to be completed. This indicates significant improvement in the performance of *phaser*[4] in the later versions.

The problems experienced with the earlier version of *phaser*[4] are further confirmation that the most intense spots are more significant in determining solutions in molecular replacement.

4.7.7 Summary of conclusions

The results indicate that low resolution data are not of particular importance to molecular replacement and obtaining the 5 % most intense reflections across the diffraction pattern is relatively more important, both in the initial experimental phase of gathering results, and in subsequent computational analysis.

These experiments also indicate that, whilst both maximum likelihood and Patterson-based methods are far more susceptible to issues with the loss of 5 % most intense reflections than to the very low resolution data being lost, the former, as implemented in *phaser*[4], are better able to cope with this loss.

Chapter 5 Conclusions and Perspectives

5.1 Conclusions

This thesis describes three different but related pieces of research in the field of structural determination through X-ray diffraction. Methods development in structure determination has been taken forward through this work in different ways. Firstly, the structure determination and understanding of cytokine receptor-like factor 3 (CRLF3) has helped a scientist working in the medical field to characterise this protein more fully, and suggestions have been made to assist in the interpretation of the importance of mutations in understanding disease. Secondly, the development of the *NaCelleS* program will assist scientists working with difficult datasets in the determination of molecular structures by offering alternative approaches to the discovery of non-crystallographic symmetry (NCS) in the data. Thirdly, the relative importance of low resolution data and high intensity reflections has been investigated leading to the conclusion that the latter are more significant. This will aid the design of experiments in this field. More detailed conclusions are found at the end of each chapter, but the key findings are summarised here.

There were many successful outcomes in the CRLF3 project. Firstly, the plasmid construct design allowed high expression levels and purification of part of the protein. This led to crystallisation of residues 174-442 of the protein of interest. Experimental phasing of mercury-containing derivative crystals of this construct then permitted a successful solution of the composition of this part of the protein. This then provided a basis to allow rationalisation of some of the mutations that can lead to blood clotting disorders in certain individuals. This work also permitted a fuller understanding of the procedures and techniques used in structural determination.

The *NaCelleS* program was developed with the aim of improving and extending the capabilities of *phenix.find_ncs_from_density*[2]. A novel approach to finding NCS through a two-sphere search was developed. This allows the electron density around the site of an anomalous scatterer to be compared with that of other anomalous scatterer sites and determines whether they are NCS copies of each other. A significant advantage of this approach is that, unlike geometry methods which work well when there are three or more sites per copy of the protein, this new method works well when there are only one or two sites. This has been shown to be successful, allowing identification of clear distinctions between sites that are linked and are not linked by NCS. A new one-sphere approach with *NaCelleS* had mixed success. Unlike the algorithm used in *phenix.find_ncs_from_density*[2], it uses all available data and performs a rotation search first to identify the most likely rotations. Comparative outcomes were better in some cases, but worse in others. When the program was unable to identify a full set of NCS operators, a further innovation was developed in iterating the *NaCelleS* protocol, which was shown to improve determination of these operators in some instances where other methods had failed.

The importance of low resolution and high intensity data in molecular replacement was investigated. A lack of low resolution data was examined to determine if this was uniquely important, and a comparison was made with the loss of an equivalent proportion of the greatest intensity reflections. Using several measures, it was found that low resolution data were not uniquely important and that it is more important to accurately collect the most intense reflections. This has addressed an assumption that has been held for some time.

A brief exploration was also made into the difference between maximum likelihood methods and Patterson-based methods, as implemented in *phaser*[4], and the effect of removing different reflections on the solution of molecular replacement test cases. Both methods exhibited similar characteristics on exclusion of specific data from the dataset, confirming the conclusions noted in the last paragraph. It was found that maximum likelihood methods coped better than Patterson-based methods in this implementation. Maximum likelihood methods appear to be better at accounting for errors in the data and so can deal with data deformities better, and *phaser*[4] appears to take full advantage of this.

5.2 **Future Directions**

Future direction for work in the three linked areas of research are considered in the next sections in turn. Work on CRLF3 concentrates on extending the determination of the structure of the protein, and in improving the understanding of the role it plays in the blood clotting process. The novel approaches taken in *NaCelleS* have significant potential for solving NCS problems in molecular crystallography, and a variety of options are presented. The work on data elimination answered a specific question, and can be extended in several ways to clarify the relative importance of different elements of the diffraction pattern.

5.2.1 Future work on CRLF3

The goals originally set out for the CRLF3 project were achieved. Furthermore, the results allowed insight into the mutations and effect on protein function. Gaps remain in this understanding, as the full protein could not be crystallised to determine its full structure.

Additional work could be undertaken to investigate other forms of crystallisation of the complete protein molecule, but the overall sequence statistics proposed for the protein suggest that it will remain difficult to crystallise. Further investigations into the sections of the protein that were successfully replicated in the construct are unlikely to yield much new information, but bound constructs are worth further work to investigate protein structure under these conditions. It is also possible that the full length protein will adopt a slightly different conformation when compared to the construct.

A section of the protein has not been modelled successfully as it appeared to inhibit crystallisation. The proposed constructs to investigate this section that were not successfully duplicated could also be tried in other cell lines, as protein expression was poor in the line used for this work and limited the work. In particular, a mammalian cell line might be required to properly express the full length protein, as has been seen with other proteins[294]. Crystallisation trials could then be carried out with the other constructs proposed in this work. Options of binding the section of the protein that was not included in the current construct to other proteins to permit crystallisation of a new construct should be investigated further.

In order to understand the function of this protein more fully, binding assays could be performed to determine the interaction of the CRLF3 protein with other proteins in the clotting pathways. Functional studies to directly assay activity would provide further data. This information will then help to define the role of the CRLF3 protein in the clotting pathway, and help to explain why mutations might lead to the clotting diseases.

The effect of known disease mutations could be explored more fully. The BRIDGE consortium has identified these, and this work has assisted in interpreting the effects these have on the structural folding of the protein. Creating mutant constructs of the protein containing the identified changes would permit investigation of the specific effects more fully in measuring protein stability or through functional assays. In particular, this work might then identify approaches and therapeutics to be developed to overcome clotting disorders.

5.2.2 Future work on NaCelleS

Research on the *NaCelleS* project has identified a large number of ways in which algorithms can improve the identification of NCS in macromolecular crystallography.

The one-sphere approach could be further refined to improve NCS outcome reliability. Starting from the initial NCS solution given by *phenix.find_ncs_from_density*[2], the algorithms in *NaCelleS* could generate refined results using the iterative approach, but further work is required in this area. This refinement would be both in generating better values for the operators, and potentially in the identification of additional operators. There is currently no iteration possible in *phenix.find_ncs_from_density*[2].

The novel two-sphere approach has potential to significantly aid elucidation of protein structure, and should be particularly helpful in situations where only one or two centres in a protein are known amongst a list of anomalous scatterers. Geometric approaches are known to struggle under these conditions. This method also facilitates identification of NCS when several of the anomalous scatterers only appear in some of the copies in the asymmetric unit. This method has shown a clear ability to distinguish the correct sites of the anomalous scatterers from the incorrect, with only a small region of uncertainty.

More complete automation of the two-sphere approach in choosing the NCS operators that are correct should be investigated further. Currently the validity of the NCS operators returned has to be determined manually, but using ideas relating to grouping of TFZ scores a more automatic approach should be possible.

An aspect that has not been fully explored in the two-sphere approach is determining whether NCS operators predicted from various different sites can be combined to form more complete NCS operators. Creation and coding of an algorithm in this area should result in improved NCS outcomes enhancing the method, but this remains an area of research.

For both the one and two-sphere approach, there is currently no automated way of altering the sphere radius in *NaCelleS* to adapt to the data received. Further research could be performed to see if dynamic alteration of the sphere size in response to the characteristics of the available data affects the outcome. These characteristics of the data could include data quality, resolution, size of the unit cell, quality of the phases in the input map and the size of the monomer.

When considering the averaging of electron densities when using the discovered NCS operators for a second pass through the data, the quality of fit of each operator could be checked to exclude incorrect or inaccurate results that reduce the overall quality of the new search volume, or different averaging techniques could be applied. A number of weighting approaches could be taken, depending on the quality statistics, among others. This would be particularly helpful in taking forward the iterative approach.

phaser[4] does not account for any of the input phases in scoring of the rotation function, except when determining the sites around which to cut out the electron density. This means that half the data is not used, and that the list of rotations generated is not ideal, and even misses the most appropriate rotations on occasion. *NaCelleS* uses *phaser*[4] to place its search volume within the electron density map of the asymmetric unit and to generate the score for this position, so its performance is limited by this incomplete rotation function. Any means of further exploiting the phase information in the input map would enhance *NaCelleS.*

The final refinement of the search hits is performed using the translation function that also does not use phase information in *phaser*[4], which limits performance. This could be replaced by a phased likelihood function in due course when this is implemented in *phaser*[4] to improve the NCS operators generated by *NaCelleS*. The current implementation of the phased translation function does not allow for errors or uncertainties in the search and target maps. Addition of this functionality would improve the determination of the correct solution.

Further uses of the one-sphere approach could be considered in extending the functionality of *phaser*[4] into electron microscopy (EM). This extension in the use of *phaser*[4] is an important focus of future research.

There are two broad ways this extension could be carried out. The first is using a PDB structure as a model to be fitted into an electron density map from an electron microscopy experiment. The phased translation function was developed some time ago and has now been partially incorporated into *phaser*[4]. It can exploit the phases already in the electron density map as extra information. It would be possible to get a comprehensive set of tests from *PDBeSHAPE* test data, and use this set to compare the capability of *phaser*[4] against other software[300–303]. The second approach uses EM data to help with the molecular replacement solutions. Various earlier software packages have attempted to do this but there have been difficulties. One of the most challenging problems[304–306] was that electron microscopy data contain significant information at very low resolution, whereas X-ray crystallography data were far higher resolution. However, this is becoming less of a problem as the resolution limit for electron microscopy continues to improve [307]. For this EM output, the ideas developed for and within *NaCelleS* of incorporating known phase information into *phaser*[4] will be useful and can be developed further.

phaser[4] can also manage two problems that can exist in electron microscopy maps. It can refine the magnification, which is usually poorly defined and can lead to incorrect cell parameters. It can also refine the molecular weight of the electron density being used and so refine the completeness of the model with respect to the full asymmetric unit. The use of PDB models in interpreting EM data is thus likely to be more efficient with *phaser*[4] than some current methods.

There are distinct advantages for using maximum likelihood methods such as *phaser*[4] in exploiting EM data in molecular replacement solutions. Electron microscopy maps sometimes provide better models than the current PDB models, or indeed are the only available models for molecular replacement, which would improve the likelihood of solving the structure.

In order to see how well *phaser*[4] can use electron microscopy maps with protein X-ray crystallography data, a test data set would need to be created. However, in order to proceed, electron microscopy maps would have to be matched up with protein X-ray diffraction data. This is potentially very challenging as electron density maps tend to be of large complexes, whereas X-ray structures tend to be of smaller components.

5.2.3 Future work on data pathologies

Further work could be considered to investigate how different inadequacies affect the data (e.g. lacking low resolution data at other resolutions). Prediction of the expected LLG (eLLG) for a given search model and target dataset could be explored. This would require some particular consideration when the most intense reflections are removed, as the change in the normalisation will affect the expected LLG considerably, even though it has been seen that the actual solution is still successful. This would aid the understanding of the LLG measure in dealing with data inadequacies.

The modern version of *phaser*[4] is more resilient in handling data pathologies, so new, more challenging data sets might be required if further conditions are to be tested. These could be generated by revisiting the information in Dr Rob Oeffner’s data set but applying tighter criteria such as lower original TFZ scores and smaller size molecules.

Bibliography

1. Ringwald, M. *et al.* The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium. *Nucleic Acids Res.* **39,** D849–D855 (2011).
2. Terwilliger, T. C. Finding non-crystallographic symmetry in density maps of macromolecular structures. *J. Struct. Funct. Genomics* **14,** 91–5 (2013).
3. Davies, G. J. Simple example of the molecular replacement technique: The structure determination of 3-phosphglycerate kinase from Bacillus stearothermophilus. *Jt. CCP4 ESF-EACBM Newsl. Protein Crystallogaphy* **28,** 60–67 (1993).
4. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
5. Marengo-Rowe, A. J. Structure-function relations of human hemoglobins. *Proc. (Bayl. Univ. Med. Cent).* **19,** 239–45 (2006).
6. Gubernator, K. & Böhm, H.-J. *Structure-based Ligand Design*. (1998).
7. Takenaka, T. Classical vs reverse pharmacology in drug discovery. *BJU Int.* **88 Suppl 2,** 7-10; discussion 49-50 (2001).
8. Lazo, J. S. Rear-view mirrors and crystal balls: a brief reflection on drug discovery. *Mol. Interv.* **8,** 60–3 (2008).
9. Convertino, M., Das, J. & Dokholyan, N. V. Pharmacological Chaperones: Design and Development of New Therapeutic Strategies for the Treatment of Conformational Diseases. *ACS Chem. Biol.* **11,** 1471–1489 (2016).
10. Breda, A., Valadares, N. F., Souza, O. N. de & Garratt, R. C. in *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach* (National Center for Biotechnology Information (US), 2007). at <https://www.ncbi.nlm.nih.gov/books/NBK6824/>
11. Aasland R, Abrams C, Ampe C, Ball LJ, Bedford MT, Cesareni G, Gimona M, Hurley JH, Jarchau T, Lehto VP, Lemmon MA, Linding R, Mayer BJ, Nagai M, Sudol M, Walter U, W. S. A One-Letter Notation for Amino Acid Sequences*. Tentative Rules. *Eur. J. Biochem.* **5,** 151–153 (1968).
12. Alberts, B. *et al. Molecular Biology of the Cell. 4th edition.* (Garland Science, 2002). at <https://www.ncbi.nlm.nih.gov/books/NBK26830/#_A389_>
13. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (80-.).*

181, 223–230 (1973).

14. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **37,** 205–11 (1951).
15. Priestle, J. P. *et al.* Improved dihedral-angle restraints for protein structure refinement. *J. Appl. Crystallogr.* **36,** 34–42 (2003).
16. MacArthur, M. W. & Thornton, J. M. Influence of proline residues on protein conformation. *J. Mol. Biol.* **218,** 397–412 (1991).
17. Joseph, A. P., Srinivasan, N. & de Brevern, A. G. Cis-trans peptide variations in structurally similar proteins. *Amino Acids* **43,** 1369–81 (2012).
18. MacArthur, M. W. & Thornton, J. M. Deviations from Planarity of the Peptide Bond in Peptides and Proteins. *J. Mol. Biol.* **264,** 1180–1195 (1996).
19. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7,** 95–99 (1963).
20. Bhuyan, M. S. I. & Gao, X. A protein-dependent side-chain rotamer library. *BMC Bioinformatics* **12 Suppl 1,** S10 (2011).
21. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. The penultimate rotamer library. *Proteins* **40,** 389–408 (2000).
22. Pence, H. E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **87,** 1123–1124 (2010).
23. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 486–501 (2010).
24. Merritt, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277,** 505–24 (1997).
25. Surridge, C. Astbury and the α-helix. *Nat. Struct. Biol.* **6,** 210–211 (1999).
26. Eisenberg, D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11207–10 (2003).
27. Dunitz, J. D. Pauling’s Left-Handed α-Helix. *Angew. Chemie Int. Ed.* **40,** 4167–4173 (2001).
28. KENDREW, J. C. *et al.* Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. *Nature* **185,** 422–427 (1960).
29. Hol, W. G. Effects of the alpha-helix dipole upon the functioning and structure of proteins and peptides. *Adv. Biophys.* **19,** 133–65 (1985).
30. Pauling, L. & Corey, R. B. The pleated sheet, a new layer configuration of polypeptide

chains. *Proc. Natl. Acad. Sci. U. S. A.* **37,** 251–6 (1951).

31. BLAKE, C. C. F. *et al.* Structure of Hen Egg-White Lysozyme: A Three-dimensional Fourier Synthesis at 2 Å Resolution. *Nature* **206,** 757–761 (1965).

32. Ho, B. K. & Curmi, P. M. G. Twist and shear in β-sheets and β-ribbons. *J. Mol. Biol.* **317,** 291–308 (2002).

33. Chothia, C. Conformation of twisted β-pleated sheets in proteins. *J. Mol. Biol.* **75,** 295–302 (1973).

34. Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6,** 1425–1436 (1968).

35. Lewis, P. N., Momany, F. A. & Scheraga, H. A. Chain reversals in proteins. *Biochim. Biophys. Acta* **303,** 211–29 (1973).

36. Chou, P. Y. & Fasman, G. D. Beta-turns in proteins. *J. Mol. Biol.* **115,** 135–75 (1977).

37. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34,** 167–339 (1981).

38. Barlow, D. J. & Thornton, J. M. Helix geometry in proteins. *J. Mol. Biol.* **201,** 601–19 (1988).

39. Vieira-Pires, R. S. & Morais-Cabral, J. H. 310 helices in channels and other membrane proteins. *J. Gen. Physiol.* **136,** (2010).

40. Pal, L. & Basu, G. Novel protein structural motifs containing two-turn and longer 310-helices. *Protein Eng. Des. Sel.* **12,** 811–814 (1999).

41. Enkhbayar, P., Hikichi, K., Osaki, M., Kretsinger, R. H. & Matsushima, N. 310-helices in proteins are parahelices. *Proteins Struct. Funct. Bioinforma.* **64,** 691–699 (2006).

42. Steinberg, I. Z., Harrington, W. F., Berger, A., Sela, M. & Katchalski, E. The Configurational Changes of Poly-L-proline in Solution. *J. Am. Chem. Soc.* **82,** 5263–5279 (1960).

43. Horng, J.-C. & Raines, R. T. Stereoelectronic effects on polyproline conformation. *Protein Sci.* **15,** 74–83 (2006).

44. Adzhubei, A. A. & Sternberg, M. J. E. Left-handed Polyproline II Helices Commonly Occur in Globular Proteins. *J. Mol. Biol.* **229,** 472–493 (1993).

45. Shi, Z., Woody, R. W. & Kallenbach, N. R. Is polyproline II a major backbone conformation in unfolded proteins? *Adv. Protein Chem.* **62,** 163–240 (2002).

46. Siligardi, G. & Drake, A. F. The importance of extended conformations and, in particular, the PII conformation for the molecular recognition of peptides. *Biopolymers*

37, 281–292 (1995).

47. Kelly, M. A. *et al.* Host-guest study of left-handed polyproline II helix formation. *Biochemistry* **40,** 14376–83 (2001).

48. Rath, A., Davidson, A. R. & Deber, C. M. The structure of unstructured regions in peptides and proteins: Role of the polyproline II helix in protein folding and recognition. *Biopolymers* **80,** 179–185 (2005).

49. Wang, J. & Feng, J.-A. Exploring the sequence patterns in the α-helices of proteins. *Protein Eng. Des. Sel.* **16,** 799–807 (2003).

50. Blaber, M., Zhang, X. J. & Matthews, B. W. Structural basis of amino acid alpha helix propensity. *Science* **260,** 1637–40 (1993).

51. Chakrabartty, A., Kortemme, T. & Baldwin, R. L. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **3,** 843–852 (1994).

52. Pace, C. N. & Scholtz, J. M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75,** 422–7 (1998).

53. Jacob, J., Duclohier, H. & Cafiso, D. S. The Role of Proline and Glycine in Determining the Backbone Flexibility of a Channel-Forming Peptide. *Biophys. J.* **76,** 1367–1376 (1999).

54. Doig, A. J. & Baldwin, R. L. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci.* **4,** 1325–36 (1995).

55. Doig, A. J., MacArthur, M. W., Stapley, B. J. & Thornton, J. M. Structures of N-termini of helices in proteins. *Protein Sci.* **6,** 147–55 (1997).

56. Serrano, L. & Fersht, A. R. Capping and α-helix stability. *Nature* **342,** 296–299 (1989).

57. Yan, R., Song, J., Cai, W. & Zhang, Z. in *Pattern Recognition in Computational Molecular Biology* 97–113 (John Wiley & Sons, Inc, 2015). doi:10.1002/9781119078845.ch6

58. Rost, B. Review: Protein Secondary Structure Prediction Continues to Rise. *J. Struct. Biol.* **134,** 204–218 (2001).

59. Chou, P. Y. & Fasman, G. D. Prediction of protein conformation. *Biochemistry* **13,** 222–45 (1974).

60. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41,** W349–W357 (2013).

61. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Genet.* **23,** 566–579 (1995).

62. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32,** W500–W502 (2004).

63. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).

64. Andersen, C. A. F., Palmer, A. G., Brunak, S. & Rost, B. Continuum secondary structure captures protein flexibility. *Structure* **10,** 175–84 (2002).

65. Berg, J. M., Tymoczko, J. L. & Stryer, L. in *Biochemistry* (W H Freeman, 2002). at <https://www.ncbi.nlm.nih.gov/books/NBK22375/#_ncbi_dlg_citbx_NBK22375>

66. Kaddis, C. S. *et al.* Sizing large proteins and protein complexes by electrospray ionization mass spectrometry and ion mobility. *J. Am. Soc. Mass Spectrom.* **18,** 1206–16 (2007).

67. Rose, A. S. & Hildebrand, P. W. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.* **43,** W576-9 (2015).

68. Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* **20,** 391–6 (2012).

69. Burley, S. K. *et al.* in 627–641 (Humana Press, New York, NY, 2017). doi:10.1007/978-1-4939-7000-1_26

70. IUCr. Commission on Biological Macromolecules. *Acta Crystallogr. Sect. A Found. Crystallogr.* **45,** 658–658 (1989).

71. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* **13,** 387–8 (2016).

72. Doreleijers, J. F. *et al.* BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR* **26,** 139–46 (2003).

73. Drenth, J. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2003).

74. Bragg, W. H. Bakerian Lecture: X-Rays and Crystal Structure. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **215,** 253–274 (1915).

75. Smyth, M. S. & Martin, J. H. x Ray crystallography. *Mol. Pathol.* **53,** 8–14 (2000).

76. Gernert, K. M., Smith, R. & Carter, D. C. A simple apparatus for controlling

nucleation and size in protein crystal growth. *Anal. Biochem.* **168,** 141–7 (1988).

77. Carter, C. W. Protein crystallization using incomplete factorial experiments. *J. Biol. Chem.* **254,** 12219–23 (1979).

78. Jaeger, J. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2005). at <https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0002723>

79. Carugo, O. & Argos, P. Protein-protein crystal-packing contacts. *Protein Sci.* **6,** 2261–3 (1997).

80. Schmidt, A. & Lamzin, V. S. Internal motion in protein crystal structures. *Protein Sci.* **19,** 944–53 (2010).

81. Gerstel, M., Deane, C. M. & Garman, E. F. Identifying and quantifying radiation damage at the atomic level. *J. Synchrotron Radiat.* **22,** 201–12 (2015).

82. Weik, M. *et al.* Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 623–8 (2000).

83. Meents, A., Gutmann, S., Wagner, A. & Schulze-Briese, C. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 1094–9 (2010).

84. Garman, E. F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 339–51 (2010).

85. Holton, J. M. A beginner’s guide to radiation damage. *J. Synchrotron Radiat.* **16,** 133–42 (2009).

86. Hope, H. Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallogr. B.* **44 (Pt 1),** 22–6 (1988).

87. Chinte, U. *et al.* Cryogenic (&lt;20 K) helium cooling mitigates radiation damage to protein crystals. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63,** 486–492 (2007).

88. Meents, A. *et al.* Reduction of X-ray-induced radiation damage of macromolecular crystals by data collection at 15 K: a systematic study. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63,** 302–309 (2007).

89. Alcorn, T. & Juers, D. H. Progress in rational methods of cryoprotection in macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 366–373 (2010).

90. Liu, H. & Spence, J. C. H. XFEL data analysis for structural biology. *Quant. Biol.* **4,** 159–176 (2016).

91. Levantino, M., Yorke, B. A., Monteiro, D. C., Cammarata, M. & Pearson, A. R. Using synchrotrons and XFELs for time-resolved X-ray crystallography and solution

scattering experiments on biomolecules. *Curr. Opin. Struct. Biol.* **35,** 41–48 (2015).

92. Neutze, R., Brändén, G. & Schertler, G. F. Membrane protein structural biology using X-ray free electron lasers. *Curr. Opin. Struct. Biol.* **33,** 115–125 (2015).

93. Neutze, R. Opportunities and challenges for time-resolved studies of protein structural dynamics at X-ray free-electron lasers. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **369,** 20130318 (2014).

94. Zhu, D. *et al.* A single-shot transmissive spectrometer for hard x-ray free electron lasers. *Appl. Phys. Lett.* **101,** 034103 (2012).

95. Spence, J. C. H. XFELs for structure and dynamics in biology. *IUCrJ* **4,** 322–339 (2017).

96. Amin, M., Askerka, M., Batista, V. S., Brudvig, G. W. & Gunner, M. R. X-ray Free Electron Laser Radiation Damage through the S-State Cycle of the Oxygen-Evolving Complex of Photosystem II. *J. Phys. Chem. B* **121,** 9382–9388 (2017).

97. Amin, M., Badawi, A. & Obayya, S. S. Radiation Damage in XFEL: Case study from the oxygen-evolving complex of Photosystem II. *Sci. Rep.* **6,** 36492 (2016).

98. Morgan, A. J. *et al.* High numerical aperture multilayer Laue lenses. *Sci. Rep.* **5,** 9892 (2015).

99. Kabsch, W. & IUCr. A pattern-recognition procedure for scanning oscillation films. *J. Appl. Crystallogr.* **10,** 426–429 (1977).

100. Powell, H. R., Johnson, O. & Leslie, A. G. W. Autoindexing diffraction images with iMosflm. *Acta Crystallogr. D. Biol. Crystallogr.* **69,** 1195–203 (2013).

101. Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. Robust indexing for automatic data collection. *J. Appl. Crystallogr.* **37,** 399–409 (2004).

102. Kabsch, W. XDS. *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 125–32 (2010).

103. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. *iMOSFLM* : a new graphical interface for diffraction-image processing with *MOSFLM*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67,** 271–281 (2011).

104. McAuley, K. E. *et al.* A quick solution: ab initio structure determination of a 19 kDa metalloproteinase using ACORN. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1571–8 (2001).

105. Aubert, E. & Lecomte, C. Illustrated Fourier transforms for crystallography. *J. Appl. Crystallogr.* **40,** 1153–1165 (2007).

106. Green, D. W., Ingram, V. M. & Perutz, M. F. The Structure of Haemoglobin. IV. Sign Determination by the Isomorphous Replacement Method. *Proc. R. Soc. A Math. Phys.*

Eng. Sci. **225,** 287–307 (1954).

107. Groth, P. *Chemische Kristallographie: Vol. 1*. (Leipzig: Engelmann, 1908).

108. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181,** 662–6 (1958).

109. Bragg, L. & Perutz, M. F. The Structure of Haemoglobin. VI. Fourier Projections on the 010 Plane. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **225,** 315–329 (1954).

110. Hendrickson, W. A., Lattman, E. E. & IUCr. Representation of phase probability distributions for simplified combination of independent phase information. *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **26,** 136–143 (1970).

111. Helliwell, J. R. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2010). at <https://onlinelibrary.wiley.com/doi/pdf/10.1038/npg.els.0003109>

112. Hendrickson, W. A., Ogata, C. M., Navaza, J. & Saludjian, P. *Macromolecular Crystallography Part A. null* **276,** (Elsevier, 1997).

113. Ogata, C. M. MAD phasing grows up. *Nat. Struct. Biol.* **5 Suppl,** 638–40 (1998).

114. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254,** 51–8 (1991).

115. Boggon, T. J. & Shapiro, L. Screening for phasing atoms in protein crystallography. *Structure* **8,** R143–R149 (2000).

116. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J.* **9,** 1665–72 (1990).

117. Rose, J. P. *et al.* Native SAD is maturing. *IUCrJ* **2,** 431–440 (2015).

118. Ravelli, R. B. G., Leiros, H.-K. S., Pan, B., Caffrey, M. & McSweeney, S. Specific Radiation Damage Can Be Used to Solve Macromolecular Crystal Structures. *Structure* **11,** 217–224 (2003).

119. de Sanctis, D., Zubieta, C., Felisaz, F., Caserotto, H. & Nanao, M. H. Radiation-damage-induced phasing: a case study using UV irradiation with light-emitting diodes. *Acta Crystallogr. Sect. D, Struct. Biol.* **72,** 395–402 (2016).

120. Bourenkov, G. P. & Popov, A. N. Optimization of data collection taking radiation damage into account. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 409–419 (2010).

121. Leal, R. M. F. *et al.* Experimental procedure for the characterization of radiation damage in macromolecular crystals. *J. Synchrotron Radiat.* **18,** 381–386 (2011).

122. Vogeley, L. & Luecke, H. Crystallization, X-ray diffraction analysis and SIRAS/molecular-replacenent phasing of three crystal forms of Anabaena sensory

rhodopsin transducer. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.* **62,** 388–91 (2006).

123. Hendrickson, W. A. Anomalous diffraction in crystallographic phase evaluation. *Q. Rev. Biophys.* **47,** 49–93 (2014).

124. Abbani, M. A. *et al.* Structure of the cooperative Xis-DNA complex reveals a micronucleoprotein filament that regulates phage lambda intasome assembly. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 2109–14 (2007).

125. Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. Improving radiation-damage substructures for RIP. *Acta Crystallographica Section D: Biological Crystallography* **61,** 1227–1237 (2005).

126. Zwart, P. H., Banumathi, S., Dauter, M. & Dauter, Z. Radiation-damage-induced phasing with anomalous scattering: substructure solution and phasing. *Acta Crystallographica Section D: Biological Crystallography* **60,** 1958–1963 (2004).

127. Rossmann, M. G. The molecular replacement method. *Acta Crystallogr. A.* **46 (Pt 2),** 73–82 (1990).

128. Grosse-Kunstleve, R. W. & Adams, P. D. Patterson correlation methods: a review of molecular replacement with CNS. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **57,** 1390–1396 (2001).

129. Brünger, A. T. *et al.* Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54,** 905–921 (1998).

130. Read, R. J. Molecular Replacement. *CIMR Protein Crystallography Course* (2000).

131. Fujinaga, M. & Read, R. J. Experiences with a new translation-function program. *J. Appl. Crystallogr.* **20,** 517–521 (1987).

132. Read, R. J. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **57,** 1373–1382 (2001).

133. Storoni, L. C., McCoy, A. J. & Read, R. J. Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D. Biol. Crystallogr.* **60,** 432–8 (2004).

134. Read, R. J. & Schierbeek, A. J. A phased translation function. *J. Appl. Crystallogr.* **21,** 490–495 (1988).

135. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. Likelihood-enhanced fast translation functions. *Acta Crystallogr. D. Biol. Crystallogr.* **61,** 458–64 (2005).

136. Patterson, A. A Fourier Series Method for the Determination of the Components of

Interatomic Distances in Crystals. *Phys. Rev.* **46,** 372–376 (1934).

137. Navaza, J. AMoRe : an automated package for molecular replacement. *Acta Crystallogr. Sect. A Found. Crystallogr.* **50,** 157–163 (1994).
138. Glykos, N. M. & Kokkinidis, M. Molecular replacement with multiple different models. *J. Appl. Crystallogr.* **37,** 159–161 (2004).
139. McCoy, A. J. Liking likelihood. *Acta Crystallogr. D. Biol. Crystallogr.* **60,** 2169–83 (2004).
140. Wilson, A. J. C. & IUCr. The probability distribution of X-ray intensities. *Acta Crystallogr.* **2,** 318–321 (1949).
141. Luzzati, V. & IUCr. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallogr.* **5,** 802–810 (1952).
142. Sim, G. A. & IUCr. The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. *Acta Crystallogr.* **12,** 813–815 (1959).
143. Kleywegt, G. J. & Read, R. J. Not your average density. *Structure* **5,** 1557–69 (1997).
144. Leslie, A. G. W. A reciprocal-space method for calculating a molecular envelope using the algorithm of B.C. Wang. *Acta Crystallogr. Sect. A Found. Crystallogr.* **43,** 134–136 (1987).
145. Wang, B. C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115,** 90–112 (1985).
146. Abrahams, J. P. & Leslie, A. G. Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr. D. Biol. Crystallogr.* **52,** 30–42 (1996).
147. Wilson, C. & Agard, D. A. PRISM: automated crystallographic phase refinement by iterative skeletonization. *Acta Crystallographica Section A: Foundations of Crystallography* **49,** 97–104 (1993).
148. Zhang, K. Y. J. & Main, P. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallogr. Sect. A Found. Crystallogr.* **46,** 41–46 (1990).
149. Zhang, K. Y. J. & Main, P. The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. *Acta Crystallogr. Sect. A Found. Crystallogr.* **46,** 377–381 (1990).
150. Bricogne, G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallogr. Sect. A* **30,** 395–405 (1974).
151. Rossmann, M. G. & Blow, D. M. The detection of sub-units within the

crystallographic asymmetric unit. *Acta Crystallogr.* **15,** 24–31 (1962).

152. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–42 (2000).

153. Wang, X. & Janin, J. Orientation of non-crystallographic symmetry axes in protein crystals. *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 505–12 (1993).

154. Lu, G. FINDNCS : a program to detect non-crystallographic symmetries in protein crystals from heavy-atom sites. *J. Appl. Crystallogr.* **32,** 365–368 (1999).

155. Terwilliger, T. C. Rapid automatic NCS identification using heavy-atom substructures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58,** 2213–2215 (2002).

156. Vonrhein, C. & Schulz, G. E. Locating proper non-crystallographic symmetry in low-resolution electron-density maps with the program GETAX. *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 225–9 (1999).

157. Pai, R., Sacchettini, J. & Ioerger, T. Identifying non-crystallographic symmetry in protein electron-density maps: a feature-based approach. *Acta Crystallogr. D. Biol. Crystallogr.* **62,** 1012–21 (2006).

158. Richards, F. M. The matching of physical models to three-dimensional electron-density maps: a simple optical device. *J. Mol. Biol.* **37,** 225–30 (1968).

159. Jones, T. A. & IUCr. A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11,** 268–272 (1978).

160. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A.* **47 (Pt 2),** 110–9 (1991).

161. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the *PHENIX AutoBuild* wizard. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64,** 61–69 (2008).

162. Cowtan, K. The *Buccaneer* software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62,** 1002–1011 (2006).

163. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3,** 1171–9 (2008).

164. Langer, G. G., Hazledine, S., Wiegels, T., Carolan, C. & Lamzin, V. S. Visual automated macromolecular model building. *Acta Crystallogr. D. Biol. Crystallogr.* **69,** 635–41 (2013).

165. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation

1 1Edited by J. Thornton. *J. Mol. Biol.* **285,** 1735–1747 (1999).

166. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D. Biol. Crystallogr.* **67,** 355–67 (2011).

167. Afonine, P. V *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D. Biol. Crystallogr.* **68,** 352–67 (2012).

168. Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res.* **32,** W615–W619 (2004).

169. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375–W383 (2007).

170. Brünger, A. T. [19] Free R value: Cross-validation in crystallography. *Methods Enzymol.* **277,** 366–396 (1997).

171. Tickle, I. J., Laskowski, R. A. & Moss, D. S. R_{free} and the R_{free} ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **56,** 442–450 (2000).

172. Arendall, W. B. *et al.* A test of enhancing model accuracy in high-throughput crystallography. *J. Struct. Funct. Genomics* **6,** 1–11 (2005).

173. Lovell, S. C. *et al.* Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins Struct. Funct. Bioinforma.* **50,** 437–450 (2003).

174. Word, J. M. *et al.* Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms 1 1Edited by J. Thornton. *J. Mol. Biol.* **285,** 1711–1733 (1999).

175. Croll, T. I. The rate of *cis* – *trans* conformation errors is increasing in low-resolution crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71,** 706–709 (2015).

176. Williams, C. J., Richardson, D. C. & Richardson, J. S. Through the Ramachandran Haze: Ca-Parameters Reveal Secondary Structure at Low Resolution. *Biophys. J.* **104,** 19a–20a (2013).

177. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D. Biol. Crystallogr.* **67,** 235–42 (2011).

178. Tronrud, D. E. TNT refinement package. *Methods Enzymol.* **277,** 306–19 (1997).

179. Fokine, A. & Urzhumtsev, A. Flat bulk-solvent model: obtaining optimal parameters. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58,** 1387–1392 (2002).

180. Jiang, J.-S. & Brünger, A. T. Protein Hydration Observed by X-ray Diffraction. *J. Mol.*

Biol. **243,** 100–115 (1994).

181. Riek, R. *et al.* NMR structure of the mouse prion protein domain PrP(121–231). *Nature* **382,** 180–182 (1996).

182. Krishnan, V., Rupp, B., Krishnan, V. & Rupp, B. in *eLS* (John Wiley & Sons, Ltd, 2012). doi:10.1002/9780470015902.a0002716.pub2

183. Niimura, N. *et al.* Neutron protein crystallography: beyond the folding structure of biological macromolecules. *Acta Crystallogr. Sect. A Found. Crystallogr.* **64,** 12–22 (2008).

184. Blakeley, M. P., Langan, P., Niimura, N. & Podjarny, A. Neutron crystallography: opportunities, challenges, and limitations. *Curr. Opin. Struct. Biol.* **18,** 593–600 (2008).

185. Hazemann, I. *et al.* High-resolution neutron protein crystallography with radically small crystal volumes: application of perdeuteration to human aldose reductase. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **61,** 1413–1417 (2005).

186. Afonine, P. V *et al.* Joint X-ray and neutron refinement with phenix.refine. *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 1153–63 (2010).

187. Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **161,** 450–7 (2015).

188. Glaeser, R. M. How good can cryo-EM become? *Nat. Methods* **13,** 28–32 (2015).

189. Fernandez-Leiro, R. & Scheres, S. H. W. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537,** 339–346 (2016).

190. Merk, A. *et al.* Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell* **165,** 1698–1707 (2016).

191. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the *CCP-EM* software suite. *Acta Crystallogr. Sect. D Struct. Biol.* **73,** 469–477 (2017).

192. Afonine, P. V. *et al.* Real-space refinement in *PHENIX* for cryo-EM and crystallography. *Acta Crystallogr. Sect. D Struct. Biol.* **74,** 531–544 (2018).

193. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).

194. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

195. Boulay, J.-L., O’Shea, J. J. & Paul, W. E. Molecular phylogeny within type I cytokines and their cognate receptors. *Immunity* **19,** 159–63 (2003).

196. Rawlings, J. S., Rosler, K. M. & Harrison, D. A. The JAK/STAT signaling pathway. *J. Cell Sci.* **117,** 1281–3 (2004).

197. Yang, F. *et al.* Cloning and characterization of a novel intracellular protein p48.2 that negatively regulates cell cycle progression. *Int. J. Biochem. Cell Biol.* **41,** 2240–2250 (2009).

198. Dang, C. *et al.* Identification of dysregulated genes in cutaneous squamous cell carcinoma. *Oncol. Rep.* **16,** 513–9 (2006).

199. Costa, R. M. & Silva, A. J. Review Article : Molecular and Cellular Mechanisms Underlying the Cognitive Deficits Associated With Neurofibromatosis 1. *J. Child Neurol.* **17,** 622–626 (2002).

200. Jenne, D. E. *et al.* Complete physical map and gene content of the human NF1 tumor suppressor region in human and mouse. *Genes, Chromosom. Cancer* **37,** 111–120 (2003).

201. Hashimoto, Y. *et al.* Uncovering genes required for neuronal morphology by morphology-based gene trap screening with a revertible retrovirus vector. *FASEB J.* **26,** 4662–4674 (2012).

202. Yang, F., Zhang, Y., Cao, Y.-L., Wang, S.-H. & Liu, L. Establishment and utilization of a tetracycline-controlled inducible RNA interfering system to repress gene expression in chronic myelogenous leukemia cells. *Acta Biochim. Biophys. Sin. (Shanghai).* **37,** 851–6 (2005).

203. Hahn, N. *et al.* The Insect Ortholog of the Human Orphan Cytokine Receptor CRLF3 Is a Neuroprotective Erythropoietin Receptor. *Front. Mol. Neurosci.* **10,** 223 (2017).

204. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244-8 (2005).

205. Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21,** 951–960 (2005).

206. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).

207. Li, Y. *et al.* Structural insights into the TRIM family of ubiquitin E3 ligases. *Cell Res.* **24,** 762–5 (2014).

208. Schiefner, A., Gebauer, M. & Skerra, A. Extra-domain B in oncofetal fibronectin structurally promotes fibrillar head-to-tail dimerization of extracellular matrix protein. *J. Biol. Chem.* **287,** 17578–88 (2012).

209. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* **292,** 195–202 (1999).

210. Buchan, D. W. A. *et al.* Protein annotation and modelling servers at University College

London. *Nucleic Acids Res.* **38,** W563–W568 (2010).

211. Deller, M. C., Kong, L. & Rupp, B. Protein stability: a crystallographer's perspective. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* **72,** 72–95 (2016).

212. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **337,** 635–645 (2004).

213. Jahandideh, S., Jaroszewski, L. & Godzik, A. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70,** 627–635 (2014).

214. Jaroszewski, L. *et al.* Genome Pool Strategy for Structural Coverage of Protein Families. *Structure* **16,** 1659–1667 (2008).

215. Slabinski, L. *et al.* The challenge of protein structure determination-lessons from structural genomics. *Protein Sci.* **16,** 2472–2482 (2007).

216. Slabinski, L. *et al.* XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* **23,** 3403–3405 (2007).

217. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196,** 641–56 (1987).

218. Hobohm, U. *et al.* Selection of representative protein data sets. *Protein Sci.* **1,** 409–417 (2008).

219. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).

220. Guruprasad, K., Reddy, B. V & Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **4,** 155–61 (1990).

221. Harper, S. & Speicher, D. W. in *Current Protocols in Protein Science* **Chapter 6,** 6.6.1-6.6.26 (John Wiley & Sons, Inc., 2008).

222. Aishima, J. *et al.* High-speed crystal detection and characterization using a fast-readout detector. *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 1032–5 (2010).

223. Winter, G. *et al. xia2* : an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43,** 186–190 (2010).

224. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).

225. Pflugrath, J. W. The finer things in X-ray diffraction data collection. *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 1718–25 (1999).

226. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.* **66,** 213–221 (2010).

227. Man, D. *et al.* Solution Structure of the C-terminal Domain of the Ciliary Neurotrophic Factor (CNTF) Receptor and Ligand Free Associations among Components of the CNTF Receptor Complex. *J. Biol. Chem.* **278,** 23285–23294 (2003).

228. Bunkóczi, G. & Read, R. J. Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr. D. Biol. Crystallogr.* **67,** 303–12 (2011).

229. Bravo, J., Staunton, D., Heath, J. K. & Jones, E. Y. Crystal structure of a cytokine-binding region of gp130. *EMBO J.* **17,** 1665–74 (1998).

230. Aritomi, M. *et al.* Atomic structure of the GCSF-receptor complex showing a new cytokine-receptor recognition scheme. *Nature* **401,** 713–7 (1999).

231. Chow, D., He, X., Snow, A. L., Rose-John, S. & Garcia, K. C. Structure of an extracellular gp130 cytokine receptor signaling complex. *Science* **291,** 2150–5 (2001).

232. Wang, X., Rickert, M. & Garcia, K. C. Structure of the quaternary complex of interleukin-2 with its alpha, beta, and gammac receptors. *Science* **310,** 1159–63 (2005).

233. Mendiratta, S. S. *et al.* A novel alpha-helix in the first fibronectin type III repeat of the neural cell adhesion molecule is critical for N-glycan polysialylation. *J. Biol. Chem.* **281,** 36052–9 (2006).

234. McLellan, J. S. *et al.* Structure of a heparin-dependent complex of Hedgehog and Ihog. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 17208–13 (2006).

235. Goihberg, E. *et al.* A single proline substitution is critical for the thermostabilization of Clostridium beijerinckii alcohol dehydrogenase. *Proteins* **66,** 196–204 (2007).

236. Kavran, J. M., Ward, M. D., Oladosu, O. O., Mulepati, S. & Leahy, D. J. All mammalian Hedgehog proteins interact with cell adhesion molecule, down-regulated by oncogenes (CDO) and brother of CDO (BOC) in a conserved manner. *J. Biol. Chem.* **285,** 24584–90 (2010).

237. Grebien, F. *et al.* Targeting the SH2-kinase interface in Bcr-Abl inhibits leukemogenesis. *Cell* **147,** 306–19 (2011).

238. Carpenter, B. *et al.* Structure of the human obesity receptor leptin-binding domain reveals the mechanism of leptin antagonism by a monoclonal antibody. *Structure* **20,** 487–97 (2012).

239. Nakayama, T. *et al.* Structural features of interfacial tyrosine residue in ROBO1 fibronectin domain-antibody complex: Crystallographic, thermodynamic, and

molecular dynamic analyses. *Protein Sci.* **24,** 328–40 (2015).

240. Schumacher, M. A., Chinnam, N., Ohashi, T., Shah, R. S. & Erickson, H. P. The structure of irisin reveals a novel intersubunit β-sheet fibronectin type III (FNIII) dimer: implications for receptor activation. *J. Biol. Chem.* **288,** 33738–44 (2013).

241. Khan, J. A. *et al.* Developing Adnectins that target SRC co-activator binding to PXR: a structural approach toward understanding promiscuity of PXR. *J. Mol. Biol.* **427,** 924–942 (2015).

242. Alonso-García, N. *et al.* Combination of X-ray crystallography, SAXS and DEER to obtain the structure of the FnIII-3,4 domains of integrin α6β4. *Acta Crystallogr. D. Biol. Crystallogr.* **71,** 969–85 (2015).

243. Bunkóczi, G. *et al.* Phaser.MRage: automated molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **69,** 2276–86 (2013).

244. Terwilliger, T. C. *et al.* phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. *J. Struct. Funct. Genomics* **13,** 81–90 (2012).

245. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5,** 725–738 (2010).

246. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43,** W174-81 (2015).

247. Yang, J. & Zhang, Y. Protein Structure and Function Prediction Using I-TASSER. *Curr. Protoc. Bioinforma.* **52,** 5.8.1-15 (2015).

248. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2015).

249. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins* **82 Suppl 2,** 1–6 (2014).

250. Terwilliger, T. C. *et al.* Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr. D. Biol. Crystallogr.* **65,** 582–601 (2009).

251. Zwart, P., H., Grosse-Kunstleve, R., W. & Adams, P., D. Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newsl.* 7 (2005). at <http://www.ccp4.ac.uk/newsletters/newsletter43/content.html>

252. Grosse-Kunstleve, R. W. *et al.* Substructure search procedures for macromolecular structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **59,** 1966–1973 (2003).

253. Terwilliger, T. C. *et al.* Maximum-likelihood density modification. *Acta Crystallogr.*

Sect. D Biol. Crystallogr. **56,** 965–972 (2000).

254. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 133–144 (2010).

255. Diederichs, K., McSweeney, S. & Ravelli, R. B. G. Zero-dose extrapolation as part of macromolecular synchrotron data reduction. *Acta Crystallogr. D. Biol. Crystallogr.* **59,** 903–9 (2003).

256. Merritt, E. A. X-ray Anomalous Scattering. at <http://skuld.bmsc.washington.edu/scatter/AS_index.html>

257. Terwilliger, T. C. & Berendzen, J. Evaluation of macromolecular electron-density map quality using the correlation of local r.m.s. density. *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 1872–7 (1999).

258. Terwilliger, T. C. & Berendzen, J. Discrimination of solvent from protein regions in native Fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods. *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 501–5 (1999).

259. Terwilliger, T. C. & Berendzen, J. Bayesian Correlated MAD Phasing. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **53,** 571–579 (1997).

260. Terwilliger, T. C. & Berendzen, J. Correlated Phasing of Multiple Isomorphous Replacement Data. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **52,** 749–757 (1996).

261. DeLano Scientific LLC, Palo Alto, California, U. The PyMOL Molecular Graphics System.

262. Urzhumtseva, L., Afonine, P. V, Adams, P. D. & Urzhumtsev, A. Crystallographic model quality at a glance. *Acta Crystallogr. D. Biol. Crystallogr.* **65,** 297–300 (2009).

263. Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45,** D190–D199 (2017).

264. Fraser, J. S., Yu, Z., Maxwell, K. L. & Davidson, A. R. Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit. *J. Mol. Biol.* **359,** 496–507 (2006).

265. Koide, A., Bailey, C. W., Huang, X. & Koide, S. The fibronectin type III domain as a scaffold for novel binding proteins. *J. Mol. Biol.* **284,** 1141–1151 (1998).

266. Masters, S. L. *et al.* Genetic Deletion of Murine SPRY Domain-Containing SOCS Box Protein 2 (SSB-2) Results in Very Mild Thrombocytopenia. *Mol. Cell. Biol.* **25,** 5639–5647 (2005).

267. Kuang, Z. *et al.* SPRY Domain-Containing SOCS Box Protein 2: Crystal Structure and Residues Critical for Protein Binding. *J. Mol. Biol.* **386,** 662–674 (2009).

268. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*

536, 285–291 (2016).

269. Fraczkiewicz, R. & Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19,** 319–333 (1998).

270. Hayryan, S., Hu, C.-K., Skřivánek, J., Hayryane, E. & Pokorný, I. A new analytical method for computing solvent-accessible surface area of macromolecules and its gradients. *J. Comput. Chem.* **26,** 334–343 (2005).

271. Deane, J. E. *et al.* Insights into Krabbe disease from structures of galactocerebrosidase. *Proc. Natl. Acad. Sci.* **108,** 15169–15173 (2011).

272. Demydchuk, M. *et al.* Insights into Hunter syndrome from the structure of iduronate-2-sulfatase. *Nat. Commun.* **8,** 15786 (2017).

273. Chen, S., Jancrick, J., Yokota, H., Kim, R. & Kim, S.-H. Crystal structure of a protein associated with cell division from Mycoplasma pneumoniae (GI: 13508053): a novel fold with a conserved sequence motif. *Proteins* **55,** 785–91 (2004).

274. Ducros, V. M.-A. *et al.* Crystal structure of GerE, the ultimate transcriptional regulator of spore formation in Bacillus subtilis. *J. Mol. Biol.* **306,** 759–771 (2001).

275. Shin, D. H. *et al.* Crystal structure of a phosphatase with a unique substrate binding domain from *Thermotoga maritima*. *Protein Sci.* **12,** 1464–1472 (2003).

276. Huang, C. -c., Smith, C. V., Glickman, M. S., Jacobs, W. R. & Sacchettini, J. C. Crystal Structures of Mycolic Acid Cyclopropane Synthases fromMycobacterium tuberculosis. *J. Biol. Chem.* **277,** 11559–11569 (2002).

277. Turner, M. A. *et al.* Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. *Nat. Struct. Biol.* **5,** 369–76 (1998).

278. Gordon, E. *et al.* Crystal Structure of UDP-N-acetylmuramoyl-L-alanyl-D-glutamate: meso-Diaminopimelate Ligase from Escherichia Coli. *J. Biol. Chem.* **276,** 10999–11006 (2001).

279. Chen, C. C. H. *et al.* Degradation pathway of the phosphonate ciliatine: crystal structure of 2-aminoethylphosphonate transaminase. *Biochemistry* **41,** 13162–9 (2002).

280. Ostermeier, C. & Brunger, A. T. Structural basis of Rab effector specificity: crystal structure of the small G protein Rab3A complexed with the effector domain of rabphilin-3A. *Cell* **96,** 363–74 (1999).

281. Jiang, J., Zhang, Y., Krainer, A. R. & Xu, R. M. Crystal structure of human p32, a doughnut-shaped acidic mitochondrial matrix protein. *Proc. Natl. Acad. Sci. U. S. A.*

96, 3572–7 (1999).

282. Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. Structure of cyanase reveals that a novel dimeric and decameric arrangement of subunits is required for formation of the enzyme active site. *Structure* **8,** 505–14 (2000).

283. Brunger, A. T., Sutton, R. B., Fasshauer, D. & Jahn, R. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature* **395,** 347–353 (1998).

284. Eicken, C. *et al.* Crystal Structure of Lyme Disease Variable Surface Antigen VlsE of Borrelia burgdorferi. *J. Biol. Chem.* **277,** 21691–21696 (2002).

285. Muchmore, C. R. A., Krahn, J. M., Smith, J. L., Kim, J. H. & Zalkin, H. Crystal structure of glutamine phosphoribosylpyrophosphate amidotransferase from *Escherichia coli*. *Protein Sci.* **7,** 39–51 (1998).

286. Daniels, D. L., Cohen, A. R., Anderson, J. M. & Brünger, A. T. Crystal structure of the hCASK PDZ domain reveals the structural basis of class II PDZ domain target recognition. *Nat. Struct. Biol.* **5,** 317–25 (1998).

287. Shin, D. H. *et al.* Crystal structure of TM1457 from Thermotoga maritima. *J. Struct. Biol.* **152,** 113–117 (2005).

288. Esser, L. *et al.* Synapsin I is structurally similar to ATP-utilizing enzymes. *EMBO J.* **17,** 977–984 (1998).

289. Sharma, V. *et al.* Structure of isocitrate lyase, a persistence factor of Mycobacterium tuberculosis. *Nat. Struct. Biol.* **7,** 663–668 (2000).

290. Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science* **271,** 72–7 (1996).

291. Matthews, B. W. Solvent content of protein crystals. *J. Mol. Biol.* **33,** 491–7 (1968).

292. Kantardjieff, K. A. & Rupp, B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12,** 1865–71 (2003).

293. Joachimiak, A. *Structural genomics Part C*. (Academic Press, 2009).

294. Taberman, H. Radiation Damage in Macromolecular Crystallography—An Experimentalist's View. *Crystals* **8,** 157 (2018).

295. Strauss, M. G. *et al.* CCD-based detector for protein crystallography with synchrotron X-rays. *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect.*

Assoc. Equip. **297,** 275–295 (1990).

296. DECTRIS. PILATUS 3 S and M series specification. (2014).

297. Mueller, M., Wang, M. & Schulze-Briese, C. Optimal fine ϕ-slicing for single-photon-counting pixel detectors. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68,** 42–56 (2011).

298. Oeffner, R. D., Bunkóczi, G., McCoy, A. J., Read, R. J. & IUCr. Improved estimates of coordinate error for molecular replacement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69,** 2209–2215 (2013).

299. Dauter, Z. in *Methods in Molecular Biology (Clifton, N.J.)* **1607,** 165–184 (2017).

300. Allen, G. S. & Stokes, D. L. Modeling, docking, and fitting of atomic structures to 3D maps from cryo-electron microscopy. *Methods Mol. Biol.* **955,** 229–41 (2013).

301. Zhou, Z. H. Atomic resolution cryo electron microscopy of macromolecular complexes. *Adv. Protein Chem. Struct. Biol.* **82,** 1–35 (2011).

302. Rawat, U. *et al.* Interactions of the release factor RF1 with the ribosome as revealed by cryo-EM. *J. Mol. Biol.* **357,** 1144–53 (2006).

303. Rayment, I. *et al.* Structure of the actin-myosin complex and its implications for muscle contraction. *Science* **261,** 58–65 (1993).

304. Scapin, G. Molecular replacement then and now. *Acta Crystallogr. D. Biol. Crystallogr.* **69,** 2266–75 (2013).

305. Xiong, Y. From electron microscopy to X-ray crystallography: molecular-replacement case studies. *Acta Crystallogr. D. Biol. Crystallogr.* **64,** 76–82 (2008).

306. Navaza, J. Combining X-ray and electron-microscopy data to solve crystal structures. *Acta Crystallogr. D. Biol. Crystallogr.* **64,** 70–5 (2008).

307. Hashem, Y. *et al.* High-resolution cryo-electron microscopy structure of the Trypanosoma brucei ribosome. *Nature* **494,** 385–9 (2013).