Multi-dataset electron density analysis methods for X-ray crystallography

Nicholas Pearce

St Cross College
University of Oxford

Supervisors

Prof. Frank von Delft	*University of Oxford* *Diamond Light Source*
Prof. Charlotte Deane	*University of Oxford*
Dr Jiye Shi	*UCB Pharma*
Dr Sebastian Kelm	*UCB Pharma*

A thesis submitted for the degree of
Doctor of Philosophy in Systems Approaches to Biomedical Science
Trinity Term 2016

To all those that helped,

And all those that hindered.

It wouldn’t have been the same without either.

Multi-dataset electron density analysis methods for X-ray crystallography
Nicholas Pearce, St Cross College
Submitted for the degree of DPhil. Systems Approaches to Biomedical Science
Trinity Term, 2016

Abstract

X-ray crystallography is extensively deployed to determine the structure of proteins, both unbound and bound to different molecules. Crystallography has the power to visually reveal the binding of small molecules, assisting in their development in structure-based lead design. Currently, however, the methods used to detect binding, and the subjectivity of inexperienced modellers, are a weak-point in the field.

Existing methods for ligand identification are fundamentally flawed when identifying partially-occupied states in crystallographic datasets; the ambiguity of conventional electron density maps, which present a superposition of multiple states, prevents robust ligand identification. In this thesis, I present novel methods to clearly identify bound ligands and other changed states in the case where multiple crystallographic datasets are available, such as in crystallographic fragment screening experiments. By applying statistical methods to signal identification, more crystallographic binders are detected than by state-of-the-art conventional approaches.

Standard modelling practice is further challenged regarding the modelling of multiple chemical states in crystallography. The pervading modelling approach is to model only the bound state of the protein; I show that modelling an ensemble of bound and unbound states leads to better models. I conclude with a discussion of possible future applications of multi-datasets methods in X-ray crystallography, including the robust identification of conformational heterogeneity in protein structures.

Author Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in text.

Nicholas Pearce
1st October 2016

Acknowledgements

I am completely indebted to both Frank and Charlotte for helping me through this project; it would not have been possible without both of them. Frank constantly forced me to think of the (insatiable) needs of the user and in doing so has greatly improved the tools that have emerged from the project. Charlotte always made time for the problems of the token crystallographer. I am also extremely grateful to all the staff at the SGC and Diamond who provided me with the mountain of data that made this project feasible.

To my family, my flatmates, my supervisors, and all those at the SGC and in OPIG, thank you for making the last years an enjoyable experience.

And finally, to Lindsay. She helped too.

I'm sorry for always working during our Skype calls.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

General

ASU	Asymmetric unit
DLS	Diamond Light Source
NCS	Non-crystallographic symmetry
NMR	Nuclear magnetic resonance
SBLD	Structure-based lead design
SGC	Structural Genomics Consortium
FBLD	Fragment-based lead design
RMS	Root mean-square

Density Metrics

RSCC	Real-space correlation coefficient
RSR	Real-space R
RSZD	Real-space Z-difference score
RSZO	Real-space Z-observed score

Model Metrics

RMSD	Root-mean-squared deviation

Proteins Acronyms

BAZ2B	Bromodomain adjacent to Zinc-finger domain 2B
BRD1	Bromodomain-containing protein 1
JMJD2D	Lysine demethylase 4D
SP100	Nuclear auto-antigen SP100

PanDDA Acronyms

BDC	Background density correction

Statistical Acronyms

AIC	Akaike information criterion
ROC	Receiver operator characteristic
AUC	Area under the (receiver operator characteristic) curve
GLM	Generalised linear model

Thesis Outline

This thesis is chiefly concerned with methods for identifying and modelling ligands in macromolecular crystallographic datasets. In particular, I focus on crystallographic fragment screening experiments, which aim to identify small weakly-binding molecules that bind to a target protein.

Chapter 1 covers the foundational material required for the thesis, including a description of current methods for ligand identification and automated modelling. In Chapter 2 I study the viability of applying current methods to the detection of crystallographic ligands.

Having found that current ligand-detection methods are fundamentally hindered by map quality, in Chapter 3 I describe a novel multi-dataset approach to improving the sensitivity and clarity of ligand identification; I apply these methods in Chapter 4.

In Chapter 5, I compare my novel method to a state-of-the-art conventional ligand-identification approach, and see a significant increase in the number of identified ligands, and improved clarity of previously-identified ligands. Additionally, I question current modelling omissions in X-ray crystallography, such as the failure to simultaneously model bound and unbound states in crystallographic structures.

In Chapter 6, I show preliminary results of another multi-dataset approach to identify structural heterogeneity in crystallographic datasets, showing that multi-dataset methods have further applicability outside of ligand identification and hold a great deal of promise for future algorithmic approaches to crystallographic modelling.

Chapter 1

Introduction

This chapter introduces the foundational material that is used throughout this thesis to discuss ligand identification in crystallographic fragment screening datasets. I begin with an introduction to proteins and protein structure (section 1.1) and the use of protein-ligand complex structures in drug discovery (section 1.2). Following this, I describe the methods required for macromolecular structure determination and validation using X-ray crystallography (section 1.3), and conclude with a discussion of current approaches for the detection, modelling and validation of small molecules in crystallographic datasets (section 1.4).

1.1 Introduction to proteins

Proteins are macromolecules formed of linear chains of covalently-bonded amino acid residues (Figure 1.1). There are 20 natural amino acids used in proteins (Figure 1.2), and a sequence of these defines a protein. Amino acids comprise two components: a backbone and a sidechain. The backbone atoms are common to all amino acids, and the non-hydrogen atoms consist of N, C, C_α and O (Figure 1.1). Sidechain atoms vary between amino acids, and endow them with different hydrophilicities, polarizabilities, and charge. The particular energetic preferences of residues to be e.g. exposed to solvent, or buried in the centre of the protein, drive proteins to fold spontaneously and consistently into well-defined three-dimensional structures (Alberts et al. 2014).

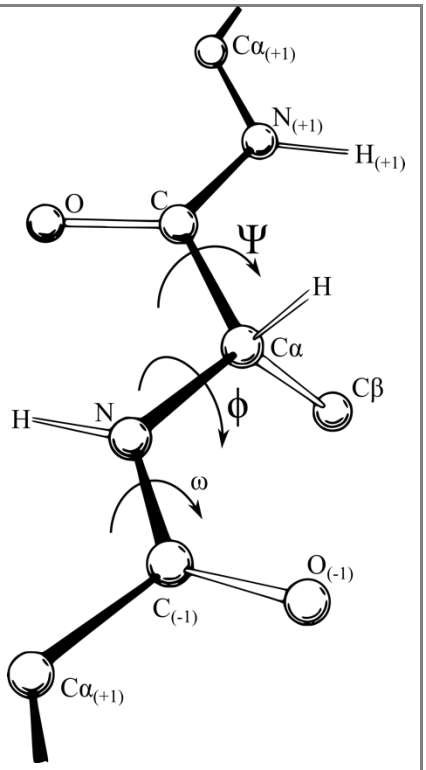
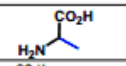
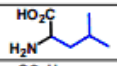
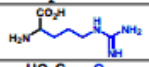
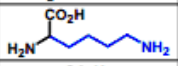
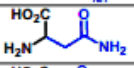
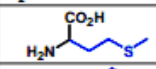
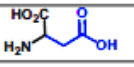
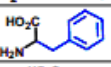
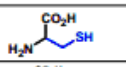
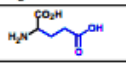
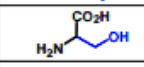
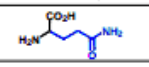
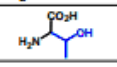
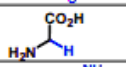
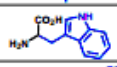
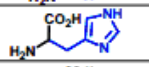
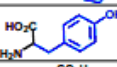
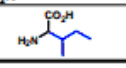
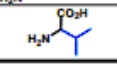


Figure 1.1. ***Amino acid atom names and angle definitions.*** *The n-1th, nth, and n+1th amino acid residues of a protein are shown, as indicated by subscripts in brackets after the atom element. Numbering of residues increases from the N-terminus to the C-terminus, in the same direction as the protein is translated. The non-H backbone of an amino acid is composed of the N, C, C_α and O atoms. The sidechains are composed of variable atoms; these different configurations endow the different types of amino acids with a variety of properties. The naming of carbon atoms is incremental (C, C_α, C_β, C_δ, etc); all carbons other than C and C_α occur in the sidechain of the residue. The sidechain for the pictured (nth) amino acid consists of only a C_β atom, making it an alanine. All natural protein side-chains are shown in Figure 1.2. The three principal dihedral angles of amino acids are indicated by the curved arrows (top to bottom: psi, phi and omega angles). (Image: Wikipedia; en.wikipedia.org/wiki/Dihedral_angle)*

The 20 natural amino acids

Name	Abbrev.	Structure*	**	Name	Abbrev.	Structure*	**
Alanine	Ala, A		H	Leucine	Leu, L		H
Arginine	Arg, R		B	Lysine	Lys, K		B
Asparagine	Asn, N		P	Methionine	Met, M		H
Aspartic Acid	Asp, D		A	Phenylalanine	Phe, F		H
Cysteine	Cys, C		H	Proline	Pro, P		H
Glutamic Acid	Glu, E		A	Serine	Ser, S		P
Glutamine	Gln, Q		P	Threonine	Thr, T		P
Glycine	Gly, G		H	Tryptophan	Trp, W		H
Histidine	His, H		B	Tyrosine	Tyr, Y		P
Isoleucine	Ile, I		H	Valine	Val, V		H

*Side Chain shown in Blue.

**Letters denote side chain properties: H = Hydrophobic, P = Polar, A = Acidic, B = Basic.

Figure 1.2. ***The 20 natural amino acids used in proteins, and some of their properties.*** *Amino acids are shown in the unconnected monomeric form; each COOH reacts with the NH_2 another amino acid to produce the peptide bond and H_2O, elongating the peptide chain.*

1.1.1 Preferred amino acid backbone conformations

Amino acid backbones in a protein have a limited number of conformations that can be adopted without causing energetically unfavourable clashes between atoms. The three principal angles that define an amino acid backbone conformation are shown in Figure 1.1. The ω angle is constrained to be either 0° (cis, rare) or 180° (trans, common); this planarity permits energetically favourable pi-orbital delocalization. Ψ and ϕ are less constrained, but allowed values are still restricted due to steric clashes with other backbone atoms; the energetically favoured and disfavoured angle regions are displayed as a Ramachandran plot (Ramachandran et al. 1963). Different amino acids have different preferred/allowed conformations due to their sidechains; proline, where the sidechain is bonded back to the backbone (making it an imino acid), is very restricted, whereas glycine, whose sidechain is a hydrogen atom, is very flexible.

1.1.2 Protein structure-function dependency

Proteins are crucial components of living systems, carrying out a highly diverse range of functions that are necessary for the continuation of life. The structure of proteins is intimately related to their function: specific arrangements of functional atoms are required for e.g. catalysis. The three-dimensional structures of proteins are correspondingly diverse, since function-specific structural arrangements of protein atoms are required for each task (Berg et al. 2002).

Elucidation of a protein structure can thus allow an analysis of the interactions that a protein makes with regulatory molecules, chemical substrates, and/or other proteins, revealing how it performs that function (Williams & Daviter 2013). Some proteins display *allosteric* sites, which may be distant from the *main* binding site, but still maintain the ability to modulate protein function. The binding of regulatory molecules

at allosteric sites can affect the activity of the protein *in vivo*, either increasing or inhibiting its ability to perform its function (Liu & Nussinov 2016).

Binding partners interact with proteins via a range of interactions, including hydrogen bonding and charged interactions. Binding molecules require a number of energetically-favourable contacts to the binding site to allow for high binding affinities. The availability of a protein structure with the ligand bound enables both a visual and a quantitative analysis of such binding events (Williams & Daviter 2013).

1.2 Protein structures in rational drug design

Physically inhibiting or modulating a protein's function with a small molecule that binds strongly and selectively *in vivo* – a drug – can be a useful tool for medical treatment. The design of compounds that bind to specific sites on a specific protein greatly benefits from knowledge of the protein structure, whether determined by X-ray crystallography and NMR or through modelling (e.g. Verlinde & Hol 1994; Congreve et al. 2005).

1.2.1 Structure-based lead design

Modern structure-based lead design (SBLD) extensively deploys X-ray crystallography to determine the structure of ligands bound to a "target" protein (e.g. Blundell et al. 2002; Tickle et al. 2004). In SBLD, ligand-bound structures are used to inform the chemical elaboration and optimisation of initially identified binding "hits" into "leads"; the intention is to increase affinity and other favourable traits, such as selectivity for the protein of interest and stability *in vivo,* whilst reducing effects such as toxicity (Hughes et al. 2011). Structural information guides the decisions of medicinal chemists in optimising contacts with the protein or in making new interactions within the binding site. Thus, structural information enables "rational" structure-based lead design.

Historically, structure-based lead design programmes have relied on high-throughput screening (HTS) techniques to identify binding ligands, involving biophysical screening of millions of large drug-like molecules against a protein (Macarron et al. 2011; Hughes et al. 2011). Structures of the most potent compounds are subsequently determined in complex with the protein, and cycles of compound optimisation begin (Anderson 2003). However, since the HTS compounds screened are already large drug-like molecules, optimisation of such compounds is difficult, contributing to the high attrition rates of molecules in drug-discovery pipelines (Murray & Rees 2009).

1.2.2 Fragment-based lead design

Fragment-based lead design (FBLD) efforts within SBLD, in contrast to HTS-based approaches, utilise libraries of small "fragment" molecules of approximately 150-300Da, commonly composed of approximately ten non-hydrogen atoms (Carr et al. 2005; Murray & Rees 2009; Scott et al. 2012).

Binding fragments, due to their size, bind only weakly to the target protein, in the mM-μM affinity range (Murray & Rees 2009). However, binding fragments often make very favourable interactions with the protein, resulting in a high ligand-efficiency (binding energy per atom of the ligand) relative to HTS hits (Murray & Rees 2009).

High-quality interactions are a common feature of fragment binding as the relatively large entropic loss of the ligand upon binding must be compensated for by only a small number of protein-ligand interactions (Chilingaryan et al. 2012): the interactions must therefore generally be highly favourable for binding to occur at all. The entropic penalty to binding may however be compensated for by the release of other bound solvent molecules into solution.

Fragments that bind to the target protein – "hits" – serve as starting points for elaboration into "leads" by medicinal chemists, using SBLD. Based on structural information, compound development can be performed either by "growing" fragments through chemical elaboration, or by chemically "linking" fragments that bind to distinct but proximate regions on the protein (Murray & Blundell 2010).

The weakness of fragment binding is not indicative of utility, and weak fragments can be developed into effective binders after only a small number of synthetic steps (Schiebel, Radeva, et al. 2016). This further speaks to the high quality of protein-fragment interactions, which may be preserved during compound elaboration (Zerbe et al. 2012; Kozakov et al. 2015). Key interactions may indicate "hotspots" on the protein surface, and they contribute disproportionately to binding; fragment screening experiments are capable of discovering these hotspots (Ludlow et al. 2015).

Since fragments bind only weakly to the protein, a sensitive detection method of binding is required. Prior to the foundation of Astex Pharmaceuticals (see section 1.2.3), it had historically been considered too difficult to perform crystallographic screening of complete libraries of compounds; other methods were (and still are) applied to pre-screen the fragments before crystallographic structure determination.

Biophysical assay (bioassay) methods are able to determine weak binding, and can obtain the high throughput required to efficiently screen ~1000 compounds against a protein (Renaud et al. 2016). Screening cascades, which require fragments to be detected by more than one biophysical method – to reduce the number of false-positives – are amongst the most common methods of pre-screening (e.g. Silvestre et al. 2013).

However, the overlap between compounds identified by biophysical techniques and crystallography is low (e.g. Schiebel, Radeva, et al. 2016), implying that screening cascades are not the most efficient way of generating what is sought: crystallographic ligand-bound structures for use in SBLD.

1.2.3 Direct crystallographic fragment screening

Companies such as Astex Pharmaceuticals (formerly Astex Technology Ltd.) have pioneered the use of routine primary crystallographic fragment screening, whereby a whole fragment library is screened directly against a protein by X-ray crystallography (Carr & Jhoti 2002; Hartshorn et al. 2005), rather than simply using crystallography to determine the structures of a set of hits identified by other methods.

Rather than screening with one compound per crystal, which was until recently prohibitively costly and time-consuming, a cocktailing approach was applied: multiple compounds are added to the same crystal, reducing the required number of crystallographic datasets several-fold (Nienaber et al. 2000; Hartshorn et al. 2005). However, recent advances in the automation of macromolecular beamlines enable the collection of single-compound datasets (crystals where only one compound has been added) for full fragment screening libraries (e.g. Schiebel, Krimmer, et al. 2016), including currently unpublished work from Diamond Light Source, beamline i04-1.

I will now describe current methods for the determination of structures by X-ray crystallography (section 1.3), before returning to protein-ligand crystallography and the task of identifying bound ligands in crystallographic datasets (section 1.4). Ligand identification is a key stage of fragment screening by X-ray crystallography.

1.3 Macromolecular crystallography

Protein structures are extensively determined experimentally in order to study them, the effects of perturbations to the structures, and the interactions of proteins with their binding partners. One of the most commonly-used methods of structure determination is macromolecular crystallography (MX), comprising 89% (on 2016-08-09) of the experimentally-determined protein structures in the PDB (Berman et al. 2000; Berman et al. 2003), a databank of freely-available protein structures.

X-ray crystallography is a powerful method for the determination of the atomic structures of molecules. By shining a beam of X-rays on molecules regularly arranged in a lattice (crystals), diffraction patterns are formed. These diffraction patterns can be measured and used to derive the electron density for the repeated building blocks of the crystal (unit cells), allowing a model for the atoms in the crystal to be built.

1.3.1 Crystals & crystallisation

Crystallography, unsurprisingly, requires crystals. The fundamentals of crystal lattices and crystal symmetry are covered in Appendix A.1. The chirality of proteins reduces the possible crystal symmetry operations to translational symmetry, rotational symmetry and screw axes (the combination of the previous two) only; protein crystals adopt one of the 65 space groups that satisfy this constraint.

Protein crystals, unlike mineralogical crystals, are formed of loose networks of large molecules, held together by weak non-covalent interactions. Obtaining protein crystals is a difficult and time-consuming process, as the proteins must be placed into the right condition for them to spontaneously self-assemble into crystals. Determining these conditions is a hit-and-miss process, requiring the sampling of a large number of

different conditions (McPherson & Gavira 2014). Further considerations of the difficulties concerning protein crystallisation are not considered here.

1.3.2 Theory of diffraction and the phase problem

Once a crystal has been obtained, diffraction experiments are performed to determine the atomic composition of the unit cell of the crystal. When a beam of light is incident on a crystal, scattering from the regularly-arranged molecules leads to distinct, regularly-spaced spots far from the crystal due to interference – a diffraction pattern. The theoretical basis of the diffraction pattern is derived in Appendix A.2.

The strength of interference at the diffraction spots is described by a complex number, $\mathrm{A}(\boldsymbol{h}) = \mathrm{A}_0 e^{i\phi}$, where $\boldsymbol{h}$ is the Miller index of the reflection. In practice, however, we do not measure the complex number $\mathrm{A}(\boldsymbol{h}) = \mathrm{A}_0 e^{i\phi}$, from which the structure factors, $\mathrm{F}(\boldsymbol{h})$, are calculated, but rather the intensities, $\mathrm{I}(\boldsymbol{h}) = |\mathrm{A}_0|^2$, the squared-magnitude of the amplitude. This loss of information constitutes the *phase problem* in crystallography; only half of the information required to reconstruct the crystallographic electron density is measured. Since the phases are not determined in the primary diffraction experiment, other methods must be applied in order to determine a set of phases, and allow density reconstruction (Taylor 2010).

Methods for "phasing" of a dataset – the estimation of an initial set of phases – are not covered in depth here, but approaches include: utilising anomalous diffraction, where the absorption and delayed elastic re-emission from certain types of atom results in phase retardation; isomorphous replacement, where heavy-atom derivatives of the crystal system offer perturbed sets of structure factors which allow an initial phase

estimate to be derived; and molecular replacement (MR), where a "phase transplant" is performed from a previously-solved protein structure.

Once an initial set of phases has been calculated/chosen, the model-building process can begin. However, the use of an imperfect or incomplete set of phases leads to a new problem: *phase bias*.

1.3.3 Phase bias in crystallography

Phases have a significant effect on the reconstructed electron density: the electron density strongly resembles the model from which the phases are derived. This phenomenon is known as *phase bias*, and is a fundamental characteristic of structure factors formed from an imperfect set of phases (discussed in detail in Appendix A.3).

When the density is biased to look like the source of the phases used, the crystallographer is at risk of modelling the object used to generate the phases, rather than the true atomic representation of the crystal, which provided the amplitudes.

To correct for phase bias, weightings are applied to the experimental and modelled amplitudes; these corrected amplitudes are then combined to create *composite maps* (Read 1986), which are heavily used in practical crystallography for model building. Methods for minimising phase bias using phase probabilities result in the figure-of-merit weighting, m, and the sigma-A weighting factor, D (Appendix A.3); these factors are used to weight the observed ($\mathrm{F_o}$) and model-calculated ($\mathrm{F_c}$) diffraction amplitudes respectively.

The weighted diffraction amplitudes, $m\mathrm{F_o}$ and $\mathrm{DF_c}$, are combined with integer multipliers and, using phase information (typically from the model), lead to the generation of composite $2m\mathrm{F_o}$-$\mathrm{DF_c}$ maps and $m\mathrm{F_o}$-$\mathrm{DF_c}$ maps (Appendix A.3.3). The

$2m\mathrm{F_o}$-$\mathrm{DF_c}$ map approximates the electron density of the crystal and the $m\mathrm{F_o}$-$\mathrm{DF_c}$ map approximates the difference between the electron density of the crystal and the model that was used to generate the $\mathrm{F_c}$ amplitudes.

These maps are minimally biased towards the source of the phases, and further increase the strength of un-modelled features, which appear weaker than expected due to inaccuracy of the phases or incompleteness of the model.

1.3.4 Usage and representation of crystallographic maps

Here, I briefly detail a few notes on the generation and representation of crystallographic maps. The $\mathrm{F_{000}}$ reflection – where $\mathrm{h} = (0,0,0)$ – corresponds to forward scattering of the electrons at an angle of $\theta = 0$. This reflection cannot be measured in crystallographic experiments, as it is swamped by the un-scattered X-ray beam. From Appendix A.2, the structure factor for $\mathrm{F_{000}}$ is

$$\mathrm{F}(\mathbf{0}) = \int_{unit\ cell} \rho(\boldsymbol{x}')\, d^3\boldsymbol{x}' = N_{electrons}, \qquad 1.1$$

where $N_{electrons}$ is the number of electrons in the unit cell. In the reconstruction of the electron density map, the reflection $\mathrm{F_{000}}$ serves to put the map on an absolute scale. Because this reflection is *never* determined, the reconstructed electron density maps have an average value of zero. Recent publications discuss the estimation of the $\mathrm{F_{000}}$ reflection (Lang et al. 2014), but such approaches are yet to see universal uptake, and it is still commonplace for maps to be rescaled internally, resulting in traditional "sigma-scaled" maps. These sigma-scaled maps, $\rho_\sigma(\boldsymbol{x})$, are calculated as

$$\rho_\sigma(x) = \frac{\rho(x)}{rms\big(\rho(x)\big)}. \qquad 1.2$$

Rescaled map values are still centred on zero, but now have an RMS of one; both $2m\mathrm{F_o}$-$\mathrm{DF_c}$ maps and $m\mathrm{F_o}$-$\mathrm{DF_c}$ maps are scaled under equation 1.2 (above).

Electron density maps are displayed as a mesh whose vertices are defined by a grid, where the mesh lines approximate electron density contours at a particular value. Sigma-scaled maps are normally viewed at standard "sigma levels": $2mF_o$-DF_c maps are typically displayed as a blue mesh contoured at approximately 1σ, and mF_o-DF_c maps are viewed as a green and red mesh contoured at $+3\sigma$ and -3σ, respectively.

These levels permit electron density visualisation whilst minimising the influence of "noisy" density that appears at low contour levels (see section 1.3.8). The 3σ level of the mF_o-DF_c map in particular is intended to reflect a conventional cut-off for "significance" of a set of values that is normally distributed. Peaks in the difference map above $\pm 3\sigma$ reflect "significant" differences between the model and the observed data, which the crystallographer should correct.

1.3.5 Model-building and refinement

Once initial estimates for the phases are obtained, the modelling process begins. The $2mF_o$-DF_c map (approximating the experimental electron density) and mF_o-DF_c map (showing differences between the experimental density and the model) are the principal maps used for modelling. The mF_o-DF_c map is used to highlight regions of the model that are unmodelled or incorrect, and the combination of the two maps is then used to guide the placement of atoms.

Modelling can be performed through visual modelling of the electron density by a crystallographer using tools such as Coot (Emsley et al. 2010), or by automated computational methods (e.g. Langer et al. 2008; Terwilliger et al. 2007).

Once all of the atoms that can be confidently modelled have been placed – given the limits imposed by the inaccuracy of the phases – the model is subjected to

crystallographic *refinement*, a process whereby the model parameters are optimised against the experimental data (Chapter 12, Rupp 2010).

Each atom placed has: three positional coordinates (x,y,z); a crystallographic occupancy, which represents the fraction of the unit cells that contain that atom at that position; and a temperature factor, which describes the disorder of this atom throughout the crystal due to thermal fluctuations. At low-to-moderate resolution, the thermal displacement, the so-called B-factor, is parameterised by a single number, modelling isotropic thermal motion; it can also be described anisotropically at higher resolution, with up to six parameters defining an ellipsoid of motion (Trueblood et al. 1996).

Refinement of a model consists of the optimisation of these model parameters against the observed data. There are several refinement programs available (e.g. Murshudov et al. 2011; Afonine et al. 2012; Bricogne et al. 2011). However, since a crystallographic model can consist of thousands of atoms, each of which is modelled by a minimum of four parameters (the occupancy is held at unity for large numbers of atoms), it is possible to have a comparable or much larger number of model parameters to observed amplitudes, especially at low resolution (where fewer data are collected).

As such, different resolution-dependent refinement procedures are applied; more complex refinement protocols, which allow more model parameters, are only applied when higher resolution data is available, such that a satisfactory data-to-parameter ratio is maintained (Chapter 12, Rupp 2010). If there is a low ratio of observed data to model parameters, careless refinement will allow us to *over-fit* the data, and this will in general lead to poor models (where a model is generated that fits the noise in the data,

rather than being a good model of the crystal). Two principal methods are applied to prevent and monitor overfitting: restraints and orthogonal validation.

Crystallographic restraints in refinement (and modelling)

To minimise the amount of overfitting that can occur, restraints are imposed on models during model-building and refinement to enforce our knowledge of physics and chemistry (e.g. Murshudov et al. 2011). Restraints account for features such as the length of a bond between two atoms, or the penalisation of a significant overlap (clash) between two unbonded atoms, or the geometrical similarity between molecules of identical chemical structure (e.g. Smart et al. 2012).

Restraints are incorporated into weighting functions along with the fit of the model to the experimental data. The task of the refinement program is find a suitable balance between optimising the atomic parameters to reproduce the experimental data, and generating a model that is "physically reasonable", as encoded by the restraints. In general, restraints are up-weighted at low resolution, as the crystallographic data become less numerous and/or informative, and down-weighted at higher resolution, where the data-parameter ratio is much higher (Chapter 12, Rupp 2010).

Model validation R-values

The agreement between the refined model and the experimental amplitudes is measured by the crystallographic R-factor, defined as

$$R = \frac{\sum \left| |\mathrm{F_{obs}}| - |\mathrm{F_{calc}}| \right|}{\sum |\mathrm{F_{obs}}|}, \qquad 1.3$$

where the amplitudes $\mathrm{F_{obs}}$ and $\mathrm{F_{calc}}$ are the experimental data and calculated from the crystallographic model respectively, after $\mathrm{F_{obs}}$ have been scaled to $\mathrm{F_{calc}}$. The R-factor

provides a percentage difference between the model and the data. However, whilst it describes the fit of the model to the data, it cannot measure overfitting. Instead, it is customary to remove a randomly chosen set of experimental amplitudes (a *free* set) from the data, and not to use these in refinement; these data subsequently provide an orthogonal measure of how well modelling is being performed (Brunger 1992).

By default, the *free* set constitutes approximately 5% of the experimental data (in CCP4 & PHENIX). The R-values, calculated for the *free* and the remaining *work* set, become the $\mathrm{R}_{\mathrm{free}}$ and the $\mathrm{R}_{\mathrm{work}}$ respectively. The $\mathrm{R}_{\mathrm{free}}$ measures the *true* agreement between the model and the data, and the difference between $\mathrm{R}_{\mathrm{free}}$ and $\mathrm{R}_{\mathrm{work}}$ measures the degree of overfitting for the model. Use of these metrics allows a monitoring of the degree of overfitting during modelling and refinement (Kleywegt & Brünger 1996).

1.3.6 Crystallographic modelling: the iterative modelling paradigm

After modelling of the model-able density and several cycles of refinement, a (hopefully) more-complete and optimised model of the protein is now available. From our refined model, we generate a new set of phases, which are used to generate a new set of $2m\mathrm{F}_{\mathrm{o}}$-$\mathrm{DF}_{\mathrm{c}}$ and $m\mathrm{F}_{\mathrm{o}}$-DF_{c} structure factors (Appendix A.3), from which we generate new maps that can be used for modelling. As further atoms are added – presuming these are largely correct – and wrongly-placed atoms are removed, the phases will improve and lead to better density for both modelled *and unmodelled* parts of the structure.

This *iterative-modelling* paradigm forms the core of the crystallographic model-building process, where sequential rounds of model-building and refinement are performed which iteratively improve the phase estimate; this process is assumed to converge on a

complete model of the crystal. However, many features of the data remain unmodelled at the end of refinement, even for "near-convergence" phases, resulting in the so-called R-factor gap (Holton et al. 2014). Unmodelled features are particularly prevalent in the poorly-ordered solvent-filled channels in the protein structures, where building of an atomic model is difficult; such unexplainable features represent the limit of current modelling techniques, leading to the aphorism that "refinement against [and modelling of] high-resolution data is never finished, only abandoned" (Sheldrick 2007).

1.3.7 Modelling of discrete heterogeneity in the crystal

Much of the difficulty in modelling during the late stages of refinement arises in poorly ordered regions of the crystal that cannot be represented by a single atomic model. The crystallographic experiment measures the average electron density across all the different unit cells that comprise the crystal. If the contents of these unit cells are different, then the obtained diffraction data, and subsequently derived electron density, is an average across all crystallographic unit cells.

For small continuous disorder between unit cells, the data remain well represented by a single atomic model, with disorder modelled by an isotropic B-factor (or an anisotropic B-factor at high resolution) – though for large displacements, B-factor refinement may not converge correctly (Kuzmanic et al. 2014). For large, distinct conformational changes, however, modelling requires the explicit declaration of alternate conformations, with appropriate occupancies attached to each conformer (Figure 1.3).

In the case shown in Figure 1.3, the density for the alternate conformations of the arginine sidechain is interpretable. However, where a sidechain adopts a large number of alternate conformations in the crystal, or where alternate conformations of the

sidechain overlap, the density superposition can be uninterpretable – often described as “noisy” or “weak” on online fora such as *ccp4bb*. Phenomena such as these lead to ambiguity in crystallographic modelling.

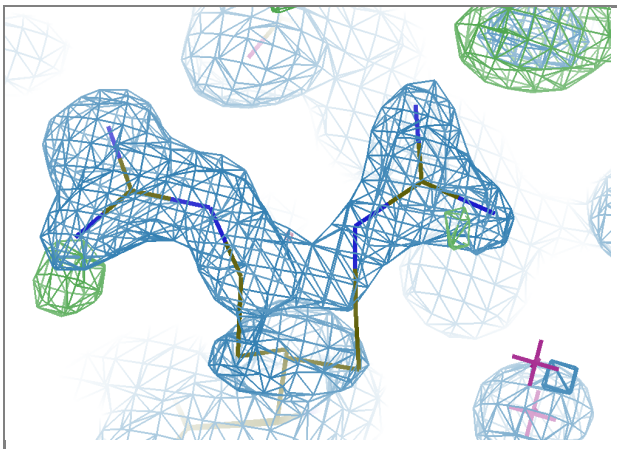


Figure 1.3. ***Multiple superposed models are required to model discrete heterogeneity in the crystal.*** *When heterogeneity is present in the crystal that cannot be modelled through thermal disorder, multiple models for the different conformations must be built. Shown is an arginine that is present in at least two conformations in the crystal (of which two are modelled). Heterogeneity is modelled using alternate conformations; each is labelled by a conformer ID (e.g. A and B) and has an associated occupancy. For a “complete” model, the occupancies of the different conformers are constrained to sum to unity. Image: 1.34Å map from dataset x443, protein JMJD2D (lysine demethylase 4D; see Chapter 4 for details). The $2mF_o$-DF_c map (blue; 0.9σ) and mF_o-DF_c map (green/red; ±3σ) are shown at typical modelling contour levels.*

1.3.8 Ambiguity, interpretation and noise in crystallography

Ambiguity is present in crystallography because information is lost when averaging over multiple states; unambiguous interpretation of crystallographic electron density is generally not possible where multiple states are present. We therefore rely on crystallographers to *interpret* the electron density: it the task of the crystallographer to identify and model all the states that are present. In regions of poor crystallographic order or low occupancy, the crystallographer can have a large influence on the model that is produced, leading to incorrect models (e.g. Stanfield et al. 2016) and phenomena such as the “ligand of desire” (Pozharski et al. 2013).

Ambiguity in modelling due to averaging over the crystal is further exacerbated by the presence of noise in the crystallographic density due to uncertainty in the

crystallographic diffraction data and phase bias, but it is failures in modelling that principally lead to the R-factor gap (Holton et al. 2014).

The choice of map contour can particularly affect the model that is constructed; contouring to a low level reveals "noisy" density which appears uninterpretable. To avoid these uninterpretable regions, a conservative modelling approach is followed as "best practice" in crystallography. However, this results in an under-representation of heterogeneity in crystallographic models: the 1σ level has been shown to be consistently significantly above the noise level of the electron density (Lang et al. 2014; Lang et al. 2010). The use of conservative contour levels results in signal – e.g. below the 1σ level – remaining unmodelled.

Furthermore, the term "sigma-scaled maps" is a misnomer, since this implies some level of statistical significance; sigma-scaled maps should correctly be called "rmsd-scaled" maps, as this is the mathematical operation being performed. The term "sigma-scaled" is particularly inappropriate for $2\mathrm{mF_o}$-$\mathrm{DF_c}$ maps, where the map values are not normally-distributed (Figure 1.4), and the "1σ" level does not correspond to a measure of statistical significance. The use of the term is thoroughly misleading.

However, the use of the sigma-scaled maps and the use of standard σ-levels persists in the culture of crystallography; accordingly, the contour levels are reported as sigma-levels for maps throughout this thesis.

New approaches are now being developed to rigorously identify and model conventionally "weak" crystallographic signal, including: methods to determine the true noise level in maps, so that appropriate contour levels may be chosen (Lang et al. 2014); sensitive methods to identify the presence of rotameric heterogeneity (Lang et al.

2010); automated multi-conformer modelling protocols (Van Den Bedem et al. 2009; Keedy et al. 2015); and the use of molecular dynamics to generate hundreds of independent superposed models of the protein to describe the heterogeneity of the crystal (Burnley et al. 2012).

However, none of these methods are applicable where density results from the superposition of multiple chemical species, such as the superposition between a ligand and a solvent molecule, which are the cases this thesis explores: partial-occupancy ligand binding.

Example histogram of $2mF_O$-DF_C map values

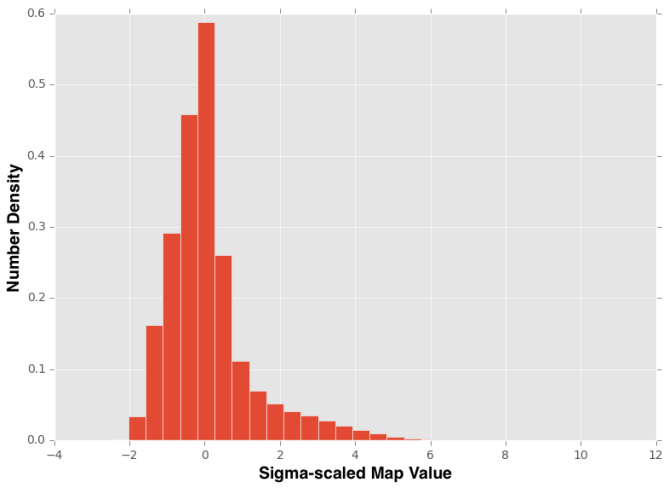


Figure 1.4. ***Map values in the $2mF_o$-DF_c map are not normally distributed****. The histogram of map values in a sigma-scaled $2mF_o$-DF_c map are skewed, and therefore not normally distributed. The sigma-levels associated with such maps therefore do not reflect the commonly-associated statistical meaning. Map distribution for dataset x434, map at 1.65Å, of protein BAZ2B (bromodomain adjacent to zinc finger 2b; see Chapter 4 for details).*

1.3.9 Model validation

One method by which we can attempt to cast an objective eye on the modelling process is through the extensive use of validation metrics. Application of these metrics allows us to check the quality of the modelling and refinement being performed.

Further to the qualitative validation that takes place during model-building, where large peaks in the $m\mathrm{F_o}$-$\mathrm{DF_c}$ maps are used to identify errors or omissions in the model, there exist a number of quantitative model-validation measures. These metrics can broadly

be divided into two categories: those which measure the agreement of the model with our knowledge of protein structure (knowledge-based metrics) and those which measure the agreement between the model and the experimental data (experimental metrics). Each category can be further sub-divided into global and local metrics, which either give a validation score for the whole or a part of a structure respectively.

An example global knowledge-based score is the RMS deviation from known bond lengths and angles over the model; models that deviate significantly from these known empirical values are unlikely to be correct. Local metrics includes measures of clashes between non-bonded atoms, and unlikely combination of phi-psi angles (Ramachandran outliers: sparsely populated or forbidden regions of the Ramachandran plot). These metrics are available in programs such as MolProbity (Chen et al. 2010) and integrated into modelling tools such as Coot (Emsley et al. 2010), but are not discussed further in this thesis.

Examples of global experimental metrics have been encountered in the previous section: the $\mathrm{R}_{\mathrm{free}}$ and $\mathrm{R}_{\mathrm{work}}$ metrics. These metrics are used to judge the overall quality of the model, but as they are global metrics they cannot identify specific regions of the structure that need correcting. More precise errors can be identified by comparing (groups of) modelled atoms to the electron density in real space.

1.3.10 Electron density quality indicators

With issues over the interpretation of electron density by crystallographers (section 1.3.8), validation of the constructed models against the experimental density is crucial for the identification of modelling errors. The electron density metrics described below are used extensively throughout this thesis for the validation of ligand models.

Scoring against the electron density requires the simulation of electron density (at the relevant resolution) for the atomic model and the comparison of this density to the experimental electron density. The density for atoms is modelled as a simple three-dimensional Gaussian, factoring in the atomic number of the element, and the appropriate blurring of the density by the B-factors (Tickle 2012).

When comparing the density of the model to the experimental data, each part of the experimental density must be assigned to an atom: for a superposition of two states, the observed density must be ascribed, in proportion, to all the atoms that could contribute to it. These considerations are taken into account by EDSTATS (Tickle 2012), which calculates a variety of measures, both new and conventional; the correct treatment of the electron density enables the assessment of superposed model states.

Conventional real-space electron density metrics

The two most commonly used electron density validation metrics are the real-space correlation coefficient (RSCC) and the real-space R-factor (RSR). The RSCC is the Pearson product-moment sample correlation coefficient for two electron densities samples, ρ, calculated as

$$\mathrm{RSCC}(\rho_1, \rho_2) = \frac{\sum_{\boldsymbol{x}}(\rho_1(\boldsymbol{x}) - \overline{\rho_1(\boldsymbol{x})})(\rho_2(\boldsymbol{x}) - \overline{\rho_2(\boldsymbol{x})})}{\sqrt{\sum_{\boldsymbol{x}}\left(\rho_1(\boldsymbol{x}) - \overline{\rho_1(\boldsymbol{x})}\right)^2 \sum_{\boldsymbol{x}}\left(\rho_2(\boldsymbol{x}) - \overline{\rho_2(\boldsymbol{x})}\right)^2}}, \qquad 1.4$$

where all sums are over the same set of points, $\boldsymbol{x}$, and $\overline{\rho(\boldsymbol{x})}$ is the average value of $\rho(\boldsymbol{x})$ over these points. The RSR (Bränden & Alwyn Jones 1990; Jones et al. 1991) also measures the agreement between the model density and the experimental density. It is calculated as

$$\mathrm{RSR} = \frac{\sum|\rho_1(x) - \rho_2(x)|}{\sum|\rho_1(\mathrm{x}) + \rho_2(x)|}. \qquad 1.5$$

Since the RSR requires the subtraction of the two densities, the densities are required to be scaled correctly to one another; the RSR thus applies best after refinement, where the model density is naturally on the same scale as the observed density (Tickle 2012).

The RSCC, however, has no requirement for the model density to be scaled to the experimental data, and can thus be used for the quick validation of any atomic model, regardless of whether it has been refined. Additionally, as the RSCC measures the similarity in the shapes of electron densities, it can be calculated between the model and the *difference* (mF_o-DF_c) density, making it useful for the automated modelling of unmodelled features, such as ligands (see section 1.4.5).

Improved real-space electron density metrics

The flaws with conventional metrics are discussed at length in Tickle (2012); conventional metrics do not measure either the accuracy or the precision of a model (Figure 1.5), but rather a combination of the two. The RSCC and RSR both possess further unfavourable traits, such as correlation with the model B-factors (Tickle 2012).

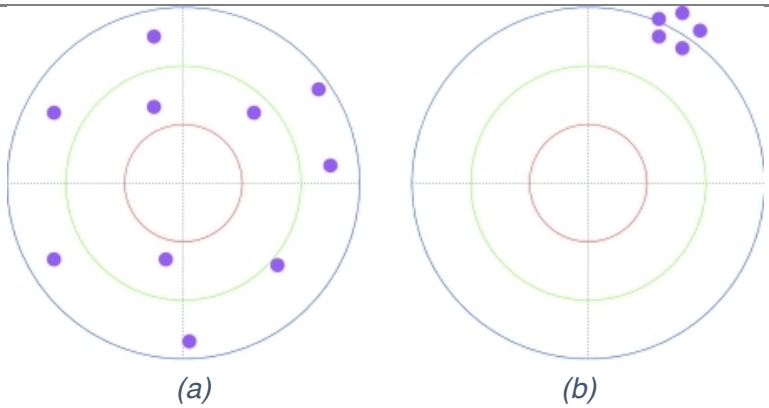


Figure 1.5. ***Simple demonstration of accuracy and precision.*** *(a) The points are centred on the centre of the circle (they are accurate) but they are highly variable (they are imprecise). (b) The points are tightly grouped (they are precise), but they are not centred on the circle (they are inaccurate). Image reproduced from Tickle (2012).*

Tickle (2012) introduced two sets of new metrics for the validation of crystallographic structures: the real-space Z-difference (RSZD) and the real-space Z-observed (RSZO) scores (described in detail in Appendix A.4). These metrics, in contrast to existing metrics, are designed to measure distinct aspects of the model: RSZD measures the significance of difference density over the model (model accuracy), and RSZO measures the diffuseness of the electron density for a model (model precision). Both scores assume the convergence of the crystallographic phases; they are, therefore, useful only towards the end of the modelling and refinement process, when a nearly complete and largely correct model is available.

1.3.11 Model-based quality indicators

There are several further quality indicators that highlight features in a model that are unlikely to occur. One suspicious feature is for atoms to display much larger B-factors than surrounding atoms: The B-factor of an atom is a measure of its mobility and it is unphysical for an atom to have a *much* larger mobility than a nearby atom. As such, atoms identified as having larger B-factors than surrounding atoms are often poorly modelled or poorly refined. We can thus define a *B-factor ratio*, calculated as

$$B_{ratio} = \frac{mean(\boldsymbol{B_{residue}})}{mean(\boldsymbol{B_{surroundings}})}, \qquad 1.6$$

where $\boldsymbol{B_{residue}}$ are the B-factors of the residue in question, and $\boldsymbol{B_{surroundings}}$ are the B-factors of a set of appropriately-selected nearby atoms. For the calculation of B_{ratio} for a non-covalently bound molecule, an appropriate choice of surroundings is the sidechains of all residues within 4Å of the molecule (Stanfield et al. 2016).

When large B-factor ratios are present, it is for one of two reasons. It can be indicative of a non-present model: when no density for an atom is present, the refinement

program will inflate the B-factors to blur the model density so that its contribution to the scattering is minimal. The second reason is that the occupancy of a model is incorrect: when the occupancy is too high, refinement once again increases the B-factors to compensate for the lack of density. The permissibility of large B-factor ratios can still result in intense debate regarding the presence/absence of crystallographic features (Stanfield et al. 2016, and the associated correspondence).

1.4 Protein-ligand complex determination

As discussed in sections 1.1 and 1.2, it is frequently desirable to determine the structure of a protein bound to a small molecule – a ligand. To determine a ligand-bound structure, the compound of interest must be introduced into a crystal, and diffraction experiments performed on that crystal. The structure for the protein must then be solved; if the compound binds, it will be present in the electron density at the location of binding. The compound must then be modelled, refined and validated.

1.4.1 Methods for introducing compounds into crystals

Two of the most common methods of introducing a compound into a crystal are co-crystallisation and soaking (Hassell et al. 2006). For co-crystallisation, the compound is introduced into the crystallisation conditions with the solubilised protein, and for soaking, compounds are added to preformed crystals.

The success of ligand-binding experiments relies on several aspects of the protein and the crystal form; for example, where the binding of a compound requires a conformational change in the protein, the crystal must be able to tolerate this change for the structure to be determined. Conversely, since the unbound state of the protein

may be less stable and more disordered than any bound complex of the protein, crystallisation may rely on such a conformational change taking place.

Other crystallographic considerations include that sites of interest on the protein surface may be blocked by crystal contacts, prohibiting binding at these locations; optimisation and generation of different crystal forms is one method for overcoming this (e.g. Price II et al. 2009). Equally, binding sites may be created between symmetry-related proteins, which are solely a function of the crystal lattice, and are thus non-biological.

In soaking experiments, compounds must be able to penetrate the crystal lattice. This requires the presence of solvent channels – connected regions of solvent that permeate the crystal – large enough to accommodate the movement of compounds; soaking experiments are only possible because the solvent content in macromolecular crystals is typically ~40% or higher (Matthews 1968).

1.4.2 Structure solution for a known crystal form

In crystallographic ligand-binding experiments, the crystal structure of the unbound protein will frequently have been previously determined; this model can be used to generate the phases for Molecular Replacement (MR). From co-crystallisation and soaking experiments, and soaking experiments in particular, the ligand-bound crystals are often the same *crystal form* – arrangement of the protein molecules in the crystal – as the originally-determined protein structure. For such data, the phases between datasets are so similar that Molecular Replacement of the protein model can rather be called “Molecular Substitution”; subsequent crystal structures can be solved by simply

refining the original model against the new diffraction data. Computational pipelines are available to perform these steps, e.g. Dimple (CCP4; Winn et al. 2011).

Most of the structure of the protein does not change between crystals of the same crystal form; the only modelling required is that of new features of the derivative crystal (such as binding ligands), the re-modelling of protein atom affected by these changes (such as sidechain reordering upon ligand binding), and occasional re-modelling of small stochastic differences between crystals.

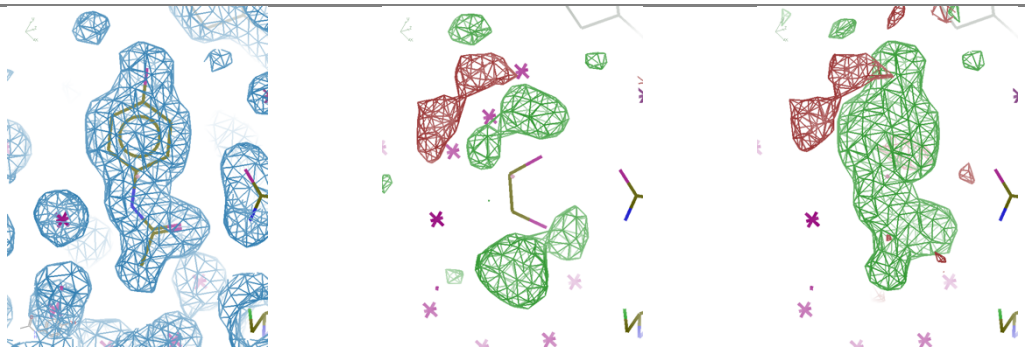
1.4.3 Ligand identification in derivative crystallographic datasets

Where the structure of a known protein has been solved by Molecular Replacement or Molecular Substitution, methods may be subsequently applied to detect electron density for any bound ligands. There are two methods used identifying electron density features: difference map methods and isomorphous-difference methods.

Difference map methods

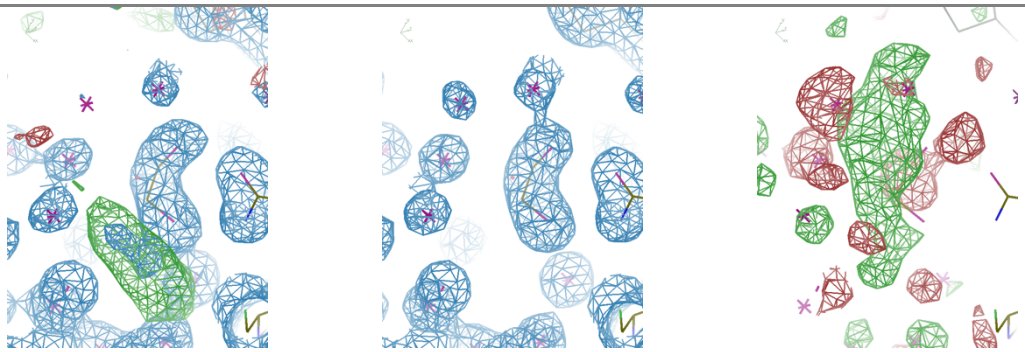
Difference map methods detect unmodelled features in crystallographic data; $m\mathrm{F_o}$-$\mathrm{DF_c}$ maps are analysed, since these minimise the error caused by incorrect phases. Difference map methods contour the $m\mathrm{F_o}$-$\mathrm{DF_c}$ map at a particular value and identify connected regions (“blobs”).

Maps can be contoured at fixed values, as in Dimple (CCP4; Winn et al. 2011), or dynamically, using more advanced “fragmentation-tree” methods (Langer et al. 2012). Regions of positive difference density may represent unmodelled features, such as a bound ligand. An example for a high-occupancy bound ligand is shown in Figure 1.6.



(a) (b) (c)

Figure 1.6. ***Identification of a ligand using difference density.*** *Different electron density maps for a strongly bound ligand (paracetamol) in dataset x575 of BAZ2B at 1.83Å (more details of the crystals of this protein can be found in Chapter 4), bound at approximately 95% occupancy. (a) The fitted $2mF_o$-DF_c map (blue; 1σ) showing the modelled pose of the bound ligand. The density was generated after refinement of a reference model obtained from a high-resolution unbound dataset against the new diffraction data. The modelled ligand pose is shown for comparison with the following images. (b) mF_o-DF_c density (green/red; ±3σ) which corresponds to the $2mF_o$-DF_c density in (a). The refined reference model is now shown, with an ethylene glycol and three waters modelled in the binding site; these bind when no ligand is present in the crystal. These molecules overlap with the ligand density and reduce the signal in the difference map; difference maps are sensitive to the placement of solvent molecules. (c) The solvent-model-sensitivity characteristic of difference maps leads to the practice of removing modelled solvent from known binding sites. The refined mF_o-DF_c density (green/red; ±3σ) with the described solvent molecules removed results in clear difference density for the ligand, as the ligand binds at high occupancy.*



(a) (b) (c)

Figure 1.7. ***Identification of a ligand using isomorphous difference maps.*** *Different maps for BAZ2B dataset x425, resolution 1.72Å, and x575 (as in Figure 1.6). (a) The $2mF_o$-DF_c (blue; 1σ) and mF_o-DF_c (green/red; ±3σ) maps for an unbound dataset, x425. The model is not perfect: residual ambiguous difference density remains. (b) The F_o map (blue; 1σ) for dataset x425, created without weightings on the diffraction amplitudes, is used for the isomorphous difference map. Contoured at the same level as (a), the density for the unmodelled feature (creating the difference density) in (a) cannot be seen. (c) The isomorphous difference map of dataset x425 and x575 (green/red; ±3σ), calculated with phenix. The presence of the ligand is clearly indicated, though the density for the sections of the ligand that overlap with reference solvent are removed in the subtraction; the density does not completely resemble the ligand.*

Difference map methods are naturally sensitive to the model used in refinement. The presence of solvent molecules in the model may obscure binding ligands, even where their placement was correct in the reference model used for refinement (Figure 1.6b). Stripping of water and other solvent molecules from the model is used to prevent this, but the phases will be degraded by the removal of correct atoms from the model, potentially decreasing the quality of the maps.

Complex decision-making pipelines have been developed to model the crystal conservatively, so that solvent molecules are not placed into possible ligand density (used in Schiebel, Krimmer, et al. 2016; method unpublished), but this conservative approach still leads to the removal of correct atoms from the model.

Where solvent molecules are removed from the model, refinement will fill the region with bulk solvent; "polder" maps have been developed to prevent this bulk solvent masking possible ligand density (phenix.polder, part of PHENIX; Adams et al. 2010).

Isomorphous difference methods

Isomorphous difference maps provide an objective measure of signal in a crystal, by contrasting putative ligand-bound datasets against a reference (typically unbound) dataset. Density in an isomorphous difference map reflects a difference between the datasets; where the only systematic difference is the addition of a ligand, regions with significant differences may indicate a binding ligand. An isomorphous difference map is calculated using structure factors

$$\mathrm{F_{iso}} = \left(|\mathrm{F_{obs}^{new}}| - \left|\mathrm{F_{obs}^{ref}}\right|\right) \cdot e^{i\phi_{calc}^{ref}}, \qquad 1.7$$

where $\mathrm{F_{obs}^{new}}$ are the amplitudes of the new ligand-bound datasets, and $\mathrm{F_{obs}^{ref}}$ and $\phi_{\mathrm{calc}}^{\mathrm{ref}}$ are the amplitudes of the reference dataset and $\phi_{\mathrm{calc}}^{\mathrm{ref}}$ are the phases from the refined

model of the reference dataset. An advantage of the isomorphous difference map is that it is fast to calculate: isomorphous difference maps do not require crystallographic refinement as the phases from the reference dataset are used. However, with the increase in computational power in recent years, speed is less of a concern.

Isomorphous difference maps have been seemingly forgotten in recent years (Rould & Carter 2003), potentially due to the development of improved maps, such as the composite $2mF_o$-DF_c and mF_o-DF_c maps (Read 1986), which overcome many of the problems that faced crystallographers at the time of their initial development.

Furthermore, with the requirement of strict isomorphism for the Fourier terms to align, the usage of isomorphous difference maps is limited in cases where the unit cell varies between crystal: isomorphous difference maps require strict isomorphism (Rould & Carter 2003). Isomorphous difference maps are not extensively deployed to identify ligands in fragment screening experiments.

1.4.4 Interpretation & validation in ligand modelling

In a continuation of section 1.3.8, ligand modelling remains an interpretative art: the presence/absence of a ligand is decided by a crystallographer. This human dependency leads to errors in models deposited in the PDB (Berman et al. 2000; Berman et al. 2003) by overzealous modellers (e.g. Stanfield et al. 2016). The missing ligands discussed in Stanfield *et al.* (2016) principally stem from disagreements about the use of the difference map to prove the presence of a ligand; some crystallographers believe it is appropriate to interpret the electron density when contoured at low values.

This problem is further exacerbated by phase bias, since it is possible for density to appear for features that are not truly present in the crystal; even composite maps

remain slightly biased towards the model used to generate the phases. “OMIT” maps (Bhat & Cohen 1984) are one method for testing the presence of ligands (Pozharski et al. 2013): the ligand is removed from the model, steps such as coordinate randomisation are applied to remove any “phase memory” from the rest of the structure, and the electron density maps are recalculated. The new maps are minimally biased towards the ligand, as all information from the ligand model has been removed from the phases. If the removed ligand is “real”, it is expected that difference density corresponding to the ligand will appear in the OMIT map; in the absence of further evidence – such as density from an isomorphous difference map – lack of OMIT density is used as evidence of an incorrectly modelled ligand (e.g. Stanfield et al. 2016).

Composite OMIT maps can be used to generate maps that are not biased by the placement of local atoms for the whole unit cell (Terwilliger et al. 2008); OMIT maps are calculated iteratively for all regions of the unit cell and subsequently pieced together to generate an OMIT map for the whole unit cell, where each local region is minimally biased towards the local model.

Modelled ligands are most often validated against the electron density using the RSCC to the refined electron density (Pozharski et al. 2013). A correlation of 0.7-0.8 or higher is generally considered to indicate good agreement between the model and the data.

1.4.5 Automated ligand-fitting methods

Once significant blobs of density have been identified, chiefly using difference map methods as described in the previous section, automated methods can be applied to generate possible models for the bound ligand in those regions.

Automated ligand-fitting methods generate potential conformations (decoy models) of ligands at the identified site, and subsequently score these to identify the best model; a model which scores well is assumed to be a good candidate for a correct model of a bound ligand. Three ligand-fitting programs from Openeye, PHENIX and Global Phasing are described below and used in this thesis; others are available from ARP/wARP (Zwart et al. 2004) and Coot (Emsley & Cowtan 2004), but were not used in this thesis due to a combination of technical issues and time constraints.

FLYNN from Openeye

FLYNN (Wlodek et al. 2006), part of the AFITT package from Openeye, identifies binding ligands by generating a series of low-energy conformations of the ligand in identified regions of electron density, scoring the models against the observed density, and optimising the highest-ranked models.

It is possible to perform the analysis on either $m\mathrm{F_o}$-$\mathrm{DF_c}$ maps or $2m\mathrm{F_o}$-$\mathrm{DF_c}$ maps: if a $2m\mathrm{F_o}$-$\mathrm{DF_c}$ map is provided, the model is masked from the map and blob-identification is performed on the remaining regions. Likely blobs are identified by iso-contouring the electron density map at various levels and finding connected volumes with approximately the same volume as the ligand.

Once likely blobs have been identified, a series of low energy conformers are generated using *Omega* (from Openeye). This quickly generates large numbers of potential decoys, which are aligned to the identified blobs by alignment of the moments of inertia. Poses are rigid-body optimised and ranked using the Tanimoto overlap between the observed electron density, ρ_{exp}, and the model electron density, ρ_{model}. The Tanimoto overlap is defined by

$$T_{overlap} = \frac{V_{shape}(\rho_{exp}, \rho_{model})}{V_{shape}(\rho_{exp}, \rho_{exp}) + V_{shape}(\rho_{model}, \rho_{model}) - V_{shape}(\rho_{exp}, \rho_{model})}, \quad 1.8$$

where

$$V_{shape}(\rho_1, \rho_2) = -\int_{model} \rho_1(x)\rho_2(x)dx \quad 1.9$$

is the "shape overlap" between the two densities, integrated over the model volume; V_{shape} is also the potential used in the rigid-body optimisation stage.

The top five to ten ranked models are then optimised to balance the internal energetic strain on the ligand and the fit to the electron density; a strained, high-energy conformation is unlikely to constitute a *real* bound pose. Ligand poses are refined to minimise the potential

$$V_{total} = V_{forcefield} + \lambda V_{shape}, \quad 1.10$$

where $V_{forcefield}$ and V_{shape} account for the energy of the ligand conformations and the fit to the electron density, respectively. The λ parameter controls the balance between the energy of the ligand pose and the fit to the electron density. The fitting procedure is applied adiabatically; the λ parameter is incrementally increased from zero to one, with minimisation at each step against the new potential, V_{total}.

The $V_{forcefield}$ term is given by the MMFF94 forcefield (Halgren 1996), and the V_{shape} term is as defined in equation 1.9. The fitting progress is stopped when the Tanimoto coefficient (equation 1.8) reaches a maximum, or plateaus, as function of λ.

Ligandfit from PHENIX

Ligandfit, from the PHENIX consortium (Terwilliger et al. 2006; Adams et al. 2010), breaks the ligand into rigid fragments, and then attempts to identify a "core fragment"

in the electron density. Once this "core" fragment is fitted, connecting fragments are fitted iteratively until the whole molecule is reconstructed.

The decomposition of the ligand into rigid units is performed such that each fragment is internally connected by only non-rotatable bonds, and is only connected to other fragments by a rotatable bond. The fitting is performed in the largest contiguous blob of density that matches the size of the ligand; the iso-contour level is chosen as the highest level that provides a connected region of this size in the mF_o-DF_c map.

The density around the identified region is extracted to form a small "pseudo-map". Each core fragment of the ligand is placed at the origin of the "pseudo-map" in different conformations and an FFT-based convolution method is applied to determine whether the ligand pose matches an area of the density. Typically, the 300 models with the highest overlap are refined into the observed density, and the top 100 of these are saved, as ranked by the RSCC of the model to the observed density.

From each of the core fragments identified, the ligand is reconstructed iteratively. The conformation of a connecting fragment is sampled at 20^o intervals around the joining rotatable bond, and the resulting model is scored – including the already-fitted fragment. The model scores are calculated as

$$Q = \frac{N \sum_i Z_i \rho_i}{\sum_i Z_i}, \qquad 1.11$$

where Z_i is the atomic number of atom i, ρ_i is the density sampled at the position of atom i, and N is the total number of atoms placed. This score generally increases with the number of atoms placed, as long as they are placed into positive density – this is an intentional feature of the score. The output for the programs is the top-five models, as ranked by this score.

RHOFIT from Global Phasing

RHOFIT, from the Global Phasing consortium (Womack et al. 2010), generates likely ligand models by simultaneously optimising the ligand geometry and the fit of the model to the density, similarly to the other methods. However, it differs from the other methods in that it also contains terms which score the interactions made between the ligand model and the protein. Once more, blobs are identified in the difference density map as connected regions that are approximately the size of the ligand. Further details are not available; the method has not been published.

Summary of automated ligand-fitting approaches

All the ligand-fitting methods are based around the same paradigm, where methods identify regions of difference density and generate models in these regions; compare the generated models to the density; and present high-scoring models to the user.

Knowledge-based considerations in the construction of ligand models

Though this thesis focusses on the identification and modelling of ligands based on the primary experimental data – the electron density obtained from diffraction experiments – other sources of information may be either explicitly or implicitly used when constructing a ligand model (e.g. Mooij et al. 2006).

As with the model validation metrics discussed in section 1.3.11, ligand models that produce atomic clashes with the surrounding model are unlikely to be correct. Conversely, analysis of the molecular interactions between a ligand model and the surrounding atoms can be used to inform on the likelihood that the model of the ligand is correct: analysis and optimisation of hydrogen bonding patterns – involving the

consideration of alternate ligand charge states and tautomers – can help to differentiate between multiple similar bound poses of a ligand.

1.4.6 Ligand identification in primary crystallographic fragment screening

In this thesis, I am chiefly interested in the identification of binding ligands in datasets arising from primary crystallographic screening experiments, which are now routinely carried out at Diamond Light Source, beamline i04-1.

In primary crystallographic screens, 10s to 1000s of crystallographic datasets are collected, with different compounds present in each crystal. Binding ligands will be present in the crystallographic density; the task is to detect and model these binding events, so that the atomic models can be used to inform compound elaboration.

The binding affinities of fragments are weak, so the occupancy in the crystal may be correspondingly weak, making ligand identification difficult. Companies such as Astex have developed internal pipelines to process crystallographic data in SBLD (Mooij et al. 2006), including fragment screening, but these pipelines are not publically available.

New approaches continue to focus on the presence of signal in the mF_o-DF_c map for the identification of binding fragments (Schiebel, Krimmer, et al. 2016). Through extensive refinement of the structure, Schiebel *et al.* decrease the noise in the difference maps and enable the detection of weakly-bound compounds. However, the interpretation of the difference maps is a labour-intensive and subjective task, and required eight crystallographers to analyse 364 datasets in the case of Schiebel *et al.* (2016). Moreover, frequently only part of the molecule was identified in the electron density.

Chapter 2

Testing current methods for ligand identification and modelling

---

"One does not simply fit a ligand into electron density"

Ligand identification and modelling is a common task in crystallographic structure-determination experiments. However, this thesis is particularly concerned with the identification of binding ligands in crystallographic fragment screening datasets; in such experiments, hundreds of datasets are collected, each containing a different ligand that may or may not bind to the protein.

The large number of datasets and the potential for ligands to bind anywhere on the protein surface prohibits the manual inspection of all datasets as a routine approach (as e.g. 500 crystallographic datasets per day may now be collected at Diamond beamline I04-1). Automated identification of ligands in crystallographic datasets, however, enables large numbers of datasets to be analysed objectively, without human intervention and, most importantly, without human subjectivity – avoiding situations such as that discussed in Stanfield et al. (2016).

Currently, the most common approach to automated ligand identification is to identify "suitable blobs" in a crystallographic difference (mF_o-DF_c) map and generate ligand conformations which match the density; density validation and chemical strain metrics are then used to rank the generated models (e.g. Mooij et al. 2006; Wlodek et al. 2006;

Womack et al. 2010; Emsley et al. 2010; Carolan & Lamzin 2014). It is assumed that a large amount of difference density and/or a high-scoring model implies a truly bound ligand, and these are presented to the user as likely hits; it is then up to the user to visually inspect the electron density and decide whether a ligand is truly bound.

I am unaware of impartial comparisons of fitting pipeline outputs in the literature to determine which – if any – are the best programs to use in different situations, such as crystallographic fragment screening. Furthermore, the success rates of the ligand-fitting programs are normally only tested as a function of the resolution of the dataset, and the size of the ligand (Terwilliger et al. 2006; Carolan & Lamzin 2014); new validation metrics now make it possible to measure the success as a function of the electron density quality of the ligand (Tickle 2012).

In this chapter, I characterise three currently-available ligand-fitting programs (FLYNN, Ligandfit & RHOFIT) and their ability to re-fit a series of crystallographic datasets containing bound ligands. I do this by generating a test set composed of ligand-bound structures, removing the bound ligands, re-refining the structures, and re-fitting the ligands. The re-fitted structures are then refined, the ligand models scored against the electron density, and compared to the reference model; this tests the ability of the ligand-fitting programs to identify the correct site for modelling, their ability to generate a good model for the ligand, and our ability to detect the correctness of an individual model by its agreement with the electron density.

At the end of the chapter, I conclude that difference-map-based methods are fundamentally hindered by the need for bound ligands to exhibit clearly-defined difference density in the refined maps. Any partially-occupied or disordered ligand will,

by its nature, produce less difference density than a fully-occupied ligand. Furthermore, partial-occupancy bound-ligand density will be observed in a superposition of states with molecules corresponding to the unbound copies of the protein in the crystal. Our conclusion is that traditional difference-map-based approaches are not suitable, in general, for ligand identification or modelling in fragment-screening experiments, or for any experiment where bound ligands are present at sub-unitary occupancy.

2.1 Preparation of the ligand-fitting dataset

A ligand-fitting dataset was assembled by collating all ligand-bound crystallographic models in the database at the Structural Genomics Consortium (SGC), Oxford (as of 23rd July 2014). Like all crystallographic models, these ligand-bound structures are dependent on the subjective interpretation of a crystallographer to determine the presence and the final pose of the ligand; however, it has been shown that the models from structural genomics centres are amongst the best publically-available models in the PDB (Brown & Ramaswamy 2007). Specifically, the models at the SGC have been validated by two experienced crystallographers: the modelled ligands can, with reasonable certainty, be taken to be correct.

A summary of the identified ligands is shown in Table 2.1; full details are included in Appendix B. Ligands are not filtered by chemical similarity and we allow for multiple binding ligands per asymmetric unit to keep the test set as large as possible: the ligands bound to different non-crystallographic symmetry (NCS) copies of proteins present distinct electron density. The distributions of three parameters for the structures and ligands in the test set are shown in Figure 2.1.

Table 2.1. ***Summary statistics of the SGC ligand dataset.*** *Proteins may be present multiple times, bound to different ligands, but each protein-ligand combination is present only once (one crystallographic dataset of each combination). Different non-crystallographic symmetry (NCS) copies of ligands are considered separately as the electron density quality varies between such copies. The average number of ligands per unit cell is 1.55 (standard deviation 1.19). Full details are shown in Appendix B.*

Number of distinct proteins	61
Number of distinct protein-ligand combinations	187
Number of distinct ligand copies (including NCS copies)	290

2.1.1 Preparing the dataset for re-fitting

Once the dataset had been assembled, structures were prepared for analysis by removing the modelled ligands and re-refining the structure (the pipeline is shown schematically in Figure 2.2). For structures with multiple ligands, each ligand was removed separately to generate a series of structures (e.g. two bound ligands lead to two models, with one ligand removed in each).

In addition to the removal of the ligand, all water molecules were removed from the structures. Removal of waters simulates the "unknown (weak) binder, unknown location" situation in fragment screening; the presence of water molecules in the structure may mask the presence of bound ligands in the difference map if they are present at the binding location (e.g. Figure 1.6b). Removing the water molecules will also degrade the quality of the maps (Schiebel, Krimmer, et al. 2016), further testing the sensitivity of ligand identification.

Water removal also poses a further challenge to ligand-fitting programs to discriminate between ligand and non-ligand density blobs. Modelled non-water solvent molecules are not removed; we do not consider the case where a binding ligand displaces one of these molecules.

B-factors for the generated structures were then reset to 20Å^2, and the structure was subjected to five cycles of refinement with REFMAC5 (Murshudov et al. 2011). This produced a refined structure that we will refer to as the "apo" structure; these refined structures were used as input for the fitting programs.

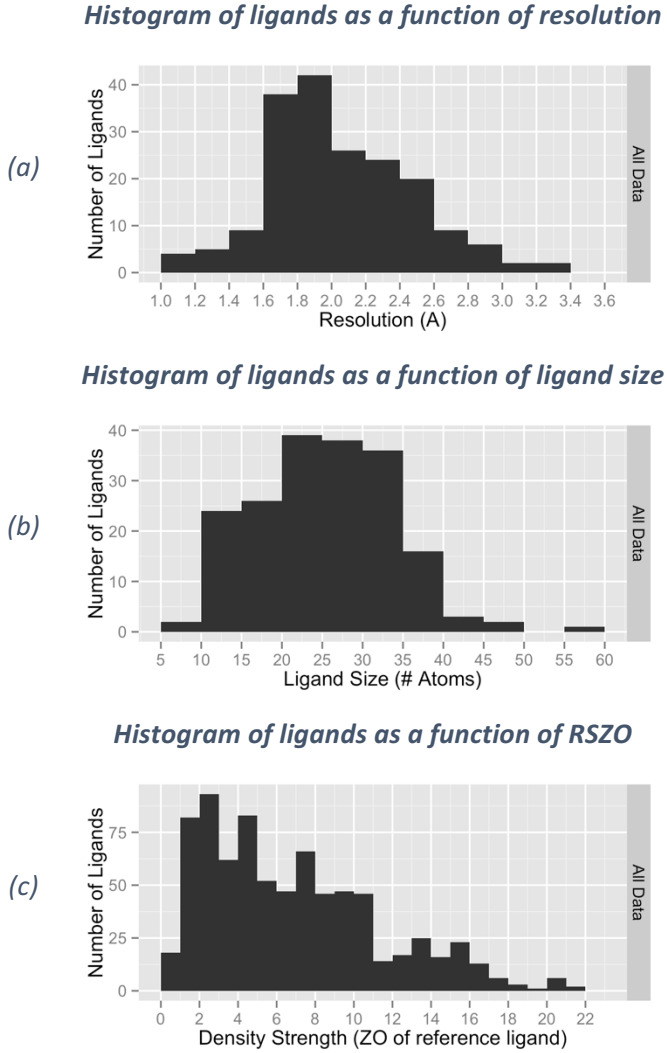


Figure 2.1. ***Statistics for the dataset of crystallographic ligands.*** *In all histograms, each ligand copy is counted separately; each crystallographic structure may be present multiple times where non-crystallographic symmetry is present. (a) Distribution of the resolutions of the datasets. The large peak in dataset numbers below 2Å is possibly due to a preference for sub-2Å crystals for ligand studies, causing optimisation of the crystal form to stop at this resolution. 82% of the ligand instances occur between high resolution limits of 1.6-2.8Å. (b) Histogram of ligand size, defined by number of non-hydrogen atoms. There is a lack of molecules with fewer than ten atoms, as the SGC focusses on the "chemical probe" and "biological molecule" region of chemical space, and the test set does not contain any fragment screening data: 91.8% of ligand instances are between 10-40 non-H atoms in size. (c) Distribution of the electron density strengths (RSZO) for the ligands in the dataset.*

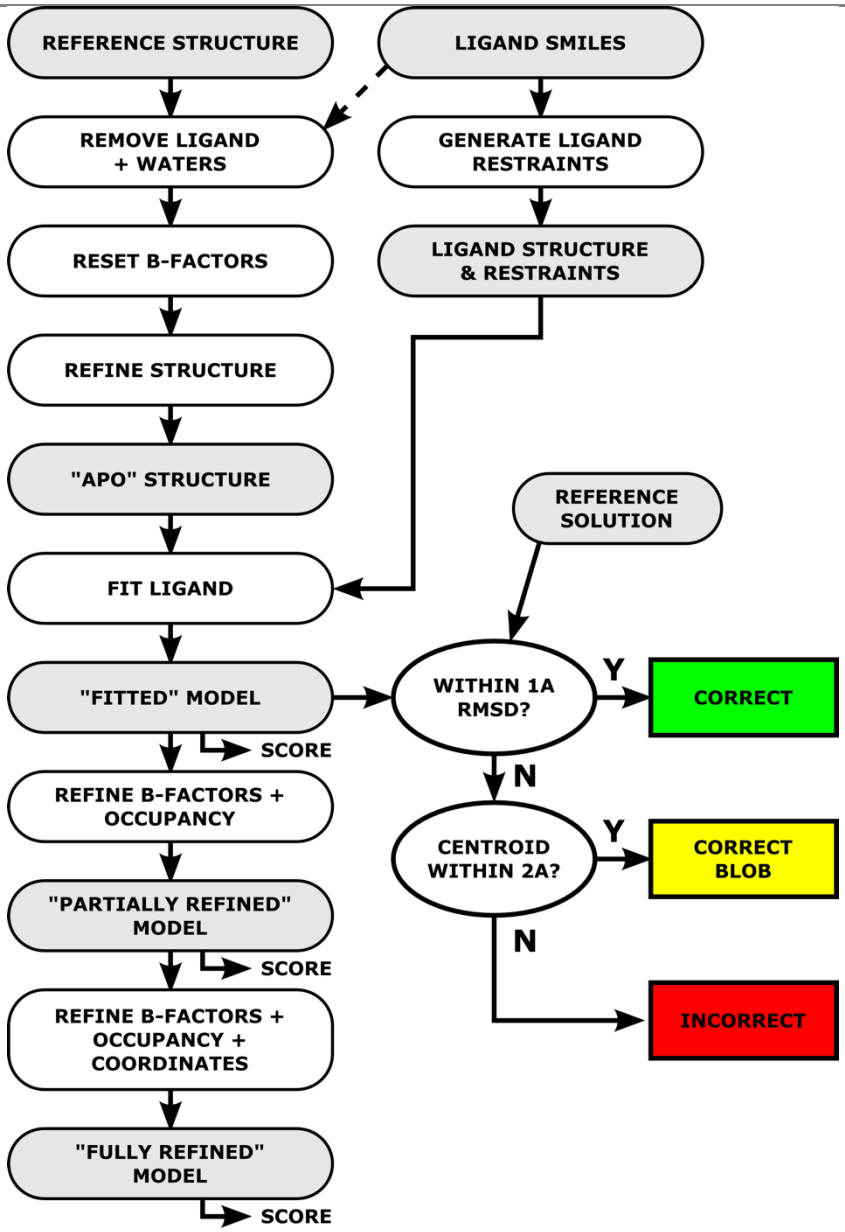


Figure 2.2. ***Schematic for the ligand re-fitting analysis pipeline.***

2.2 *Methods for re-fitting the removed ligands*

After the dataset was constructed, the ligand-fitting programs from three software packages were used to re-fit the ligands (Table 2.2). The modelling of a ligand requires a restraints file for the ligand, which defines the ideal bond lengths and angles for a ligand. Each software package contains its own ligand restraint-generation program; the restraints between programs are known to vary (personal communication with Paul Emsley), but this effect is assumed to be negligible, as we are only concerned with the

identification of the pose and the binding site of the ligand. The generation of an incorrect pose is assumed to dominate any errors in bond lengths and angles.

The smiles string for the ligand (extracted from the SGC database) was used to generate the ligand restraints with each of the ligand-restraints programs and the ligand-fitting program was used to refit the ligand to the density, using the standard settings. The approaches of each of the ligand-fitting programs are described in Chapter 1. From this process, we obtained a "fitted" model for each of the ligand-fitting programs.

Table 2.2. ***Tested ligand-fitting programs and the associated restraints generators.*** *Each software package utilises a different method of restraint generation. It was observed that the ligand-fitting programs threw errors less frequently when using their own associated restraint generation program. Although the restraints vary between the different software packages, the effects of the different restraints are assumed to be minimal compared to the errors in the modelling of the ligands. All datasets are refined with refmac, regardless of the software package used for fitting.*

Software Package	Restraints Generator	Fitting Program	Refinement
OpenEye (eyesopen.com)	WRITEDICT -	FLYNN (Wlodek et al. 2006)	REFMAC (Murshudov et al. 2011)
Phenix (phenix-online.org)	phenix.elbow (Moriarty et al. 2009)	phenix.ligandfit (Terwilliger et al. 2006)	
GlobalPhasing (globalphasing.com)	GRADE (Smart et al. 2011)	RHOFIT (Womack et al. 2010)	

2.3 Methods for refinement & analysis of the modelled ligands

After fitting, the ligand is refined in two stages. Firstly, the occupancy and B-factors of the ligand are refined for five cycles in REFMAC5 (Murshudov et al. 2011), whilst holding the coordinates fixed. Fitted ligands are modelled with fixed arbitrary B-factors and full occupancy, so this allows the profile for the fitted ligand to be refined, whilst retaining

the modelled pose; this generates a "partially refined" model. Secondly, full refinement of the ligand is performed for a further five cycles in REFMAC5, this time allowing the coordinates of the model to be variable, generating a "fully-refined" model.

Occupancy refinement of crystallographic ligands is often avoided to prevent over-fitting, since B-factors and occupancy are correlated in refinement (Bhat 1989). Furthermore, without the presence of a superposed alternate conformation a partial-occupancy ligand makes little physical sense as a model for the crystal (the rest of the crystal is effectively represented by a vacuum, or bulk solvent; discussed in Chapter 5). However, here we are interested in whether a low refined occupancy can be indicative of the correctness of a model – it is a proxy for how much refinement tries to "erase" the presence of the ligand from the model. We therefore ignore the physical implications of a partial-occupancy ligand without a superposed solvent model.

The "fitted", "partially refined", "fully-refined" and reference models are scored against the associated $2mF_o$-DF_c electron density from refinement with EDSTATS (Tickle 2012); density metrics are described in section 1.3.10. "Fitted" models are scored against the "apo" density from the initial REFMAC re-refinement of the reference structure.

Ligand-fitting programs typically generate multiple models for the ligand, and then use ranking to select the best model. The "fitted" model is simply the top-ranked model that the fitting program outputs.

We are interested in both the ability of the fitting programs to identify the correct site for the binding ligand, and to generate the correct model of the ligand. We classify the models of the fitting programs into three categories: a "correct" model, within 1Å RMSD of the reference model; a "correct blob" model where the centroid of the ligand is

within 2Å of the reference model; and an "incorrect" model, fulfilling neither of these conditions. A cutoff of 4Å on the model centroids was also used to assess whether the ligand had been placed in the correct blob, but this was found not to qualitatively change the results.

2.4 Results from re-fitting the ligand dataset

We analysed several aspects of the output models from the ligand-fitting programs, including: the failure rates and runtimes for each of the modelling programs (section 2.4.1); the frequency of the top-ranked model being correct (section 2.4.2); and whether a correct model is present in the generated decoys (2.4.3). We further investigated the extent to which electron density metrics can be used to determine the correctness of the model without need for human validation (section 2.4.4).

2.4.1 Program error rates and runtimes

The number of models for which each program output a model are shown in Table 2.3; RHOFIT models the most ligands and attains the highest success rate in modelling. The different programs display a variety of errors (Table 2.4-Table 2.6), but the fitting programs do not fail on the same ligands (Table 2.7).

Ligandfit fails most frequently, chiefly due to the inability of RDkit (used in our pipeline to analyse the constructed models) to read the ligands generated by phenix.elbow. These errors were not investigated thoroughly, but seemed to arise from what RDkit considers to be an invalid charge status based on the number of bonds between atoms; it is likely these could be rectified by assigning the correct charges to atoms in the ligand.

Table 2.3. ***Summary of the ligand re-fitting results.*** *Percentage of models is relative to the complete set of 290 models. The percentage of correctly-fitted models is calculated from the models where the individual program outputs a model.*

	FLYNN	Ligandfit	RHOFIT
Number of models (of possible 290)	**263** (91%)	**222** (77%)	**283** (98%)
Number of models correctly fitted	**137** (52%)	**120** (54%)	**173** (61%)

Table 2.4. ***Errors from the ligand fitting with FLYNN.***

Error Type	# Datasets
The model is incompatible with the FLYNN forcefield.	16
The model cannot be read by rdkit due to invalid ligand chemistry.	4
The model does not match the input smile string.	5
Refmac failed during refinement of the model.	2

Table 2.5. ***Errors from the ligand fitting with ligandfit.***

Error Type	# Datasets
No models produced by ligandfit.	3
The model cannot be read by rdkit due to invalid ligand chemistry.	50
The model does not match the input smile string.	10
No restraints generated by elbow.	5

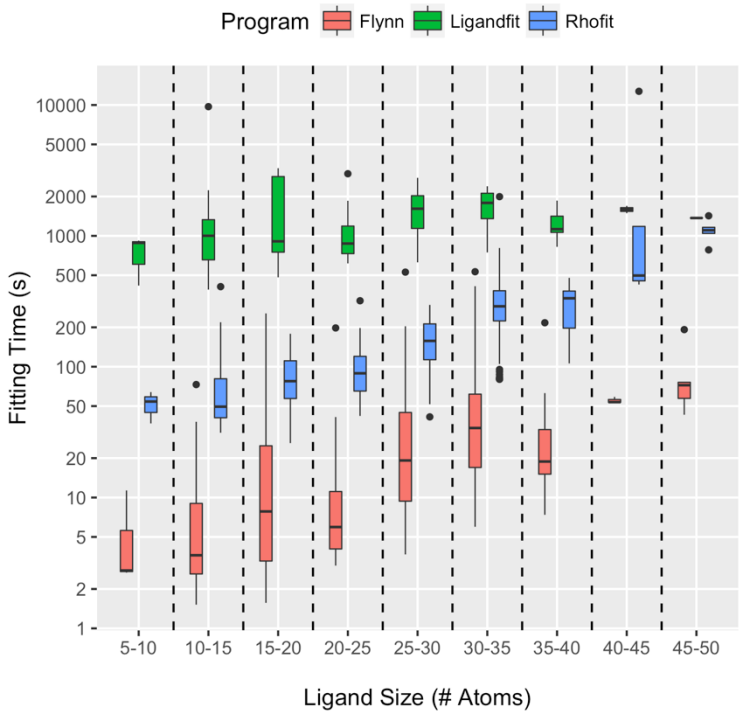
Table 2.6. ***Errors from the ligand fitting with RHOFIT.***

Error Type	# Datasets
No restraints generated by GRADE.	7

Table 2.7. ***Frequency of ligand-fitting programs failing on the same model****. Counts of whether the fitting program had succeeded or failed to generate a ligand model. This does not consider whether the generated model was correct. The two examples where all three programs failed are not included in the summary statistics in Table 2.1.*

FLYNN	Ligandfit	RHOFIT	# Datasets
SUCCESS	SUCCESS	SUCCESS	210
FAILURE	**FAILURE**	**FAILURE**	2
SUCCESS	**FAILURE**	**FAILURE**	7
FAILURE	SUCCESS	**FAILURE**	0
FAILURE	**FAILURE**	SUCCESS	15
SUCCESS	SUCCESS	**FAILURE**	0
SUCCESS	**FAILURE**	SUCCESS	46
FAILURE	SUCCESS	SUCCESS	12

The runtimes for the programs are shown in Figure 2.3; the contrast between the speed of the programs is significant (the y-axis is shown on a logarithmic scale). FLYNN is the fastest of the programs: FLYNN generates conformations systematically and uses fast ranking to discriminate, as opposed to ligandfit, which uses a computationally-intensive method (Fourier transforms) to place the first fragment of the ligand. Due to modern computational power, speed is no longer a primary concern.



(a)

Figure 2.3. ***The ligand-fitting programs exhibit large differences in runtimes****. Boxplots of the runtimes of the fitting programs; the y-axis is plotted as a logarithmic scale. FLYNN is the fastest program and ligandfit is the slowest.*

2.4.2 Top-ranked models

The overall rates for the top-ranked model being correct for FLYNN, ligandfit and RHOFIT are 52%, 54%, and 61%, respectively; the ligand-fitting success as a function of resolution and ligand size is shown in Figure 2.4. In agreement with the observations in the ligandfit paper (Terwilliger et al. 2006), ligand-fitting success is stable across the size

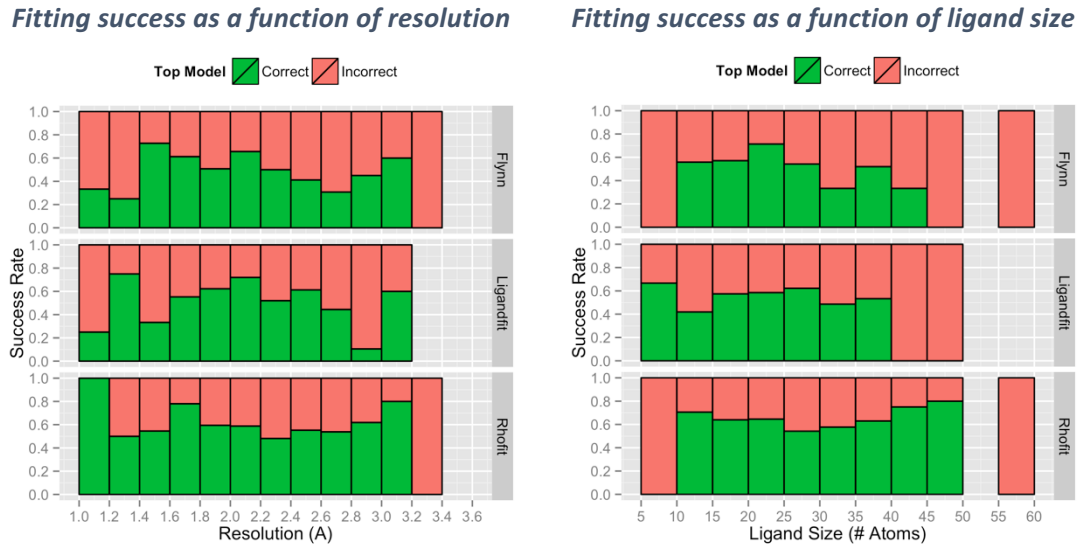
of ligand, as well as across resolutions (in the region 1.6-2.6Å, 10-40 atoms, where most of the data is contained). In the ligandfit paper, the authors also note that there is a decrease in performance for ligands with fewer than ten non-hydrogen atoms. We have too few models with ligands of this size to test this observation rigorously, as only three copies of two distinct ligands (two instances of one ligand are bound in one dataset) have fewer than ten atoms.

The ligand-fitting success as a function of electron density strength of the reference (SGC-modelled) ligand shows a distinct trend towards increased failure at lower electron density strengths (Figure 2.5). The overall success rates for FLYNN, ligandfit and RHOFIT of 52%, 54%, 61% decrease to 40%, 47%, and 44% for ligands with an RSZO of less than six. This decreases further to 36%, 36% and 33% respectively when the RSZO is less than three. This is to be expected, as weakly-bound ligands will exhibit less (or less well-defined) difference density, making both the ligand binding site and the pose of the ligand harder to identify.

Weak binding is an intrinsic problem for difference-density-based identification methods, upon which all three of these methods are based, limiting the utility of current fitting programs for the identification of binding for weakly-bound molecules.

2.4.3 Existence of a correct model in the generated decoys

We next investigate whether a correct ligand model was generated as one of the decoys, but was not the top-ranked model. This reveals whether the fitting programs are subject to ranking errors or search errors: A ranking error exists where an incorrect model scores more highly than a correct model, and a search error exists where a correct model is not generated as one of the decoys.



(a) (b)

Figure 2.4. ***Success rates for the top-ranked model from the three programs*** *as a function of (a) the resolution of the dataset, and (b) the size of the ligand (number of heavy atoms). There is no strong dependence of the success rate on the resolution of the dataset or the size of the ligand. From the distribution of datasets in Figure 2.1, most ligands are between 1.6-2.6Å resolution and 10-40 atoms, making it difficult to draw conclusions outside of these regions.*

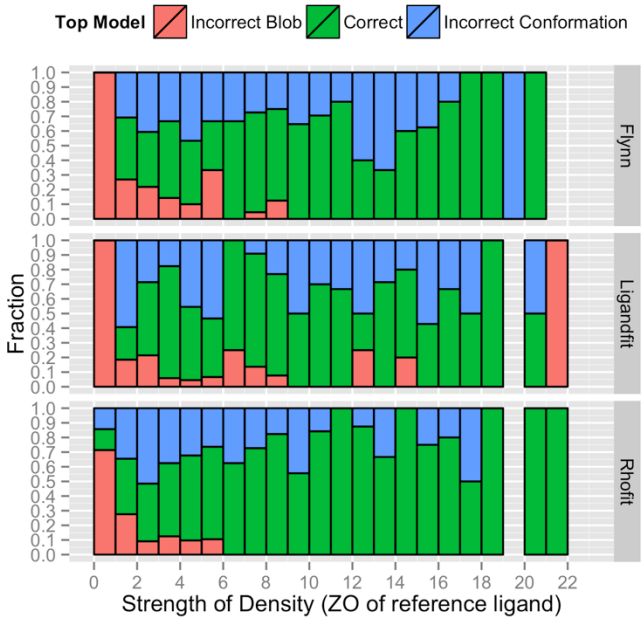


Figure 2.5. ***Correctness of the top-ranked model as a function of electron density strength of the reference ligand.*** *As the strength of the electron density (RSZO) for the ligand decreases, the chance that the top-ranked model is modelled in the wrong conformation at the correct site remains approximately constant, but the chance that the top-ranked model is found in the wrong electron density blob increases. As a tool for identifying the binding sites of ligands, this limits the usage of ligand-fitting methods in finding previously-unknown binding sites of weak ligands.*

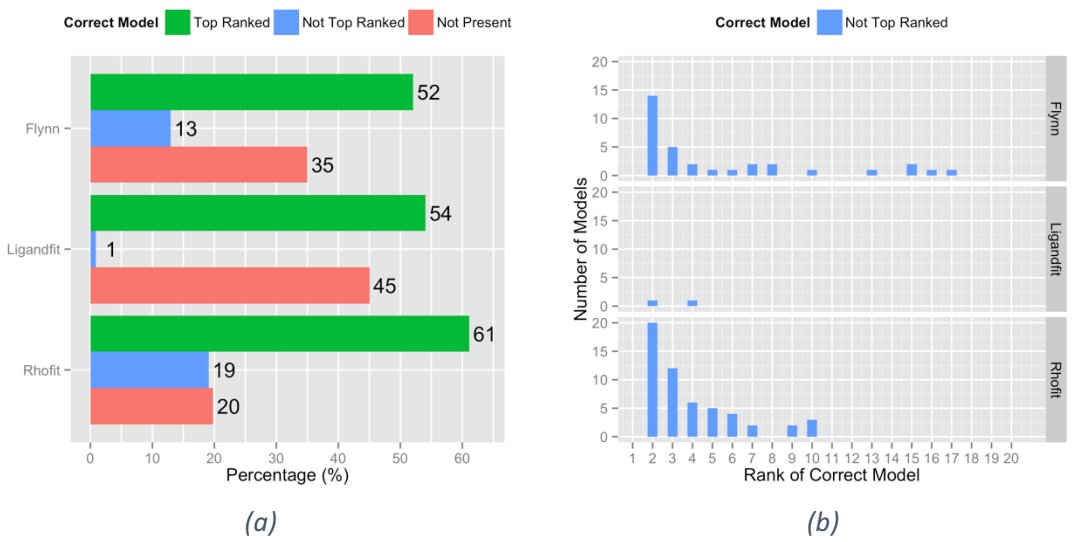


Figure 2.6. ***Rates at which a correct model is generated and ranked top, and the ranks of not-top-ranked correct models.*** *(a) Classification of the success and failure rates of the ligand-fitting programs. A correct model is defined as one that is within 1Å RMSD of the reference (SGC-modelled) ligand. "Not present" means that a correct ligand was not present in the list of generated decoys. RHOFIT has both the highest success rate in fitting (61%), and the lowest rate of search errors (20%). (b) Rank of the highest-ranked correct model when the top-ranked model is not correct. The lowest-ranked models are 35th for FLYNN (Figure 2.7), 4th for ligandfit and 10th for RHOFIT. ligandfit only generates five models, and RHOFIT only generates ten models, by default. The second-worst misranked model for FLYNN is ranked 17th. The average ranks (and standard deviations) for not-top-ranked correct models are 3.91 (2.33) for RHOFIT and 6.26 (6.91) for FLYNN. When the top-ranked correct models are included, this decreases to 1.69 (1.68) for RHOFIT and 2.05 (3.70) for FLYNN.*

An overall summary of the prevalence of ranking and search errors is shown in Figure 2.6. The ranking error rate is calculated as the percentage of cases where the correct model was generated but not selected as the top-ranked model. The search error rate is the percentage of cases where a correct model was not generated.

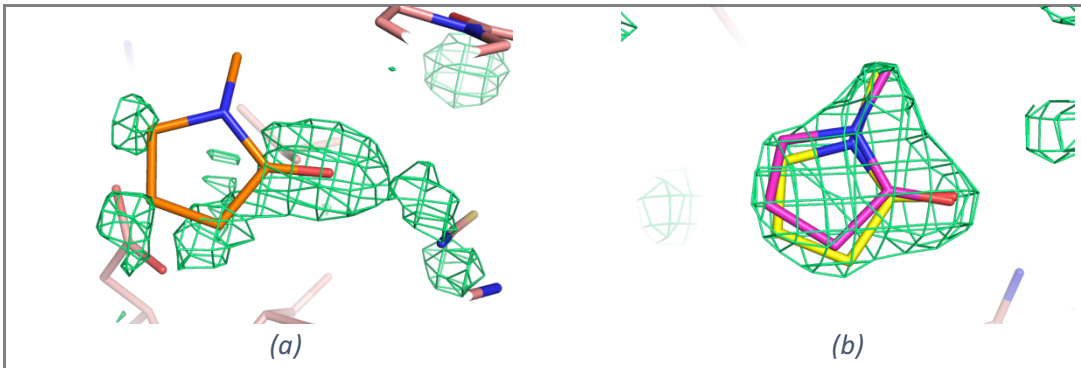
RHOFIT succeeds in generating a correct model 80% of the time, compared to 65% and 55% for FLYNN and ligandfit respectively. This is somewhat surprising, as FLYNN generates hundreds of decoys per fitting (maximum 354 decoys), where RHOFIT is limited to a maximum of ten decoys, and ligandfit to a maximum of five, by default. It would therefore be expected that both RHOFIT and ligandfit were more susceptible to

search errors than FLYNN, but this is not the case for RHOFIT. This observation suggests that RHOFIT explores a larger search space than the ten models that are output, but this cannot be determined as the method is unpublished.

Ranking error rates are calculated as the rate at which correct models are not top-ranked when they are generated. Since ligandfit generates only five models, which are selected to be diverse, it exhibits few ranking errors (only 1.8% of correct models are not top-ranked). In contrast, RHOFIT and FLYNN exhibit ranking errors of 24% and 20% respectively.

For the worst ranking failure from FLYNN, the correct model is ranked 35th of 78 models (Figure 2.7). This failure appears to be due to a down-weighting of the RSCC (between the model and the difference density) by the chemical strain component of the combined score: the correct model has an RSCC of 0.828 and the top-ranked model has an RSCC of 0.423. This characteristic of down-weighting a correct model with a high RSCC is shown more often by FLYNN than the other two programs, where most ranking errors arise between models with very similar RSCCs (Figure 2.8).

The rates of search and ranking errors, as a function of the RSZO of the reference ligand, are shown in Figure 2.9. The failures of the ligand-fitting programs are once again seen to increase as the strength of the electron density decreases. However, the rate of generating a model in the correct region of the map remains constant over the reference ligand RSZO (Figure 2.9b), suggesting that for the ligands in our test dataset, the ligand still presents more difference density than any water molecules, allowing it to be detected reliably by the ligand-fitting programs.



(a) (b)

Figure 2.7. ***An example of a ranking error from FLYNN.*** *Original models and refitted models from FLYNN for dataset CREBBPA-m004. "apo" mF_o-DF_c maps used for fitting (green/red) are contoured at ±3σ; the resolution is 1.86Å. (a) Top-ranked model (orange) from FLYNN. The RSCC of the model to the difference density, from FLYNN, is 0.423. (b) Reference (SGC-modelled) ligand (yellow) and the highest-ranked correct model from FLYNN (magenta). 78 decoys were generated; this model was ranked 35th. The RSCC of this model, from FLYNN, is 0.828. This is a ranking error caused by the FLYNN scoring function, where the crystallographic term has been down-weighted incorrectly in favour of another term. The model in (b) is clearly more likely than the model in (a), by visual inspection.*

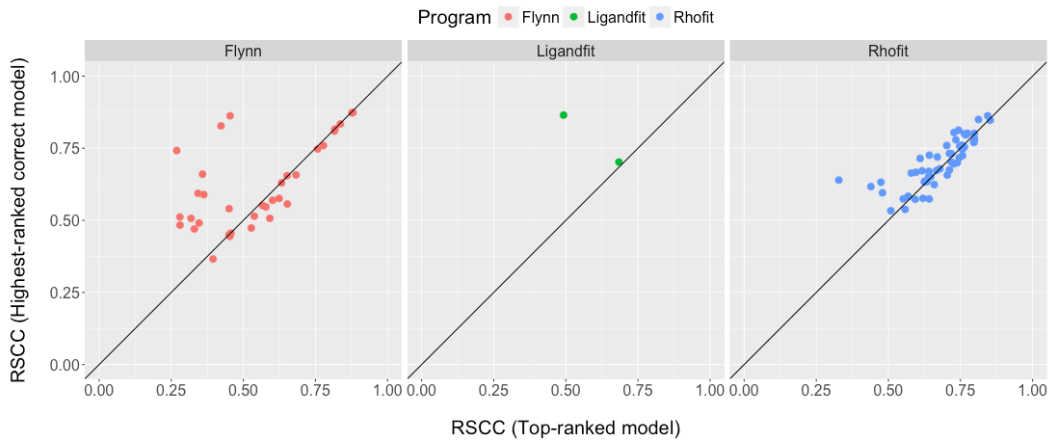


Figure 2.8. ***Most ranking errors occur between models with similar RSCCs.*** *The RSCCs for the top-ranked model and the highest-ranked correct model - as calculated by the fitting programs - are shown for the models where the top-ranked model was not correct (FLYNN: 34 models, Ligandfit: 2 models, RHOFIT: 54 models). Most points fall along the line of identity: two similarly fitting models were generated, and RSCC failed to discriminate between them. Points in the upper-left sector are cases where the correct model showed a better fit to the density than the top-ranked model. These represent significant ranking errors, where additional terms in the scoring scheme have down-weighted the contribution of the crystallographic term. Conversely, the points along the line of identity represent cases where additional terms have failed to discriminate between two equally-well-fitting models. Flynn suffers most from errors due to down-weighting of the RSCC (approx. 10 models of 34 deviate significantly from the identity line). On the other hand, RHOFIT only mis-ranks a correct model when the RSCC is very similar for both models.*

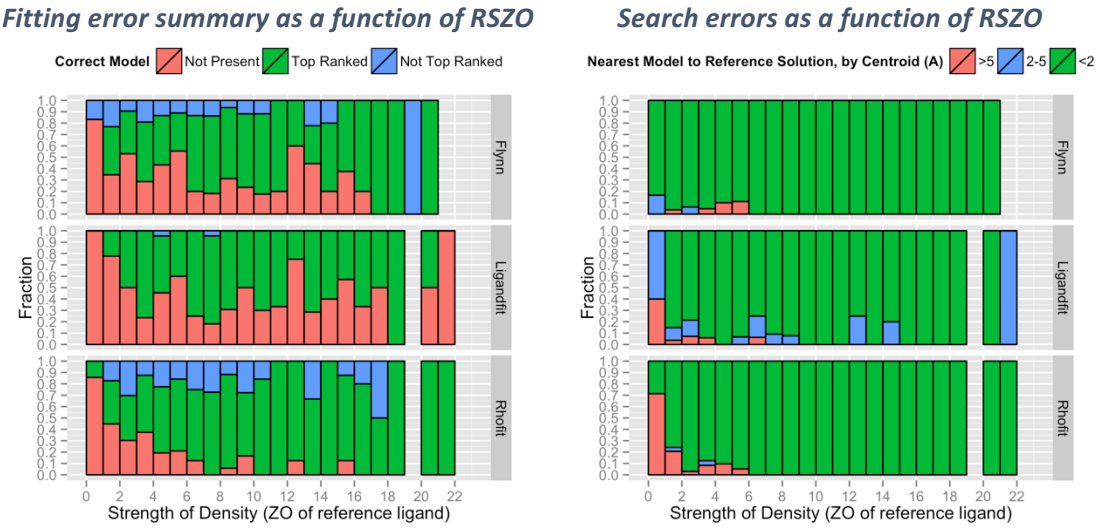


Figure 2.9. ***Fitting programs fail to generate a correct ligand model more frequently as the electron density strength for the ligand decreases.*** *(a) Generation of a correct model, as a function of electron density strength, RSZO, for the reference model. The rate of not ranking a correct model top (blue bars) remains constant as a function of RSZO. The rate of generation of a correct model decreases as a function of RSZO for RHOFIT in particular. This tendency is not as pronounced for ligandfit or FLYNN. (b) Nearest model to the reference model, as measured by the distance between the centroid of the two models, as a function of RSZO of the reference model. The rate of failing to generate a model in the correct region of the map increases as the strength of the electron density for the ligand decreases, although models are generated in the correct regions for most ligands.*

2.4.4 Classifying the correctness of the model

After model generation and ranking, we need to determine the correctness of the generated models. Here, we only look at the correctness of the top-ranked model. The ligand-fitting programs use a combination of the RSCC of the model density and the mF_o-DF_c density and other scores. We further record a range of electron density metrics for the fitted, partially-refined and fully-refined models against the associated refined $2mF_o$-DF_c density; recorded metrics are as described in section 1.3.10.

Scoring the fitted model against the $2mF_o$-DF_c density with EDSTATS before refinement of the ligand gives no information about the correctness of the ligand model (Figure 2.10). There is actually a weak anti-correlation between the RSCC from EDSTATS for the fitted model and the fitting programs (correlations are -0.36 for Flynn, -0.20 for

ligandfit, and -0.23 for Rhofit). This arises because the model being scored has arbitrary B-factors and unitary occupancy: the profile does not match the electron density and therefore scores badly against the $2mF_o$-DF_c density (this affects all metrics, not only the RSCC). However, the RSCC calculated by the fitting programs is calculated against the mF_o-DF_c density, and this proves to be more useful for discriminating between correct and incorrect models.

To identify the most useful metric for classifying the model correctness we trained a generalised linear model (GLM; McCullagh & Nelder 1989) – using a binomial error distribution and logistic regression, to predict the correctness of the models – on each of the metrics in turn (Figure 2.11). The usefulness of each metric was measured using the Akaike Information Criterion (AIC), which measures the quality of a statistical model; a smaller value for the same number of parameters reflects a better model.

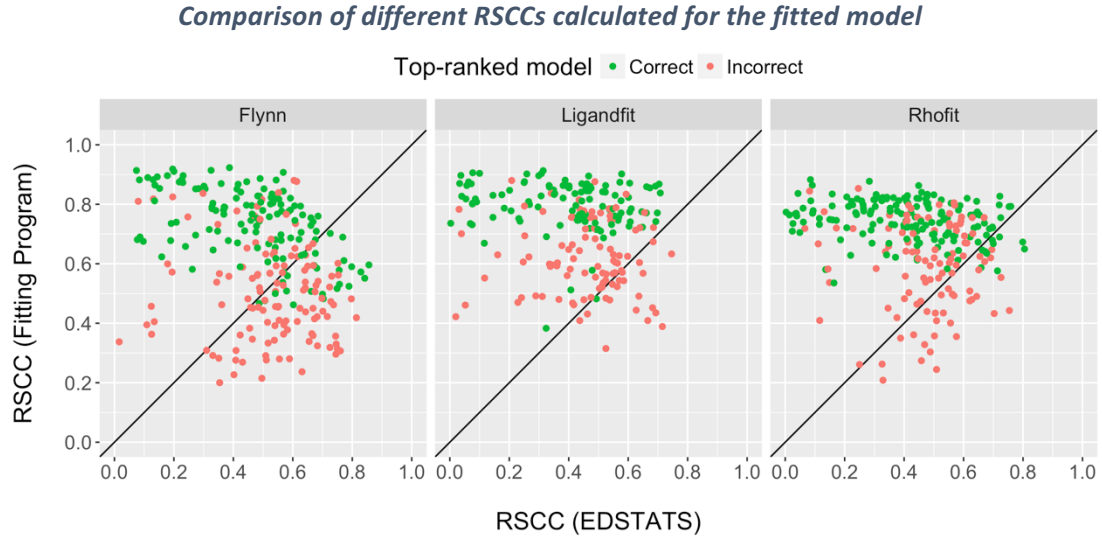


Figure 2.10. ***Scoring an un-refined model against the $2mF_o$-DF_c density by EDSTATS is not useful for determining the correctness of the model.*** *The EDSTATS population RSCC (x-axis) does not discriminate between correct models (green circles) and incorrect models (red circles). This would be shown as the green and red circles being separated into two groups along the x-axis. However, the RSCCs calculated by the ligand-fitting programs are scored against the mF_o-DF_c density, and this does allow discrimination, even though the same model is used (this is visually shown by separation of the green and red circles along the y-axis).*

Contrary to the accepted convention of using RSCC in the literature (Pozharski et al. 2013; Weichenberger et al. 2013; Deller & Rupp 2015; Schiebel, Krimmer, et al. 2016), the metric with the lowest AIC is the real-space R (RSR) for both the partially- and fully-refined models (Figure 2.11). The widespread usage of RSCC, however, is not unexpected, as it performs almost equally well to the RSR, but has one great advantage – it does not require the model density to be scaled to the experimental density.

The second- to fifth-most useful metrics remain constant between the partially- and fully-refined models. The population RSCC is only marginally less useful for classification than the RSR, and is marginally more useful than the sample RSCC, of which the population RSCC is a finite-sample-size-corrected version. The RSZO and the average B-factor both measure the strength/precision of the density for the ligand: refinement indicates that a model is not present by inflating B-factors.

The various RSZD metrics are not useful for detecting whether the model of the ligand was correct (RSZD(+) measures the significance of positive difference density only, and RSZD(-) similarly for negative difference density). As RSZD is a measure of difference density over the model, and refinement ultimately aims to minimise difference density, this may reduce the power of RSZD to classify the correctness of the model.

The high and low resolution limits serve as controls, since these can have no power to predict the correctness of the model (all datasets contained a ligand to be modelled). Lastly, the occupancy holds a small amount of predictive power when the coordinates are held fixed during refinement (for partially-refined models), reflecting perhaps that models in weak density are less likely to be correct; this is lost when the coordinates are also allowed to vary.

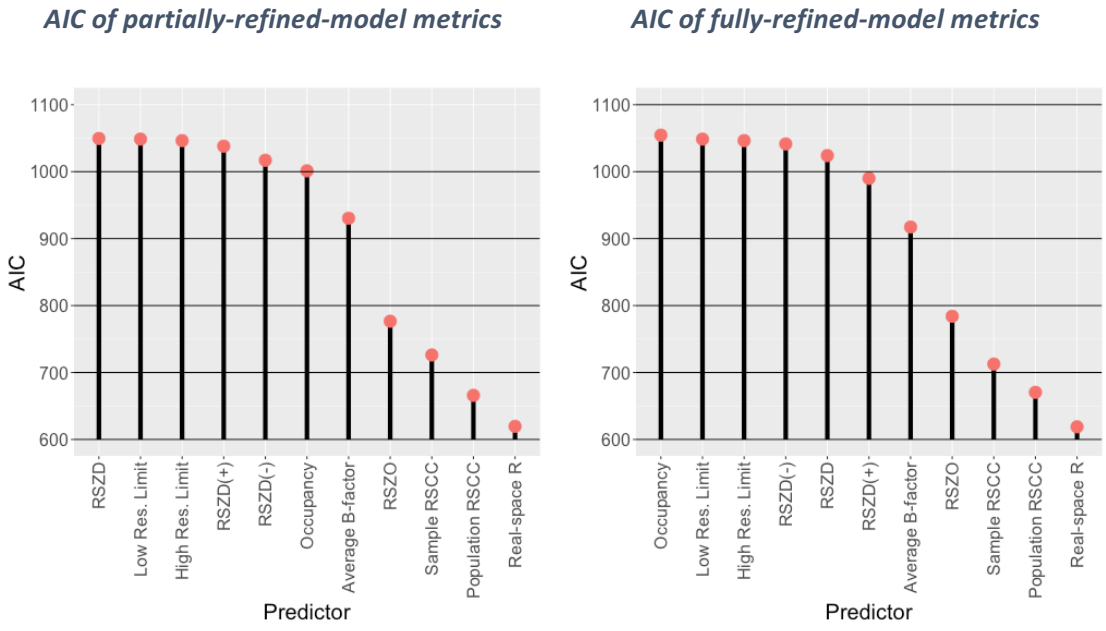


Figure 2.11. ***Power of different model metrics to predict model correctness.*** *The Akaike Information Criterion for metrics scored against (a) partially-refined and (b) fully-refined models when used to predict the correctness of the generated ligand model. The most useful metric for determining the correctness of the model is the Real-space R, contrary to the extensive use of RSCC in the literature.*

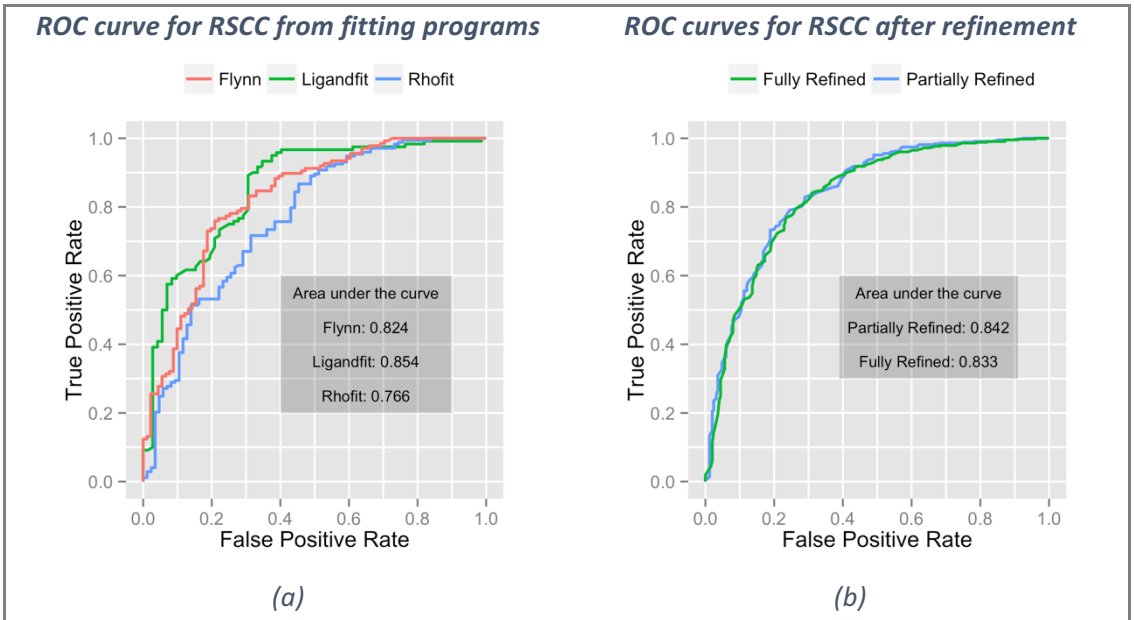


Figure 2.12. ***The RSCCs from the fitting programs are approximately as discriminatory as the RSCCs for refined models.*** *Receiver operating characteristic (ROC) plots for (a) the RSCC from the fitting programs and (b) the population RSCC for fully-refined models as calculated by EDSTATS, based on their ability to discriminate between correct and incorrect models. The areas under the ROC curves are all roughly comparable, with the notable exception being for RHOFIT, which performs marginally less well than expected.*

Combinations of multiple scores for training the GLM could be used to increase the ability to discriminate between the models. However, there are several problems in doing so. Metrics such as the B-factor and RSZO are highly resolution-dependant, precluding their use in a general scoring scheme. Furthermore, many of the metrics are highly correlated, negating their value in a combined scoring scheme. Generating a combined model with RSZD and RSR only marginally increases the discrimination performance (data not shown).

Refinement of models adds a computational burden to any ligand-discovery pipeline. If possible, it would be preferable to use the RSCC scores provided by the fitting programs, if they can discriminate on model correctness. The receiver-operator characteristic (ROC) plots (Fawcett 2006) for the RSCCs from the fitting programs, and the population RSCCs for the partially- and fully-refined models are shown in Figure 2.12; these ROC plots show the effect of filtering the generated models by a particular cutoff of RSCC and how this affects the numbers of correct models (true positives) and incorrect models (false positives).

A good classifier and cutoff will have a large true positive rate (TPR; number of correct models above the selected cutoff divided by total number of correct models) and a low false positive rate (FPR; number of incorrect models above the selected cutoff divided by the number of incorrect models).

Largely, the refined RSCC offers no more information on the correctness of the ligand model, compared to the RSCC from the fitting programs, as all ROC curves have similar areas under the curve.

2.5 Discussion

The pipeline described above represents a very simplistic approach to ligand-detection; there are several aspects of the pipeline that are considerably naïve, such as the removal of all waters prior to refinement. Recently developed methods (Schiebel, Krimmer, et al. 2016) follow much more elaborate modelling and refinement protocols, with careful water placement, and correspondingly see increased difference density for bound ligands.

2.5.1 Ligand binding locations are difficult to identify when density is weak

The ligand-fitting programs used here fail increasingly to generate the correct model when the electron density for the ligand is weak. Their success rates fall from 52-61% for all ligands to 33-36% where the RSZO is less than three. For even weaker ligands, such as might be expected in fragment screening, the success rate can be expected to fall further.

Our experiment here has made this task more difficult than necessary through the removal of water molecules from the model. Water removal both decreases the quality of the phases, making the maps noisier, and increases the RMS of the difference map, which will make fewer features appear at the 3σ level (or any other fixed level).

Difference-density-based methods require the ligand to present clear, connected difference density, such that the blob is large enough to be detected. For a ligand at low resolution, or a partial-occupancy ligand, this difference density may simply not be present above the 3σ level in the difference map; this is not a reflection of the ligand density, but rather our ability to model the rest of the crystal, which contributes to the quality of the phases and the RMS of the difference map.

2.5.2 Density metrics discriminate well between correct and incorrect models

The commonly used RSCC and the less-commonly-used RSR are both capable of identifying high-quality ligand models, with areas under the curve for RSCC of approximately 0.8. However, where the ligand is present at partial occupancy, the RSCC or RSR of a correct ligand model will be degraded if the other superposed crystal states are not modelled or accounted for in the calculation. This is a further fundamental problem when scoring ligands against the electron density, unless the ligand is present at unitary occupancy (discussed further in Chapter 5). This situation is made worse when the ligand is un-refined, as "phase-absence" makes the density for the ligand worse than expected: the ligand does not yet form part of the model, so the phases are sub-optimal.

2.5.3 Missing features from the pipeline

The pipeline used currently adopts a naïve approach to the refinement of crystallographic data: "phase memory" of the density is potentially overlooked in re-refinement of the structures after the removal of the ligand. This could be compensated for through the use of simulated annealing (Brünger et al. 1989), or by simply adding noise to the model coordinates.

Two popular programs are also missing from the ligand-fitting programs: Coot (Emsley et al. 2010) and ARP/wARP (Carolan & Lamzin 2014). The results for these two programs are likely to be very similar to the results from the three programs tested: both Coot and ARP/wARP use difference-map-based methods of ligand detection.

2.6 Chapter Summary

From the analysis above, it is my conclusion that ligand-fitting programs are in general ill-suited to the detection of weakly-bound ligands in crystallographic datasets. This is in keeping with the intended purpose of the programs: ligandfit was specifically designed to fit large drug-like molecules (personal communication).

Weaknesses lie both in methods used to detect the bound ligands – the necessity of the presence of difference density – and in the need to generate a "correct" model to be able to confirm (via e.g. a high RSCC) that the blob is an interesting blob for a bound ligand. There is currently no way of classifying the "interesting-ness" of a blob in a model-free fashion, other than the use of isomorphous difference maps, which are limited in their use by their requirement of strict isomorphism.

Chapter 3

The PanDDA method: A novel multi-dataset approach to the identification of "changed-state" crystallographic signal

---

"Wouldn't it be nice if the solvent wasn't in the way [of the ligand]."

The previous chapter has shown that the automated identification of binding ligands is impeded by our inability to accurately identify ligand-binding locations; the presence of difference density from other unmodelled molecules introduces noise which can be misidentified as the binding ligand, and a high-scoring model is required to confirm that a region corresponds to a bound ligand. Furthermore, it is not guaranteed that binding ligands will generate enough difference density to be detected; current methods based on the presence of usable difference density sometimes only see evidence of binding after extensive modelling and refinement (Schiebel, Krimmer, et al. 2016).

Even in the cases where weak binding can be detected by such methods, neither of the standard crystallographic maps will unambiguously resemble the pose of the ligand; the crystallographic density is *always* a superposition of the bound and unbound states for non-unitary occupancy ligands. This fact is rarely acknowledged in the literature and never corrected for by current methods. With no way to identify which density belongs to which state, ambiguity is always present during modelling, inviting misinterpretation.

In this chapter I describe the novel methods that I have developed to remove the dependence on the presence of difference density for ligand identification, while increasing the confidence in identified ligands; regions are only identified where there is sufficient evidence for the presence of a "changed-state" (i.e. a bound ligand).

I further present an approach for removing the superposed density that impedes map interpretation; this results in clear density for the changed-state, even for low-occupancy ligands. Finally, these novel maps are incorporated into a simple modelling approach, which naturally generates the multi-state ensemble models that are required to describe the superposition of states present in most ligand-bound crystals.

3.1 A new approach: The PanDDA paradigm

The new *PanDDA* (Pan-Dataset Density Analysis) method introduced in this chapter takes a novel approach to the detection of changed-state crystallographic signal, such as binding ligands (Figure 3.1). The PanDDA method seeks to overcome the two major obstacles to ligand identification and modelling described above: the need for accurate, sensitive identification of binding events and the removal of ambiguity introduced by the superposition of multiple states at the location of a binding ligand.

The accurate and sensitive identification of binding events is made possible by contrasting *bound* datasets against a series of *unbound* "ground-state" datasets. The principles of isomorphous difference maps (as in Rould & Carter 2003) are extended into real-space; the flexibility of real-space analysis permits changes from the unbound state to be rigorously detected whilst avoiding the restriction of strict isomorphism.

A voxel-by-voxel analysis of an ensemble of aligned crystallographic maps allows small but significant deviations from the ensemble to be detected, such as those caused by

binding ligands (Figure 3.1a). The analysis of an ensemble of unbound datasets further allows opportunities not possible for classical isomorphous difference maps: averaging over multiple datasets results in a low-noise representation of the unbound state (Figure 3.1b), and an analysis of the variation between unbound crystals can be used to suppress noise arising in variable regions (e.g. crystal contact regions). This first stage of the PanDDA method results in the identification of regions of electron density *only* where the putative bound dataset is significantly different from the unbound state.

Once these interesting "events" have been identified, the observed density remains a superposition of the bound-state and the unbound-state density (Figure 3.1c). By subtracting an appropriate fraction of the unbound-state density from the full-crystal observed density, generally clear density is revealed for only the changed state (i.e. the bound state) of the crystal (Figure 3.1c-e).

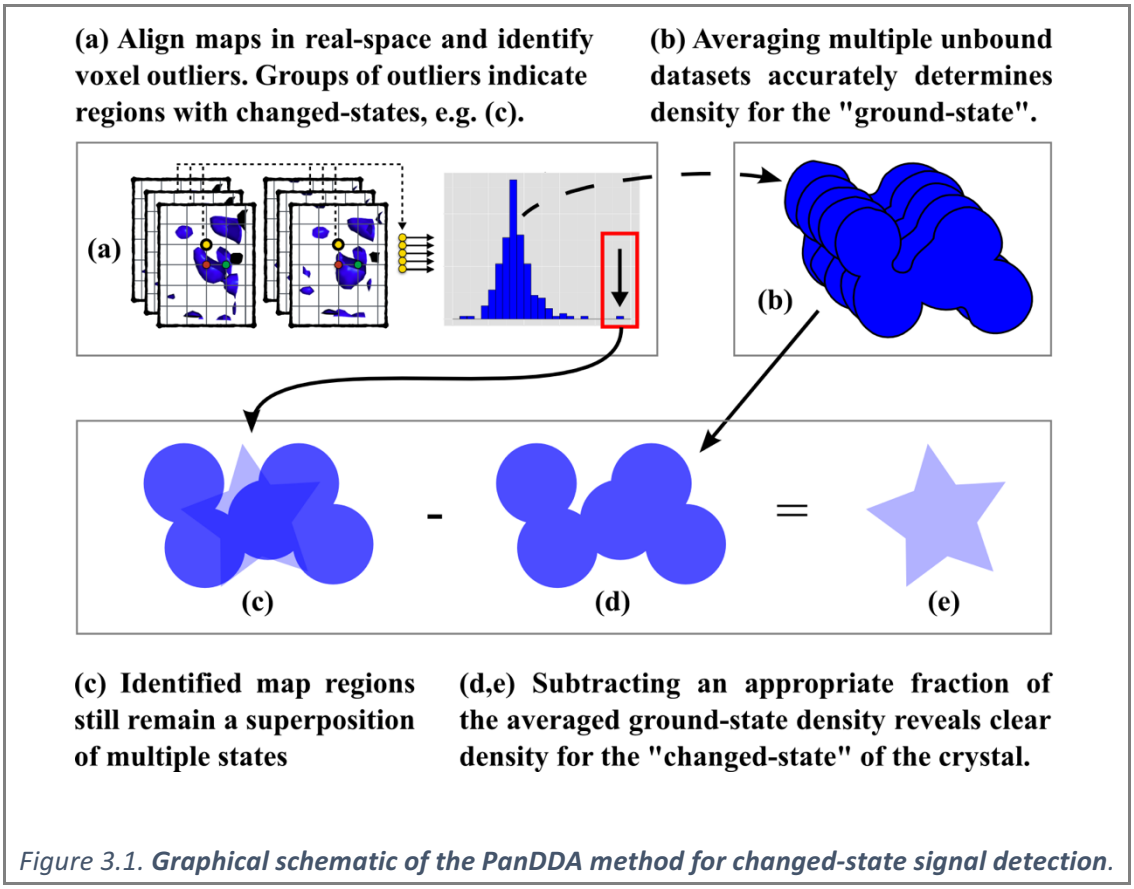


Figure 3.1. ***Graphical schematic of the PanDDA method for changed-state signal detection.***

3.2 PanDDA terminology

The terminology used in this chapter and the rest of the thesis is described in Table 3.1.

Table 3.1 ***Descriptions of PanDDA terminology.***

Terminology	***Description***
Ground-state crystal	The normal state of the crystal under experimental conditions (without a binding ligand or another changed state present). Typically, this state may be called the *apo* state of the crystal, but this is not always the case as it can include bound buffer molecules, as well as larger molecules that are always present in the crystal (e.g. a bound molecule which is co-crystallised with the protein). The ground state may contain contributions from multiple superposed states of ordered or disordered states of the protein or solvent molecules. Where multiple states are present, this distribution of states is also called the ground state.
Ground-state dataset	A crystallographic dataset obtained from a ground-state crystal.
Ground-state map	Electron density map for the ground state of the crystal.
Changed-state crystal	A crystal in which a fraction of the crystal has transitioned away from the ground state, due to e.g. the binding of a ligand.
Changed-state dataset	A dataset obtained from a changed-state crystal. This dataset consists of a superposition of the ground state and the new changed state.
Changed-state map	Electron density map for *only* the changed state of a crystal (e.g. the bound state of the protein).
Event	A statistically significant local deviation from the ground state. These may be caused by binding ligands, or by stochastic perturbations to the crystal, such as sidechain reordering.
Event map	An approximation to the changed-state map, obtained by subtracting a fraction of the ground-state map from the full-dataset map of a changed-state crystal.

Native coordinate frame	The coordinate frame from the original crystallographic dataset, based on the original crystallographic origin.
Reference coordinate frame	A frame of reference defined by a reference dataset, to which all other datasets are aligned. All analysis and changed-state modelling takes place in the reference frame, before being transformed back to the native coordinate frame for refinement. The reference frame is a programmatic artefact of the current implementation of the PanDDA algorithm.

3.3 PanDDA method overview

A PanDDA is performed on a series of refined crystallographic datasets, which must have $2mF_o$-DF_c structure factors. Analysis and comparison of $2mF_o$-DF_c maps allows a direct analysis of our best estimate of the crystal density, in contrast to the biased F_o maps used in conventional isomorphous difference maps. Although these datasets are not required to be strictly isomorphous – such as for an isomorphous difference map – they must be representable by a similar crystallographic model (same number of copies in the ASU, same identity and number of atoms): they must be of the same *crystal form*.

The PanDDA method comprises four parts, which are described in the sections below: the alignment of the datasets to a reference dataset (section 3.4); the statistical analysis of the aligned electron density maps (section 3.5); the identification of "interesting events" in datasets (section 3.6); and the subtraction of the superposed ground-state density at the identified sites to reveal the event density (section 3.7).

3.3.1 Datasets

Several examples from the application of the PanDDA method to fragment screening datasets are used in this chapter to demonstrate the PanDDA method (data was collected by other members of the SGC and DLS beamline i04-1). Each fragment screen

consists of a collection of single-compound datasets, where only one compound has been added to each protein crystal. A summary of the protein crystal forms is shown in Table 3.2. The complete results from the application of the PanDDA method to these datasets are presented in Chapter 4.

Table 3.2. ***Summaries of the datasets in the analysed fragment screens.*** *R-free and R-work are calculated after refinement of the reference model with Dimple for each dataset.*

Protein Acronym	**JMJD2D**			**BAZ2B**		
Protein Name	**Lysine-specific demethylase 4D**			**Bromodomain adjacent to zinc finger domain 2B**		
Number of Datasets	226			200		
Resolution Range (Å)	1.1-2.6			1.5-2.5		
Mean Resolution (SD) (Å)	1.45 (0.21)			1.79 (0.15)		
Space Group	P 43 21 2			C 2 2 21		
Mean Unit Cell Axes (Å)	71.42	82.17	96.57	58.03	71.42	150.41
Unit Cell Axes SD (%)	0.29%	2.02%	2.31%	1.51%	0.29%	0.25%
Unit Cell Volume SD (%)	0.80 %			3.03 %		
R-free quartiles	0.178	0.181	0.186	0.212	0.218	0.224
R-work quartiles	0.153	0.156	0.159	0.182	0.186	0.191
Domain Category	JmjN, JmjC (Jumonji)			Bromodomain		

Protein Acronym	**SP100**			**BRD1**		
Protein Name	**SP100 nuclear antigen**			**Bromodomain containing 1**		
Number of Datasets	116			292		
Resolution Range (Å)	1.3-2.7			1.5-3.6		
Mean Resolution (SD) (Å)	1.72 (0.22)			1.76 (0.33)		
Space Group	C 1 2 1			P 21 21 21		
Mean Unit Cell Axes (Å)	127.67	45.39	83.36	55.46	56.42	101.76
Unit Cell Axes SD	0.09 %	0.21 %	0.20 %	0.75 %	0.30 %	0.27 %
Unit Cell Volume SD	0.50 %			1.05 %		
R-free quartiles	0.203	0.207	0.213	0.215	0.223	0.238
R-work quartiles	0.169	0.173	0.176	0.181	0.187	0.199
Domain Category	PHD, Bromodomain			Bromodomain		

3.3.2 Assumptions for applying the PanDDA method to fragment screening data

Hit rates in fragment-screening experiments rarely exceed 25%, with the typical hit-rate in the region of 5-10%. Therefore, the large majority of obtained datasets can be considered as ground-state datasets; averaging over multiple fragment screening

datasets is assumed to approximate the ground state accurately. Where there is a high percentage of binding ligands, an iterative outlier-rejection method may be required to remove binding datasets from the ground state averaging and variation analysis.

3.4 Structure and map alignment

The first step in the PanDDA method is to align the datasets for analysis; the structures and maps need to be aligned to permit the voxel-by-voxel density comparison. Due to stochastic differences creating non-isomorphism between cryo-cooled crystals, the global crystal-averaged conformation of the protein varies slightly from dataset to dataset. These differences are typically small locally, but can culminate in significant global differences, preventing the use of conventional global alignment, or even chain-by-chain alignment: sections of the protein will remain misaligned, preventing meaningful analysis of these regions. To compensate for the global conformational variation, the protein structures are aligned to a reference structure using a custom flexible alignment method.

3.4.1 Selection of a reference dataset

The highest resolution dataset from a given set of crystallographic datasets provides the highest precision structure of the crystal form; by default, the highest resolution dataset is chosen as the reference dataset. If this dataset is subsequently detected to have a bound ligand, the analysis may need to be repeated using a different reference dataset; however, this is theoretically only necessary if the bound state in the reference dataset is present at high occupancy and causes the ground-state backbone conformation of the reference protein structure to become distorted in refinement. If the ground state

conformation is conserved in the reference structure under refinement, the dataset may still provide a good local reference for alignment.

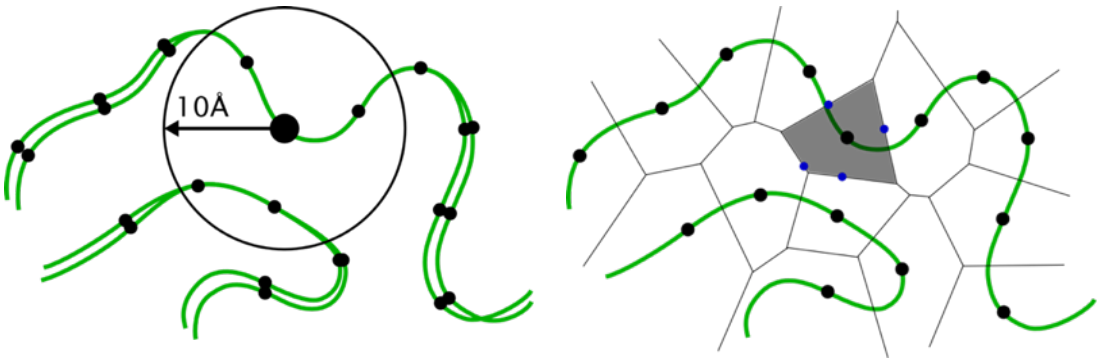
3.4.2 Flexible alignment of the protein structure

Once a reference dataset has been selected, every other dataset is aligned to it in turn. Conventional alignment algorithms preserve the structure of the protein and generate one rotation-translation alignment matrix for the whole structure. The flexible alignment algorithm generates one rotation-translation matrix *for each C-alpha* of the protein. For each C-alpha in the reference structure, all residues with a backbone atom within 10Å of the C-alpha are selected; the coordinates of the backbone atoms of these residues are aligned using a least-squares fit to the coordinates of the equivalent atoms in each dataset (Figure 3.2a).

For the alignment of coordinates between a dataset and the reference dataset, coordinates are transformed by identifying the nearest C-alpha, and using the associated alignment matrix to align the coordinates. This results in the transformation of coordinates as Voronoi cells, where the cells are centred on the C-alphas (Figure 3.2b). The application of this alignment method reduces any global variation in the structures, allowing locally similar regions to be compared (Figure 3.3).

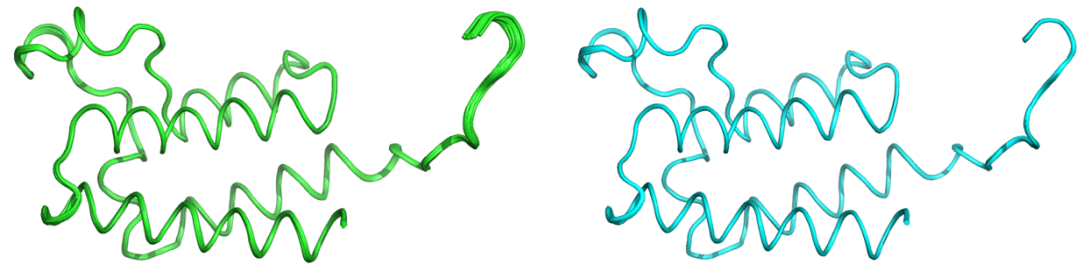
The variation in the alignment (rotation-translation) matrices for neighbouring residues is typically small; the alignment matrices of adjacent C-alphas generally vary smoothly over the protein structure (Figure 3.4). However, there also exist noticeable “hinge-points” of the protein structure, where the variation of the rotation-translation matrices between adjacent C-alphas is significantly higher across the datasets. These regions are typically in crystallographically-disordered sections of the protein, such as

loops. Further analysis of this structural variability between sets of approximately equivalent crystals is performed and discussed in Chapter 6.



(a) Local alignment *(b) Voronoi cell generation*

*Figure 3.2**. Alignment and transformation of the protein structure.** Two protein structures are shown as green lines; C-alphas are marked with black dots. (a) Generation of a local alignment for a C-alpha. For each C-alpha in the structure, all backbone atoms within 10Å are selected and aligned to the equivalent atoms in the reference structure using a least-squares fit (structures are shown aligned by these sections). The resulting alignment matrix is assigned to the selected C-alpha. (b) The C-alphas are used to define Voronoi cells. Coordinates within the cells are transformed by the transformation matrix of the associated (closest) C-alpha.*



(a) Global alignment *(b) Local alignment*

*Figure 3.3. **Example alignment of the protein structures for 99 datasets of BAZ2B**, displayed as cartoon loop using PyMOL. (a) Global alignment of the proteins, based on a least-squares fit on the protein c-alphas results in small but significant variation in some areas of the protein, mainly in the flexible N-terminus "tail" of the protein (right hand side of images). (b) Local alignment, as described in section 3.4.2, corrects for these variations, and shows that the local structure of the protein is still conserved.*

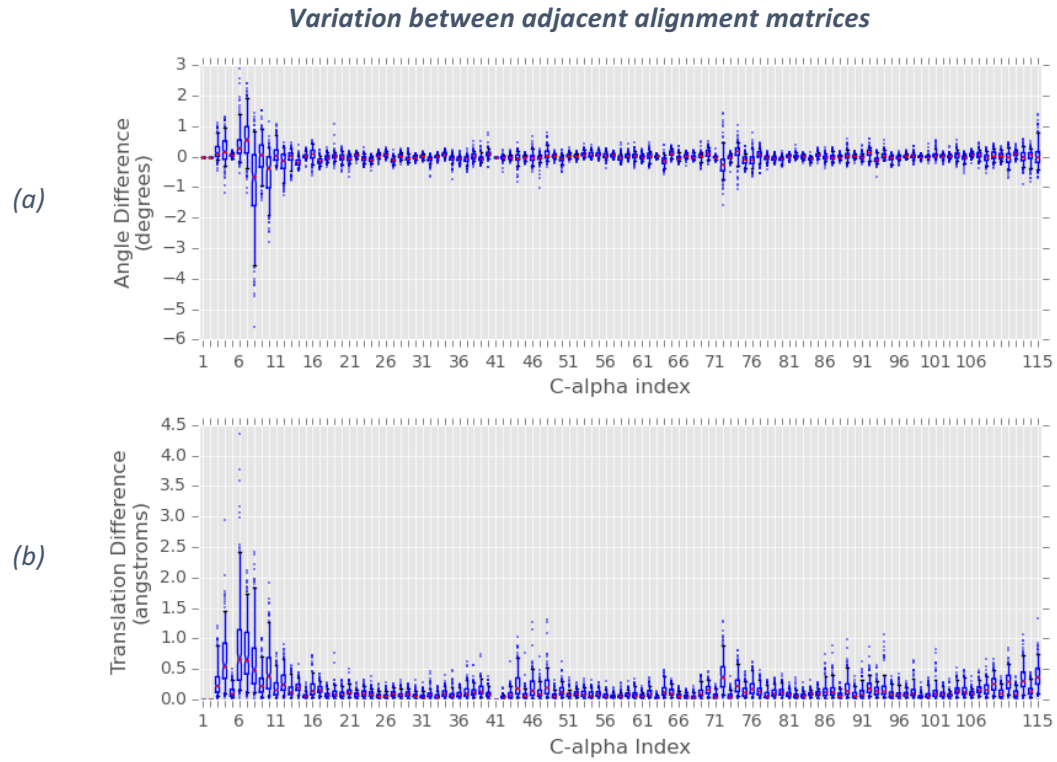


Figure 3.4. ***Variation in the flexible-alignment matrices along the protein structure.*** *Boxplots of differences between alignment matrices of adjacent residues for 200 datasets of BAZ2B. (a) Alignment matrix angle differences (rotation angle of matrix of residue n+1 – rotation angle of matrix of residue n). (b) Alignment matrix translation differences (norm of translation vector of matrix of residue n+1 – translation vector of matrix of residue n). The differences between adjacent alignment matrices are generally small, but there are "hinge-point" sections, such as the N-terminus on the left-hand-side of the plot, which exhibit large angle variations. The large translation differences in (b) exist to compensate for the larger rotations in (a); when coordinates are transformed that are distant from the coordinate origin, small differences in rotation angles result in exaggerated translations that need to be compensated for in the translation component of the matrix.*

3.4.3 Generation of crystallographic maps

For the analysis of an ensemble of crystallographic maps, all electron density must be calculated at the same resolution, such that each map contains the same amount of information. The analysis of datasets which cover a range of resolutions is described in the implementation in section 3.10.

For a set of crystallographic datasets, selected for analysis at a certain resolution, the structure factors across all datasets are first truncated to the common set of Miller

indices. This truncation prevents the inclusion of reflections in some datasets that are missing in others – this is particularly important for strong low-resolution reflections.

If strong low-resolution terms are present/missing in some datasets but not in the majority, no signal from binding ligands will be detectable in those datasets: due to the large amplitude of low-resolution structure factors, the strength of the electron density will be systematically higher/lower in some regions of the unit cell than in the ground-state, dominating the analysis of deviations from the ground-state. A solution to this problem is to replace missing low-resolution reflections with their refined estimates – maximum-likelihood DF_c values – prior to the application of the PanDDA method.

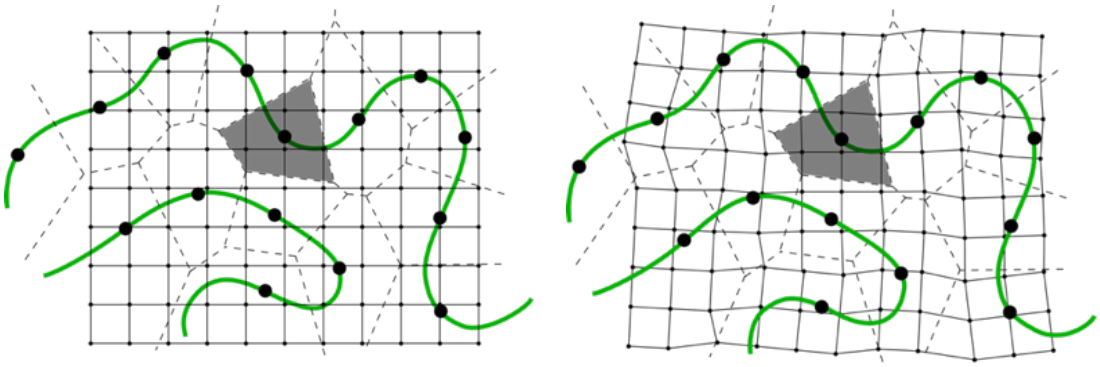
After common truncation to the common set of Miller indices, dataset structure factors are further truncated to the selected resolution, and Fourier-transformed to create electron density maps in real-space. By default, maps are rmsd-scaled. These maps are calculated in the crystallographic (native) coordinate frame of each of the datasets.

3.4.4 Alignment of crystallographic maps

The crystallographic maps are aligned in a similar manner to the structures. Firstly, an isotropic Cartesian grid is created in the reference coordinate frame to cover the aligned structures (Figure 3.5a). The default grid spacing is 0.5Å, but can be increased to reduce the computational burden for unit cells. The grid spacing is not chosen as a function of resolution since the same grid is used at multiple resolutions.

The coordinates of the grid are transformed to each dataset's *native* coordinate frame using the alignment matrix of the nearest C-alpha. The electron density in the dataset's *native* coordinate frame is then sampled at these points. The result of the transformation is a warped grid in the native coordinate frame (Figure 3.5b). Since the

grid is non-Cartesian in the dataset frame, maps are presented and modelled in the reference frame (discussed further in section 3.10.2).



(a) *Isotropic grid in the reference coordinate frame*

(b) Warped grid in the "native" dataset coordinate frame

Figure 3.5. ***Alignment of the crystallographic maps.*** *(a) An isotropic grid is created in the reference frame to cover the protein model and the surrounding space. The coordinates for this grid are transformed to the crystallographic coordinate frame of each dataset as the Voronoi cells from the structural alignment. (b) Transformation of the isotropic grid creates a non-isotropic grid in the dataset coordinate frame. Map values are sampled on the warped dataset grid and presented as maps on the isotropic grid in the reference frame.*

Once a map has been created and aligned to the reference coordinate frame, it is scaled by using a linear fit of the aligned map values to the reference map values, this generates a scaling factor and an offset. Scaling is only performed over map points within 1.8Å of a protein atom to avoid scaling over the solvent, which may be variably ordered between datasets; the density for the protein should be the most conserved feature between datasets, and therefore the best reference for scaling.

For the BAZ2B datasets, scaling factors are approximately one, and offsets are approximately zero; the datasets are already on a similar scale, as expected, since they are approximately equivalent crystals collected using the same data collection strategy.

3.5 Statistical map characterisation

To identify regions in individual datasets that deviate from the ensemble of datasets, I create and parameterise a statistical model for the density of the crystal form from the collection of datasets.

Once the crystallographic maps have been aligned and scaled, the distribution of map values at each grid point is analysed across the datasets. The electron density at a point in each dataset can be statistically modelled as an observation of the *true* electron density of the crystal at the point, with terms accounting for crystal variation and experimental error in the dataset. This can be written as

$$\rho_{i,m}^{observed} = \rho_m^{true} + \varepsilon_i, \qquad 3.1$$

where $\rho_{i,m}^{observed}$ is the observed map value in dataset i, at point m. ρ_m^{true} models the natural variation in the electron density at point m, independent of dataset, and ε_i represents the experimental uncertainty in the electron density in dataset i, independent of position.

The variability of the ρ_m^{true} term accounts for the fact that the crystal-averaged unit cell density may not be not the same for different crystals, and that small but systematic local fluctuations may exist between crystals. These areas may be in the crystal contacts, or disordered areas of the protein. ρ_m^{true} represents the "true" (unmeasurable) electron density for a crystal-averaged unit cell of the crystal form, of which each crystal (and associated dataset) is a sample.

The simplest statistical model is to assume that both the uncertainty in electron density values, and variation in electron density at a point arising from differences between the crystals, can be modelled by a normal distribution. Therefore, if

$$\rho_m^{true} \sim \mathcal{N}(\mu_m, s_m^2), \text{ and } \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad 3.2$$

then, for independent errors, the means and variances are additive, such that

$$\rho_{i,m}^{observed} \sim \mathcal{N}(\mu_m, \sigma_i^2 + s_m^2), \quad 3.3$$

where: μ_m is the mean value of the electron density at point m; s_m^2 is the variance of the "true" electron density at point m; and σ_i is the uncertainty in dataset i. Under this model, the parameters μ_m are estimated by taking the un-weighted average of all the ground-state density maps.

3.5.1 Estimation of dataset uncertainty

The obtained mean ground-state map can be used to estimate the dataset uncertainty, σ_i, for all datasets as follows. Subtracting the mean map from each dataset map we obtain a mean-difference map. By assuming that the experimental uncertainty and model phase errors in the electron density map are the major contributors to deviations from the mean map (the structural variation has been removed by the flexible alignment), the histogram of the mean-difference map values can be used to estimate the combined uncertainty of the dataset.

Calculating the quantiles of a theoretical normal distribution $\mathcal{N}(0, 1)$ and plotting them against the quantiles from the mean-difference map yields a Q-Q plot where the slope of the central portion of the map (between the ±1.5 theoretical quantiles) gives an estimate of the uncertainty of the dataset (Figure 3.6). This is equivalent to the method used in Tickle (2012) for calculating the uncertainty of an electron density map.

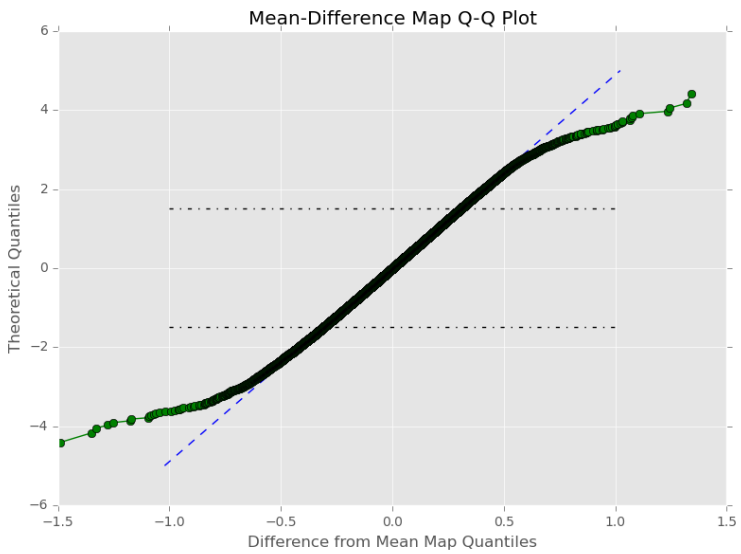


Figure 3.6. ***Characterisation of deviations from the mean map.*** *Q-Q plot against the normal distribution for the differences from the mean-map (over-plotted green and black circles), for dataset x474 of BAZ2B, analysed at 1.98Å (dataset resolution 1.975Å). The central portion of the plot, between the ±1.5 quantiles (red horizontal dot-dashed lines), is used to estimate the error in the map. The straightness of the section in the centre of the plot indicates that the small differences from the mean map are approximately normally distributed. The slope of the central portion of the graph gives the standard deviation of our (normally distributed) error distribution model for the dataset. The theoretical quantiles for this calculated error distribution lay along the blue dashed line. The areas of the Q-Q plot that deviate from the blue line show differences that are not accounted for by our model of experimental error.*

3.5.2 Estimation of point variability

After estimation of the dataset uncertainty, σ_i, which parameterises the *global* error distribution, there still remain deviations from the mean map that are not accounted for (Figure 3.6); these deviations are modelled by the point-variation term s_m. To estimate s_m, we apply a maximum likelihood method (Appendix B) to the statistical model of the density (equation 3.3), using the observed values $\rho_{i,m}^{observed}$, as well as estimates for σ_i and μ_m for the ground-state datasets.

The joint probability distribution function of the model in equation 3.3 is

$$\mathrm{JPDF}_m \sim \left(\prod_{i=1}^{n} \frac{1}{(\sigma_i^2 + s_m^2)^{1/2}} \right) * \exp\left(- \sum_{i=1}^{n} \frac{\left(\rho_{i,m}^{observed} - \mu_m\right)^2}{2(\sigma_i^2 + s_m^2)} \right), \qquad 3.4$$

for n datasets. Taking the natural logarithm of equation 3.4, we obtain

$$\ln(\mathrm{JPDF}_m) \sim \sum_{i=1}^{n} \ln \frac{1}{(\sigma_i^2 + s_m^2)^{1/2}} - \frac{1}{2} \sum_{i=1}^{n} \frac{\left(\rho_{i,m}^{observed} - \mu_m\right)^2}{(\sigma_i^2 + s_m^2)}, \quad 3.5$$

which can be simplified to

$$\ln(\mathrm{JPDF}_m) \sim -\frac{1}{2} \sum_{i=1}^{n} \ln(\sigma_i^2 + s_m^2) - \frac{1}{2} \sum_{i=1}^{n} \frac{\left(\rho_{i,m}^{observed} - \mu_m\right)^2}{(\sigma_i^2 + s_m^2)}. \quad 3.6$$

Maximising this function for s_m, yields

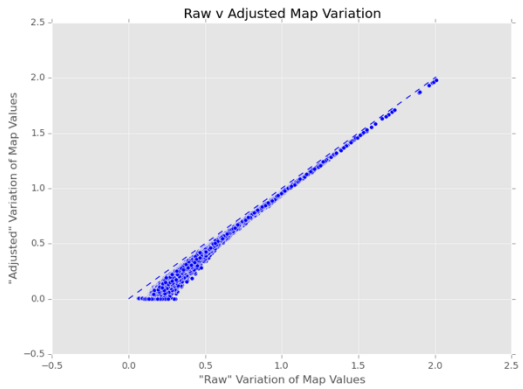
$$-s_m \sum_{i=1}^{n} \frac{1}{(\sigma_i^2 + s_m^2)} + s_m \sum_{i=1}^{n} \frac{\left(\rho_{i,m}^{observed} - \mu_m\right)^2}{(\sigma_i^2 + s_m^2)^2} = 0, \quad 3.7$$

which can be simplified, under the assumption that $s_m \neq 0$, to

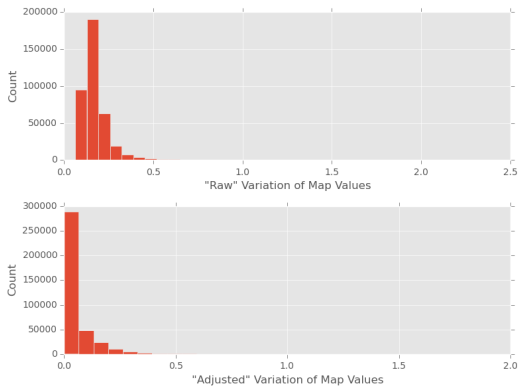
$$-\sum_{i=1}^{n} \frac{1}{(\sigma_i^2 + s_m^2)} + \sum_{i=1}^{n} \frac{\left(\rho_{i,m}^{observed} - \mu_m\right)^2}{(\sigma_i^2 + s_m^2)^2} = 0. \quad 3.8$$

Using the mean ground-state map values, μ_m, experimental values, $\rho_{i,m}^{observed}$, and estimated uncertainties, σ_i, equation 3.8 can be numerically solved to estimate s_m. The "adjusted" variation, s_m, is reduced compared to the uncorrected "raw" variation (rmsd of observed values, $\rho_{i,m}^{observed}$, from the mean value, μ_m, at point m), and the large majority of the points in the grid have a value of $s_m \approx 0$ (Figure 3.9): the majority of points in the unit cell exhibit no variation beyond the characterised experimental errors.

Those points that do exhibit variation that is not explained by our model of experimental error are variable regions of the crystal; electron density has been shown to vary between equivalent cryo-cooled crystals (Fraser et al. 2011). In the following sections, the estimation of s_m is used to suppress noise at these variable sites. The variation of the density around alternate sidechain conformations can further be used to detect heterogeneity of the protein structure (discussed in Chapter 6).



(a) Scatter plot of raw v adjusted variation



(b) Histogram of raw v adjusted variation

Figure 3.7. ***Comparison of the "raw" and the "adjusted" density variation at each point.*** *Results from the analysis of the variation in 60 BAZ2B datasets at 1.73Å. Analysis was performed on sigma-scaled maps: variation is in units of sigma. (a) Scatter plot of raw variation of each grid point plotted against the adjusted variation of each grid point (after accounting for experimental error). The adjusted value is smaller than the raw value, and for small raw values, variation can be explained by the uncertainty in the datasets only (adjusted variation ≈ 0). (b) Corresponding histograms of the raw and adjusted variation of the grid points. The raw histogram shows multiple grid points with variation between 0.07-0.40. However, the adjusted deviations show most variations are accounted for by experimental error, resulting in an adjusted value between 0-0.07.*

3.5.3 Convergence of s_m estimation

A minimum number of observations are required for the maximum-likelihood method described in section 3.5.2 to converge; this number will be dependent on the errors in the observed map values, and the inherent variability of the crystal system.

We analysed three collections of datasets to determine the minimum number of datasets required for the adjusted variation, s_m, to converge. Each collection had a minimum of 100 datasets; the number of datasets, and the analysed resolution are shown in Table 3.3 (the JARID1B data is not used elsewhere in this thesis). The best estimate of s_m is obtained by using all datasets for each collection.

To determine a convergence cutoff, available datasets were jack-knifed randomly into two halves and analysed separately. The 90, 95, and 99% quantiles of the absolute

difference between the derived s_m values for the jack-knifed halves define the 90, 95, and 99% convergence cutoffs, respectively; derived cutoffs are recorded in Table 3.3.

Multiple analyses were run with different numbers of randomly-selected datasets, and the percentage of s_m values within the 90, 95, and 99% convergence cutoffs of the best estimate (all datasets) was recorded. Convergence is achieved when X% fall within the X% cutoff of the best estimates, calculated with all datasets (e.g. when 90% of calculated values are within the 90% cutoff). This was performed five times, with different seeds.

From Figure 3.8, approximately 30 datasets are needed to reach the required cutoff, regardless of whether the 90, 95 or 99% cutoff was selected, as expected; this is the minimum number of datasets that should be used. The minimum number of datasets required for convergence will vary by crystal system, and an analysis of this number on a system-by-system basis is a potential direction of future study.

However, in the sections that follow, $s_m$is used mostly for noise-suppression, so using fewer datasets should only introduce more noise, rather than cause a loss of signal.

Table 3.3. ***S-adjusted map convergence cutoffs****. Map cutoffs are measured in sigma, as the analysis was performed on sigma-scaled maps. The analysed resolution for each protein dataset was selected manually to include the largest number of datasets for analysis.*

Protein	Resolution (Å)	Datasets Used	90% Cutoff (σ)	95% Cutoff (σ)	99% Cutoff (σ)
BAZ2B	2.00	170	0.05	0.07	0.10
JARID1B	2.90	120	0.03	0.04	0.07
JMJD2D	1.75	200	0.04	0.05	0.07

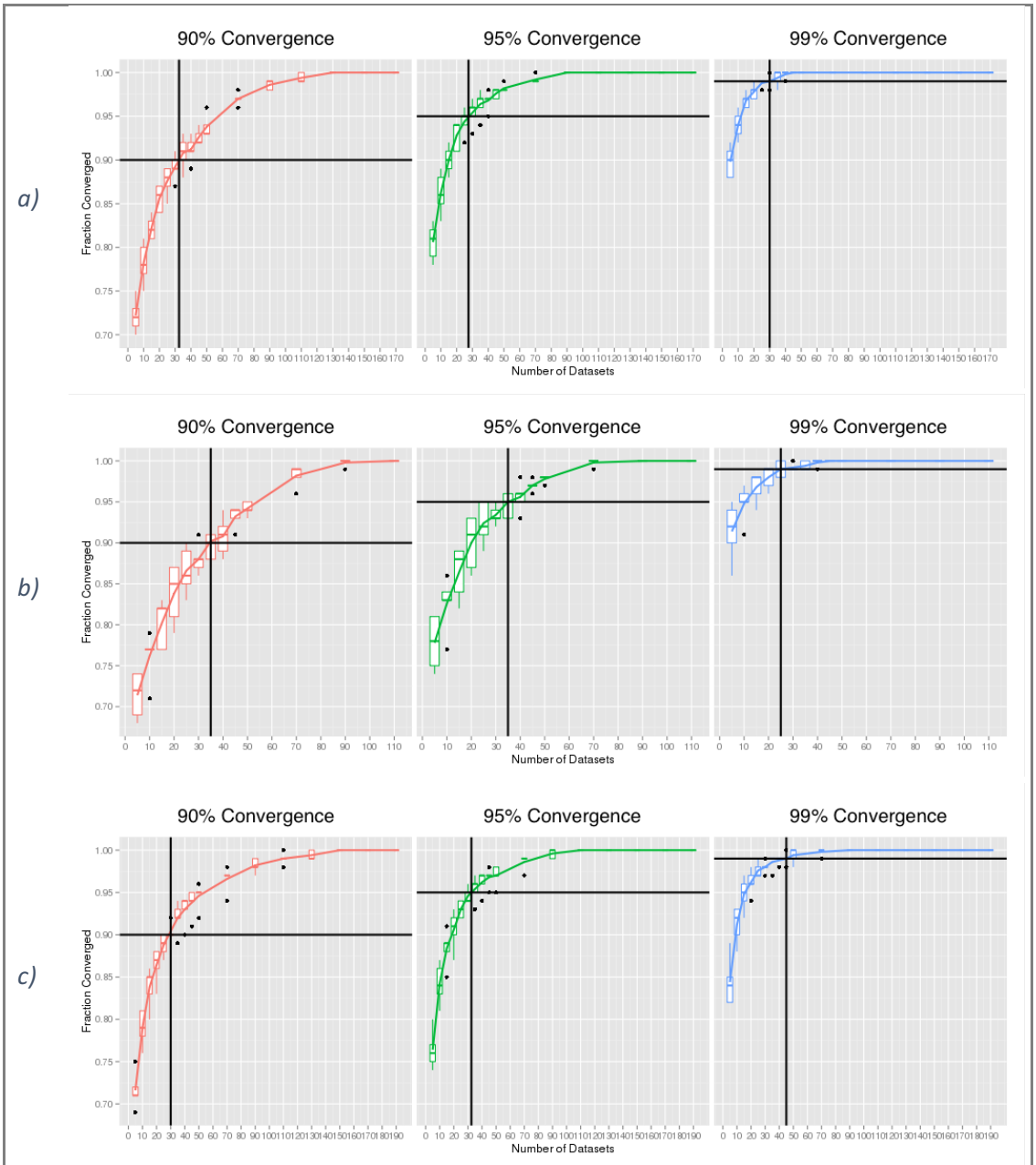


Figure 3.8. ***Convergence of adjusted s-map values.*** *Values on the Y-axis record the fraction of points within the appropriate cutoff (as listed in Table 3.3) of the best estimation of the variation (calculated using all datasets). Black horizontal lines are defined by the percentage cutoffs (90%, 95%, 99%) as appropriate. Vertical lines record number of datasets at which the horizontal black line intersects with the averaged fraction of grid points which have converged; this is the value where we consider the maximum likelihood method to have converged. Columns (left to right): 90%, 95%, 99% cutoffs. Datasets: (a) BAZ2B. (b) JARID1B. (c) JMJD2D.*

3.6 Z-map calculation

The parameterised statistical model allows the identification of areas of individual dataset maps that deviate significantly from the mean ground-state map. Significant local deviations are labelled as "events". By calculating Z-scores

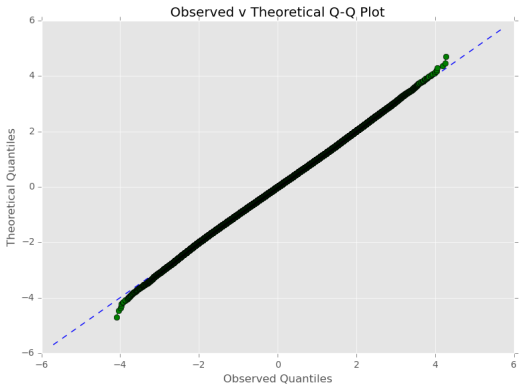
$$Z_{i,m} = \frac{\rho_{i,m}^{observed} - \mu_m}{\sqrt{\sigma_i^2 + s_m^2}}, \qquad 3.9$$

large Z-scores indicate significant deviations from the ground-state map. The distributions of Z-scores for individual datasets have improved normality compared to the simple differences from the mean, as expected (Figure 3.9a); the complete statistical model provides a good description of deviations from the averaged ground-state map for datasets where no ligand is bound. One basic characteristic of the Z-map is that in regions of high s_m values, ligands will need to bind strongly to be detected.

The existence of large Z-scores caused by the presence of a binding ligand can sometimes be seen in the Q-Q plot. For example, a binding ligand in dataset x434 of BAZ2B creates a cluster of large positive Z-scores, which cause a large deviation from the line of identity in the top-right of the Q-Q plot (Figure 3.9b).

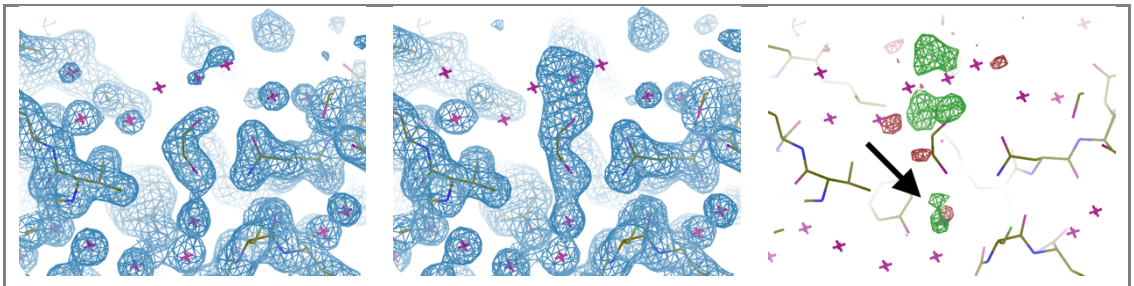
Parameters for identifying interesting regions were identified on the BAZ2B dataset, and found appropriate in the datasets presented in this thesis. Regions of individual datasets are identified as significant by contouring Z-maps at Z=2.5, and filtering remaining blobs by a minimum peak value of Z=3 and a minimum volume of 10Å^3 (volume of a water molecule is ~30Å^3). Neighbouring blobs are grouped if the minimum distance between them is less than 5Å. However, the determination of these parameters for application in the general case is the subject of future work.

The identified blobs in the Z-maps for dataset x434 of BAZ2B (as in Figure 3.9b) are shown in Figure 3.10. As high Z-values only indicate where density is different to that of the ground-state, the Z-map will not in general resemble the density for the ligand.



(a) Dataset with no events *(b) Dataset that contains a binding event*

Figure 3.9. ***Z-maps show increased normality compared to the differences from the mean map.*** *(a) Q-Q plot of Z-map values against the quantiles of the normal distribution for dataset x474 of BAZ2B (as in Figure 3.6), in which no binding ligand was detected. The plot is close to the identity (blue, dashed); the Z-map values are approximately normally distributed, and more normally distributed than in Figure 3.6. (b) Q-Q plot as in (a), but for dataset x434 of BAZ2B, analysed at 1.73Å (full resolution 1.655Å), which contains a binding ligand. The presence of large Z-values (indicating significant deviations from the ground-state) can be seen by the deviation from the identity line (blue, dashed) in the top-right of the plot.*



(a) Averaged (ground-state) map (1σ) *(b) Observed $2mF_o$-DF_c density in dataset x434 (1σ)* *(c) Z-map showing changes from the ground state (±3)*

Figure 3.10. ***Demonstration of Z-maps identifying deviations from the ground state.*** *(a) The ground-state map at a resolution of 1.73Å and contoured at 1σ, obtained by averaging 60 datasets, shows an ordered ethylene glycol bound in the binding site, as modelled. (b) The observed $2mF_o$-DF_c in dataset x434 shows more density in the binding site, although the binding pose of any ligand is not clear, as the density for the ligand overlaps with the density for the ground-state ethylene glycol. (c) The Z-map identifies regions where the maps are different from the ground-state. There is Z-density both above the ethylene glycol, and over one of the conserved waters (indicated with an arrow). Even though the binding of a molecule is indicated, the identity of the molecule cannot be determined for certain, nor can the pose of the binding molecule be unambiguously identified.*

3.7 Background Density Correction estimation and event map calculation

Once "*interesting events*" – groups of large Z-scores – have been identified, the next step is to identify and subtract an appropriate fraction of the ground-state map, thereby locally revealing the event density approximating the bound state. Far from the binding site, where the local structure remains in the ground state, subtracting a fraction of the ground-state map leaves the density largely unchanged; thus, the map subtraction reveals an approximation to the bound fraction of the crystal throughout the unit cell, both where the density is the same as the ground state, and where it is different.

The fractional multiplier of the ground-state map is named the Background-Density Correction (BDC) factor. The resulting *event map* is calculated as

$$[\textit{event}\ \text{map}] = [\textit{observed}\ \text{map}] - \text{BDC} * [\textit{ground-state}\ \text{map}]. \qquad 3.10$$

For identified events, the BDC is currently estimated to maximise contrast between the event map and the ground-state map. This minimises the correlation of the event map to the ground state map, relative to a normalising background factor, determined by the characterised uncertainty in the dataset. Different fractions of the averaged ground-state map are subtracted from the dataset map, and the correlation between the resulting map and the averaged ground-state map is calculated both globally and for the area around the event – defined by the blob identified in the Z-map expanded by 1Å (to increase connectivity of the selected region). The value for the BDC is obtained when the difference between these two correlation curves is maximised.

An example of this process is shown in Figure 3.11 for the event in Figure 3.10. Globally, the dataset map looks like the mean map, so plotting the global correlation against the

subtracted fraction yields a signal-to-noise curve, dropping off at a speed related to the noise in the dataset (green dashed line, Figure 3.11). Locally to the identified event, however, the dataset map is a superposition between something like the ground-state map and something that is unrelated (e.g. the density of a bound ligand). As more of the ground-state map is subtracted, the local correlation between the ground-state map and the resulting map (black dashed line, Figure 3.11) will decrease faster than the global correlation.

Subtracting the local correlation curve from the global correlation curve, BDC is estimated where the difference between these two correlation curves is maximised (blue line, Figure 3.11). This approach maximises the *contrast* between the partially-subtracted event map and the ground-state map. The final event map is calculated as in equation 3.10.

Estimation of the BDC for the example in Figure 3.10 results in a value of 0.69 (Figure 3.11). The resulting event map reveals the pose of the ligand unambiguously, and reveals that the ligand binds along with an ethylene glycol molecule that displaces a conserved binding site water (Figure 3.12). This displacement is not immediately obvious in either the standard $2mF_o$-DF_c density or the Z-map, but is clear in the event map; signal is further observed at a much higher contour level (2σ) in the event map than is required in the original observed map to observe an unclear superposition of the two states ($2mF_o$-DF_c map contoured at 1σ).

Estimation of the BDC by maximising contrast to the ground-state

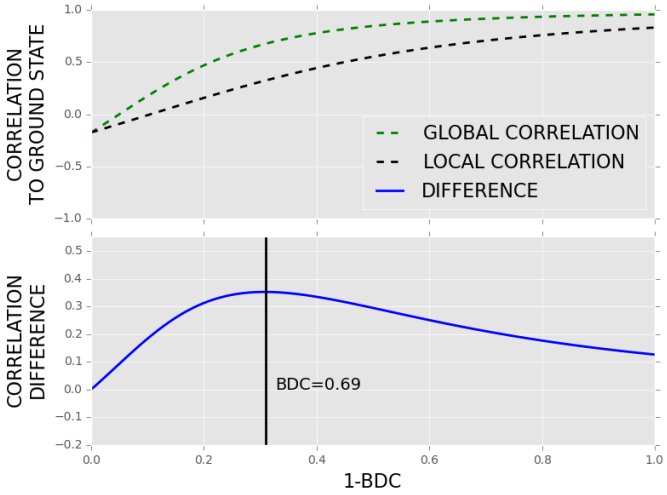
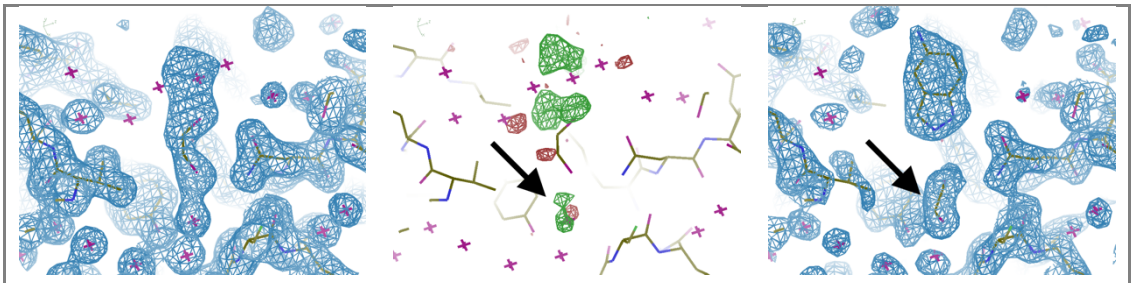


Figure 3.11. ***Example of how background density correction is estimated.*** *Estimation of BDC for dataset x434 of BAZ2B, for the region shown in Figure 3.10c. (top) Global (green, dashed) and local (black, dashed) correlations between the ground state and event maps for different values of BDC. The volumes used for the local correlations are defined by identified blobs in the Z-map, expanded by 1Å. (bottom) The BDC is estimated as the value which minimizes the local correlation compared to the global correlation; in this example, it is 0.69.*



(a) Observed $2mF_o$-DF_c density in dataset x434

(b) Z-map showing deviations from the ground state (±3)

(c) Event map (BDC = 0.69), for dataset x434

Figure 3.12. ***Demonstration of the calculation of the event map.*** *(a) The observed density (1.73Å, 1σ) indicates that something has bound, but the density is not unambiguous, allowing for interpretation, and possible misinterpretation. (b) The Z-Map (±3) gives statistical evidence that something has happened at this location in this dataset. (c) The event map (BDC=0.69, 1.73Å, 2σ) shows clear density for the bound state of the crystal; the ligand can now be confidently identified and modelled. The presence of the bound ethylene glycol (indicated with the arrows) is not immediately obvious in the standard maps in (a).*

3.8 Generation of ensemble models and refinement

The event map allows a model for the bound fraction of the crystal to be built. Globally, the event map is still largely the same as the ground-state map; atoms only need to be added/remodelled/removed where they have changed from the ground-state.

The resulting bound-state model reflects the bound copies of the protein in the crystal. The remainder of the crystal is represented by the ground-state model, known from the other datasets. As the model of the ground-state is the input to the PanDDA method, these models should be merged to form a multi-state ensemble of the crystal.

The advantages of this approach are twofold. Firstly, default incorporation of the ground-state enforces the use of prior knowledge about the crystal: that the ground state is always present in the crystal in some fraction, and should be included in refinement. The addition of the superposed ground-state model improves the model of the ligand across a range of ligand occupancies; this is demonstrated using several examples covering a range of ligand occupancies in Chapter 5. Secondly, coherent merging of distinct models allows appropriate and meaningful conformer IDs to be assigned to the multiple states. Systematic conformer labelling simplifies constrained occupancy refinement of the multiple states and enables the extraction of the individual states after refinement. For example, chemists are only interested in the bound state of the protein; the superposed ground-state model is essentially an experimental artefact. Systematic labelling of the states enables trivial removal of the superposed ground-state molecules.

After merging of the multiple states of the model, the structure can be back-transformed to the original crystallographic frame for refinement.

3.8.1 Merging of models and conformer assignment to multiple states

The process of merging the perturbed model with the ground state is shown schematically in Figure 3.13. Since the event map model is a locally edited/perturbed version of the input (ground-state) model, large parts of the structure are unaltered and

therefore identical to the ground-state model. These sections can be copied to the merged output model without change. The remaining sections of the model that are unique to either the ground-state or the bound-state model need only to be relabelled with an appropriate conformer IDs for a full model of the crystal to be built.

First, the alternate conformers that are present in the input ground-state model are identified (e.g. A and B). Atoms in the input model with a blank alternate conformer that are different in the event map model are assigned to a new unused conformer (e.g. C); atoms with pre-existing conformer IDs do not need to be changed. Secondly, all unique atoms from the event map model are re-assigned to unused conformers (e.g. " "→"D", "A"→"E", "B"→"F", etc). The relabelled atoms from both the ground-state and the event models are then transferred to the output model.

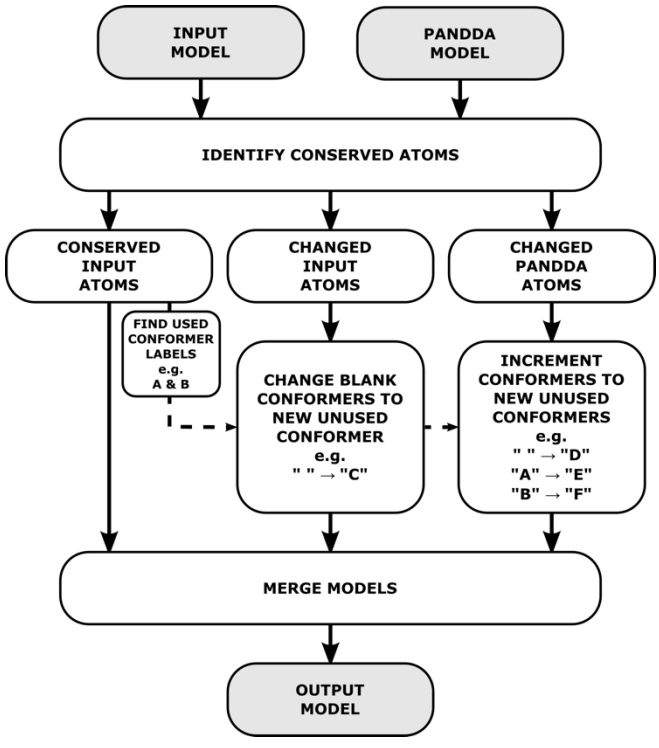


Figure 3.13. ***Schematic illustration of the processing of merging models of difference crystal states****. Conformers from the two models (input and pandda) are assigned so that different states are easily identifiable and extractable. A large part of the structures is identical; only the sections that are different need to have conformers labels added or updated.*

3.8.2 Refinement of the ensemble

Merged models still exist in the aligned *reference* coordinate frame; these models must be transformed back using to the crystallographic coordinate frame for refinement, using the alignments from section 3.4.2. As the conformers have been logically assigned to the different states, constrained occupancy refinement of the ensemble is straightforward for simple models of the protein, where the bound and unbound states are each represented by a single conformation.

To generate occupancy groups, all residues (including solvent molecules and waters) with the same conformer ID as a ligand are clustered spatially using single-linkage clustering, with a cutoff of 4Å. There may be binding events at multiple unrelated sites on the protein surface with unrelated occupancies; clustering avoids constraining distant residues to have the same occupancy.

Each local group of residues is expanded to include all alternate conformers of selected residues, as well as atoms that overlap ("clash") with one of these selected atoms (residues/atoms with a non-blank conformer ID that have an atom within 2Å of a selected atom). The resulting enlarged group of atoms is then re-partitioned by conformer ID and the occupancies of the partitions are constrained during refinement.

For more complicated superpositions of states – where either the bound or unbound states contain multiple conformations – the generated models, although simple to construct, can very quickly lead to complicated occupancy refinement situations that cannot be managed by some refinement programs. We have tried refining models with both REFMAC (Murshudov et al. 2011) and phenix.refine (Afonine et al. 2012), but in both cases it is possible for the occupancy of residues to refine to greater than unity.

Simplification of the occupancy groups and/or the model is therefore sometimes necessary to obtain valid refinements. Furthermore, the merging of models as described above can result in discontinuities in the protein chain, where a residue with conformers A & B connects to a residue with conformers C & D; there is no clear method for determining how to connect these two residues, and models may need to be further edited when this occurs. The correct management of these situations and the automated generation of occupancy groupings are the subject of further work, involving the utilisation of tools such as autoBUSTER, which are capable of handling such situations (Bricogne et al. 2011; Smart et al. 2012).

Refinement of the ensemble model of the protein is performed using occupancy-constrained groups alongside conventional resolution-dependent refinement protocols; when the input-model to PanDDA is near-complete, only a short refinement is typically needed to obtain a good ligand model.

3.9 Expanded ligand validation

Although the event maps provide clear visual indication of the binding of the ligand, ligand models must still be validated after refinement to ensure the non-visual parameters, the B-factor and the occupancy, have refined correctly. Furthermore, for very weak ligands, there will be no visual indication of the bound ligand in the full-dataset $2mF_o$-DF_c maps at conventional contour levels (and an ambiguous multi-state superposition at low contour levels); we must instead rely on the event map and the model validation scores described below to confirm the binding of the ligand.

It is well established that the commonly-used real-space correlation coefficient (RSCC) is not sufficient on its own to determine the validity of a crystallographic ligand

(Weichenberger et al. 2013; Deller & Rupp 2015). Furthermore, a low RSCC gives no information as to where the errors in a model lie. Therefore, I apply a novel combination of five validation metrics to the generated ligand models, combining conventional density metrics, new density metrics and consistency metrics; these metrics are shown in Table 3.4, and described in the introduction chapter.

The use of multiple density metrics allows scoring of distinct aspects of the model: the RSCC measures the overall agreement between the ligand and the $2mF_o$-DF_c density; the RSZD measures the significance of difference density over the model; and the RSZO measures the density strength over the model. The cutoff value of 0.7 for RSCC is a conventional correlation cutoff in statistics, representing a "strong" correlation between the observed data and the model. RSZD is a typical Z-score, so a conventional cutoff for significance of 3 is used: models must exhibit an RSZD of less than 3 to be considered "good". The RSZO is proportional to the average density over the model, and is therefore directly proportional to occupancy; dividing through by the occupancy gives a value of RSZO as if it were present at full occupancy in the crystal (further discussed in Chapter 5). Once on an absolute scale, RSZO/OCC is a density signal-to-noise ratio, for which a conventional cutoff is 2.

The B-factor ratio represents consistency with the environment; larger values than ≈2 indicate issues with the model – such as an incorrect occupancy – or absence of the model. Simultaneous inspection of the normalised RSZO and B-factor ratio allows meaningful determination of whether the refined occupancy is reasonable (e.g. Figure 3.14b). Measuring the RMSD of the ligand coordinates before and after refinement ensures that the pose of the ligand is stable during refinement. A value of RMSD which

constitutes a significant deviation, 1Å, was identified by manual inspection of the models; up to a 1Å movement of the ligand after refinement from the initial fitted model provides satisfactory discrimination between good and bad refinements.

3.9.1 Residue validation plots

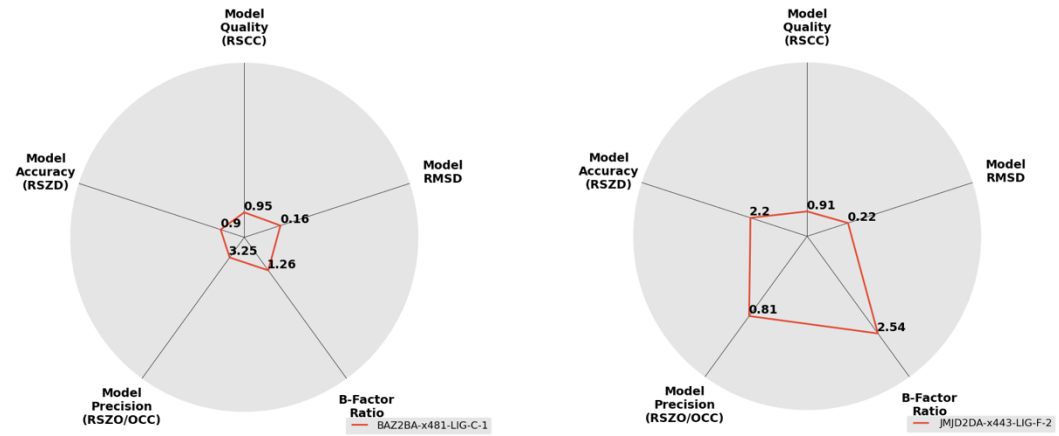
To provide a visual analysis of the quality of a ligand model, the validation metrics are presented as a "radar plot" (Figure 3.14). These plots show when the ligand validation scores deviate from the ideal/preferred values. The values at which scores begin to move from the centre of the plot, and the values at which they reach the edges of the plot are shown in Table 3.5. A high-quality model is shown in Figure 3.14a, and a model with poorly-refined occupancy and B-factors is shown in Figure 3.14b.

Table 3.4. ***Preferred values of ligand validation scores****, as defined in the main text.*

Metric	Description	Good values
RSCC	Agreement between model and data	> 0.7
RSZD	Statistical quantification of difference density	< 3
RSZO / OCC	Measure of signal-to-noise of observed density	> 2
B-Factor Ratio	Consistency of model with surrounding residues	≈ 1
RMSD	Movement of ligand model under refinement	< 1Å

Table 3.5. ***Ranges for density metrics on the validation plot.*** *The minimum value defines the value at which the line begins to move away from the centre of the plot. Values smaller than the minimum value are plotted as the minimum value. The maximum value defines the value at which the line reaches the edge of the plot. Observed values larger than the maximum continue to be plotted outside of the plot. Inverted axes swap the minimum and maximum values, so that metrics where large values are preferable (RSCC, RSZO) are shown similarly to the other metrics.*

Metric	Minimum Value	Maximum Value	Inverted Axis
RSCC	0.60	0.85	**Yes**
RSZD	1.50	4.00	No
RSZO / OCC	0	2.00	**Yes**
B-Factor Ratio	1.00	3.00	No
RMSD (Å)	0	1.50	No



(a) *(b)*

Figure 3.14. ***Example validation plots for two ligands.*** *(a) A high-quality ligand model in dataset BAZ2B-x481. All metrics score well, so the line remains close to the centre of the plot. (b) A ligand in dataset JMJD2D-x443 with possible refinement errors. The density for the ligand is weak (RSZO/OCC < 1), and the ligand has large B-factors (B-factor ratio > 2). This indicates that the refined occupancy of the ligand may be too high, and that the B-factors of the ligand are being inflated to compensate, demonstrating the problems that can arise in the simultaneous refinement of B-factors and occupancy; the validation plots can be used to highlight when such errors in refinement occur.*

3.10 The PanDDA implementation

The PanDDA method, as described as above, is implemented in Python, making extensive use of libraries from CCTBX (Grosse-Kunstleve et al. 2002). The main command line tool is called pandda.analyse. Further command line tools are also included for modelling and validation. It is freely available via pip (package name *panddas*), and the source code can be downloaded from *http://bitbucket.org/pandda*. A tutorial and help are available at *http://pandda.bitbucket.org*. It will also soon be distributed within CCP4 (Winn et al. 2011). The implementation has been evolved through extensive interaction with users of the fragment screening facility at Diamond Light Source Beamline i04-1.

3.10.1 Calculation of Z-maps and event maps: pandda.analyse

The top-level workflow of the PanDDA implementation is shown in Figure 3.15. Alignment of the structures and the crystallographic maps is performed as in section 3.4. A set of datasets typically covers a wide range of resolutions, and maps may only be compared at the same resolution; analysis is thus performed at multiple resolutions. Analysis begins at the highest resolution where a minimum number of datasets (e.g. 30) are available, and from here analysis progresses with a default resolution step of 0.5Å.

For example, for analysing datasets from 1.2⟶1.25Å, any dataset with a resolution between 1.2Å and 1.25Å will be analysed, but datasets with resolution higher than 1.2Å will also be used for determination of the statistical maps (i.e. averaged ground-state map and adjusted-variation map). All datasets will be truncated to 1.25Å and the set of common reflections, as described in section 3.4.3, and analysed. The next iteration of the analysis would then be performed on datasets from 1.25⟶1.3Å.

Significant blobs in the Z-maps of the datasets are identified as described in section 3.6 and the event maps are generated as in section 3.7. Identified events are clustered spatially to form *sites*; these are presented to the user for modelling. Sites are ordered by the number of events; events can be ordered by either the size of the Z-blob or the height of the Z-peak, the largest Z-value in the identified blob.

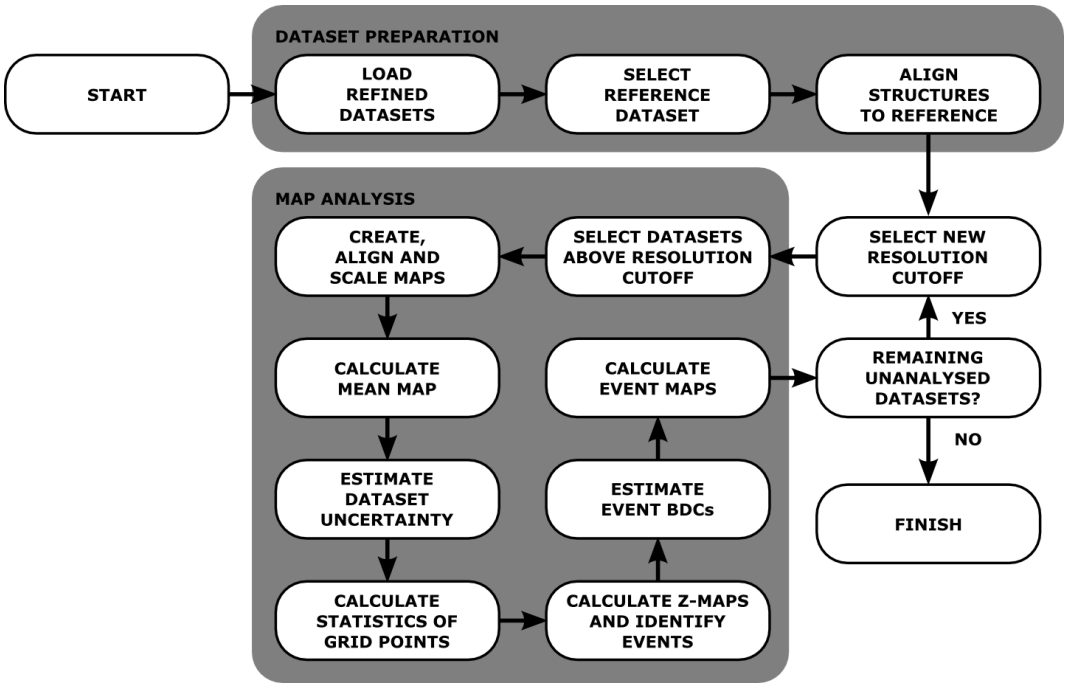


Figure 3.15. ***Top-level algorithm of the implemented PanDDA method.*** *The input to the PanDDA method is a series of refined datasets, which must contain similar structures. These structures are aligned, and these alignments are used to align the electron density maps for comparison. Datasets are analysed in resolution shells, with high resolution datasets being used multiple times to characterise lower-resolution shells. This iteration through resolution shells continues until there are no remaining unanalysed datasets.*

3.10.2 Modelling of identified events: pandda.inspect

Modelling of the identified events is performed in the reference coordinate frame, as the generated maps are not easily represented in the native crystallographic coordinate frame (see section 3.4.4). A graphical tool has been implemented within Coot (Emsley et al. 2010) to enable the easy navigation through the results (Figure 3.16). This tool allows the user to record structured meta data about the modelled ligands, such as the confidence in the generated model, as well as free-text comments.

The modelling philosophy within the PanDDA paradigm is as described in section 3.8. In the event maps, an approximation to the putative *changed-state* density is shown, whilst the input model to PanDDA should represent the *ground state* (modelled from a truly ligand-free reference dataset).

Therefore, in pandda.inspect, only a model which represents the new *changed state* needs to be built: *the user should only model what is seen in the event map*. This will involve the removal/re-modelling of *ground-state* atoms that have moved or disappeared in the bound state, and the placement of any new atoms and ligands.

After modelling of the changed state, the changed-state model and the ground-state model are merged for refinement (as in section 3.8) and back-transformed to the crystallographic coordinate frame using *pandda.export*. Occupancy groupings (generated as in section 3.8.2) are automatically output for REFMAC (Murshudov et al. 2011) and phenix.refine (Afonine et al. 2012).

3.10.3 PanDDA results summaries

The PanDDA implementation also produces several HTML summary pages so that the results can be easily presented to users. Produced summaries include: an initial summary of the parameter distributions across datasets (Figure 3.17); a summary of the identified events, grouped by site (Figure 3.18); and a summary of the modelling performed in pandda.inspect (Figure 3.19). The modelling GUI within Coot (Figure 3.16), can be used to access these summary pages, which are updated as modelling progresses.

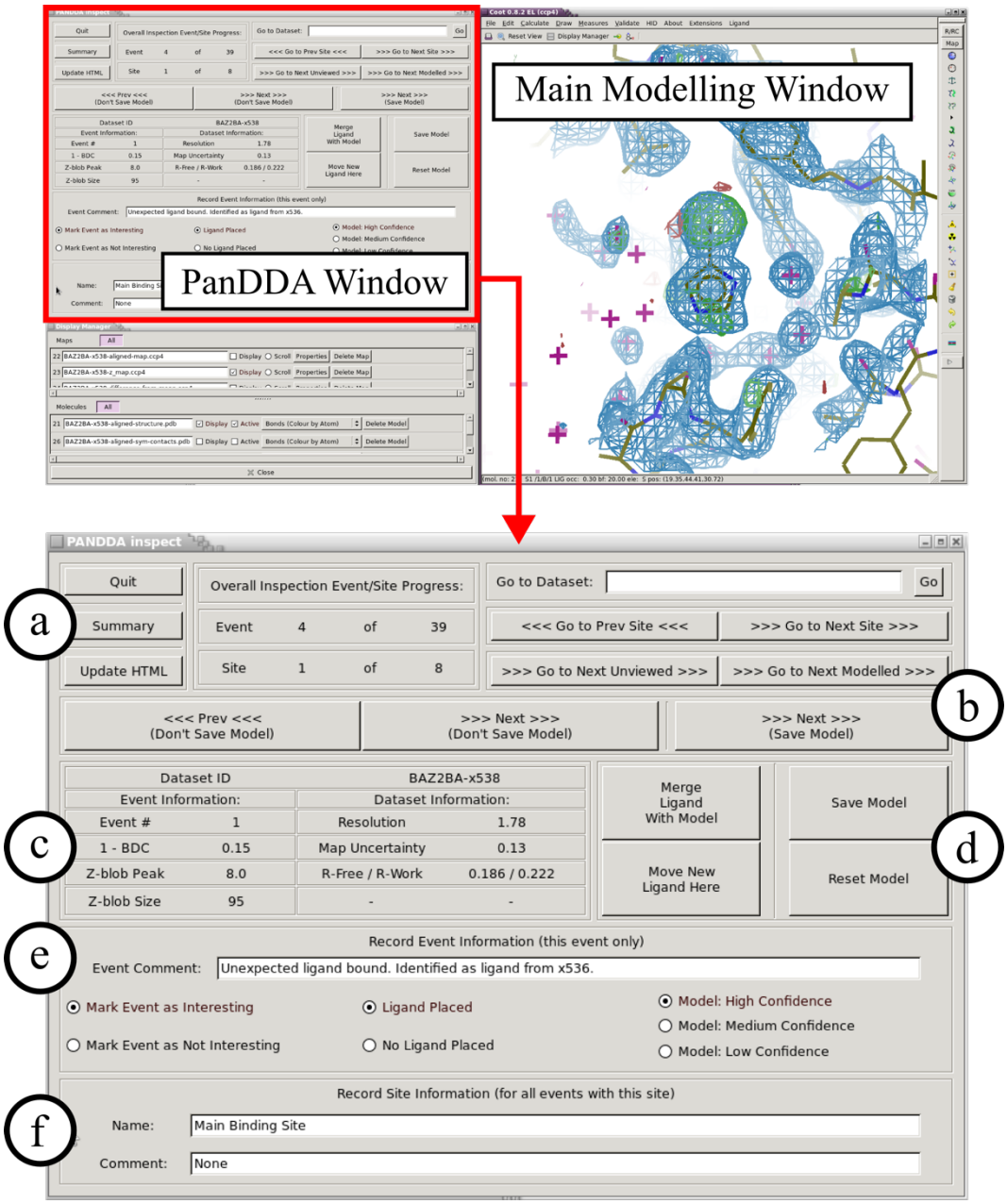


Figure 3.16. ***The PanDDA modelling GUI implemented within Coot.*** *The standard Coot windows are supplemented by the additional window (highlighted in red) that allows the user to navigate through the results. As the user navigates through the datasets, all relevant structures and maps are automatically loaded, and the view is centred on the selected event. Sections of the GUI are labelled. (a) Buttons to allow access to summary pages of the analysis (see section 3.10.3). (b) Navigation buttons allow easy movement between datasets. Navigation can be done on a site-by-site basis, or by accessing only modelled, or unviewed, datasets. (c) Summary of the currently loaded event and dataset. (d) Modelling buttons allow the user to perform standard tasks such as merging the ligand with the protein model. (e) Meta-data section for the loaded event. Standard radio buttons allow the recording of structured data, and a free-text comment box allow more detailed descriptions. (f) Comment boxes for the site (group of events), allow the naming of regions of the protein. These comments are associated with all events at this site.*

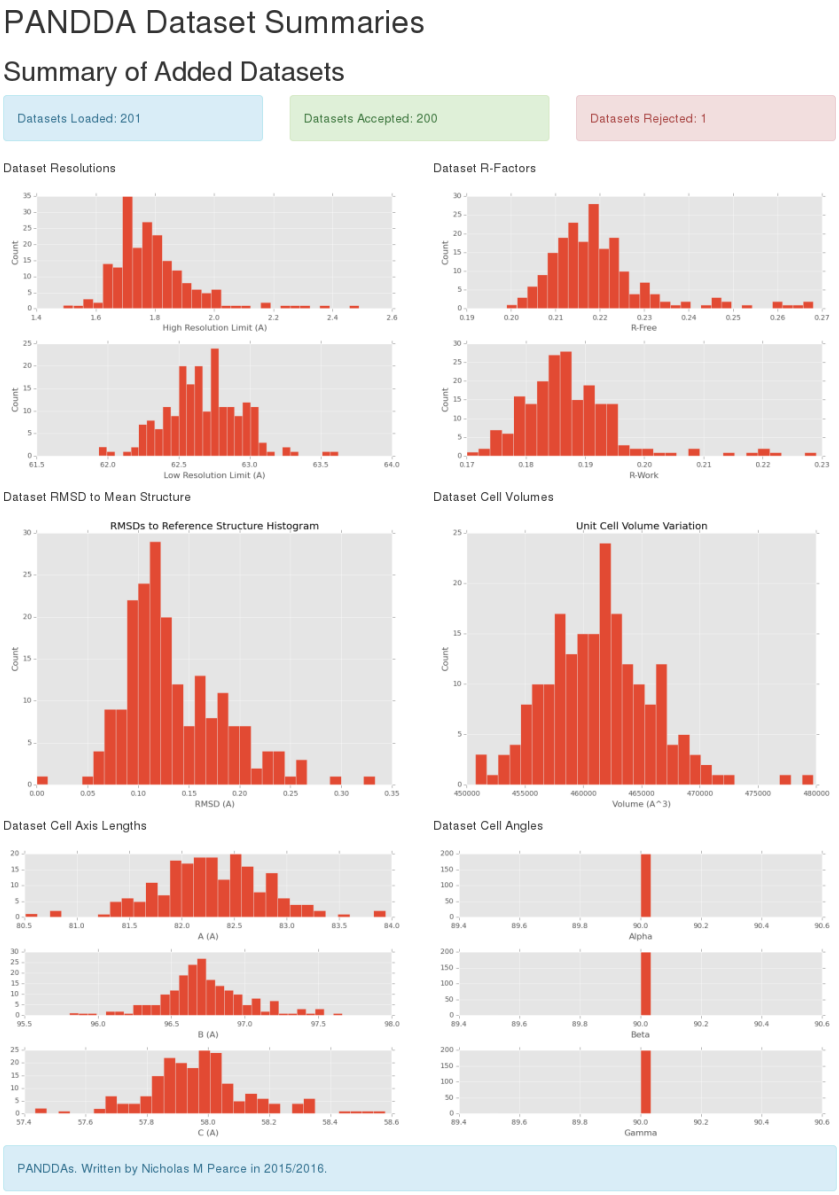


Figure 3.17. ***Dataset summary page from PanDDA.*** *Histograms of the dataset parameters (for analysed datasets, after removing low-quality or datasets from difference crystal forms) reveal the variation across the analysed crystals and datasets.*

Figure 3.18. ***Processing summary from PanDDA.*** *The bars at the top of the page give summary statistics of the analysis. The images below show the distribution of identified areas (sites) in the context of the protein and a site-by-site summary of the identified events. Finally, a detailed table of summaries of each dataset is included.*

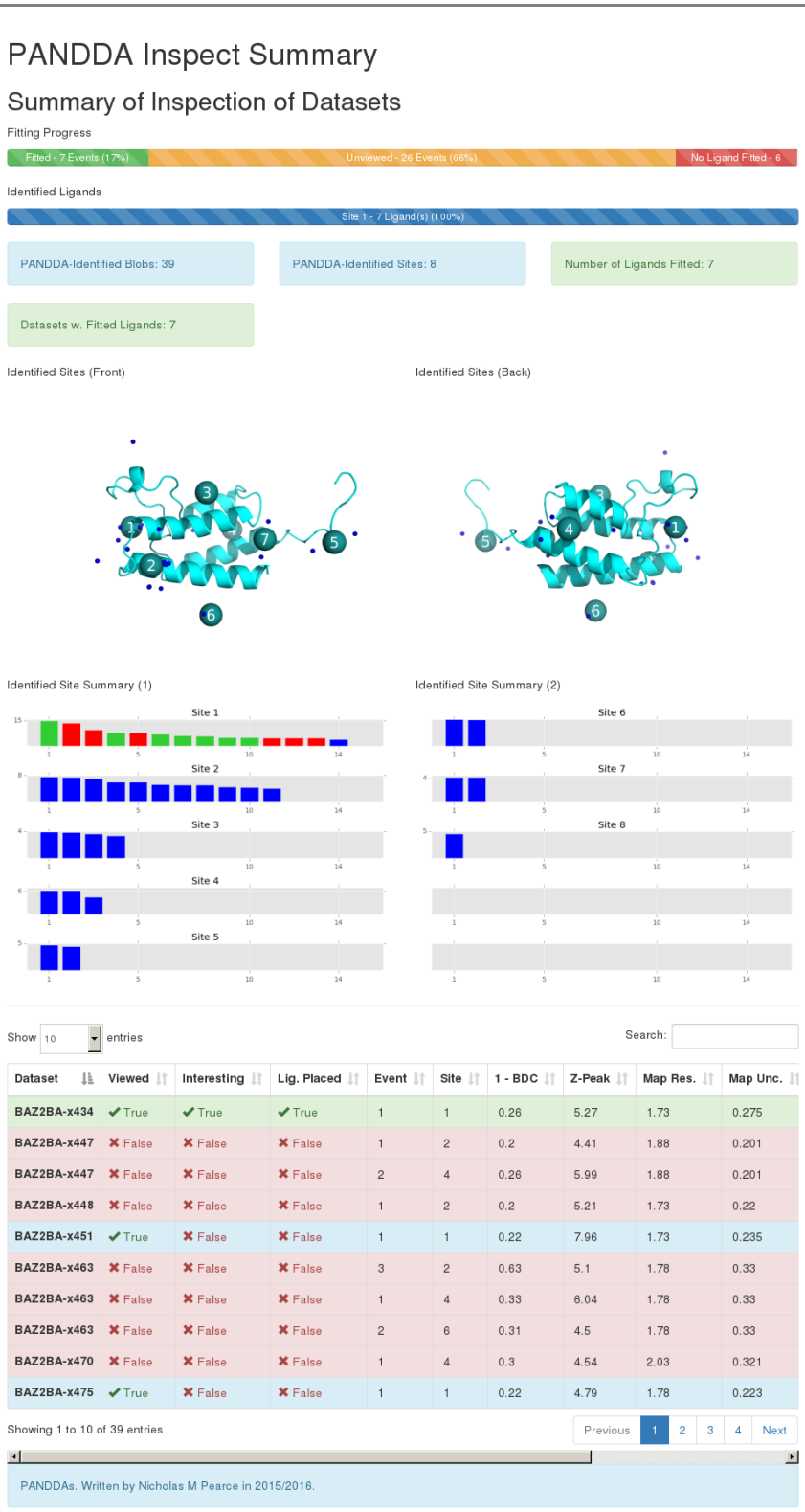


Figure 3.19. ***Modelling summary from PanDDA.*** *The bars at the top of the page show the progress of the modelling, and a summary of the modelling. The images below again show the distribution of identified areas (sites) in the context of the protein and a site-by-site summary of the identified events. Here, the bars are coloured depending on the modelling outcome of the identified event: green, modelled; red, viewed but un-modelled; blue, un-viewed. Finally, a detailed table of summaries of each event is included.*

3.11 Chapter Summary

In this chapter I have described the PanDDA method, as well as its current implementation. Binding ligands – and any other changed-states – are identified objectively by comparing single datasets to an ensemble of unbound (ground-state) datasets; the Z-maps are statistical measures and can be used to confidently determine whether a deviation is “real” or “noise”, unlike in conventional maps. The estimation of the BDC factor and corresponding subtraction of the superposed ground-state density removes the ambiguity created by the crystallographic superposition of states, improving the interpretability of the data and greatly simplifying the modelling step.

In the next chapter, I will apply the PanDDA method to a series of fragment screening datasets, leading to a significant increase in the number of identified binding fragments, as well as much clearer indications of binding; the PanDDA method makes it possible to confidently model low-occupancy crystallographic features.

Chapter 4

Results from the application of the PanDDA method to four crystallographic fragment screening datasets

---

"Modelling is going to take forever; I wish there were fewer binders."

In the previous chapter, I described the PanDDA approach to changed-state identification. Here, I apply the PanDDA method to identify bounds ligands in four crystallographic fragment screening datasets.

In these fragment-screening experiments, hundreds of crystallographic datasets were collected to find fragments which bind to the proteins; since fragments are composed of only a few atoms, often bind only weakly to the protein, and may bind at multiple sites on the protein surface, fragment screening experiments represent the extreme of crystallographic ligand-identification challenges. Minimal difference density may be presented by binding ligands, and where partial-occupancy binding is present – as it likely invariably is – the observed density will remain a superposition of states, increasing the likelihood of misinterpretation of the experimental electron density.

The PanDDA approach overcomes both issues and further results in higher hit rates than were obtained through a conventional analysis of the data. The clarity of the PanDDA maps lead to higher confidence in identified ligands, even for low-occupancy binders,

and the ensemble modelling approach leads to models that are generally of high quality, as measured by the multiple validation metrics described in Chapter 3.

4.1 Data and Methods

Four collections of fragment screening datasets are analysed in this chapter. These data were collected as part of the SGC-Diamond fragment screening collaboration, by other scientists at the SGC Oxford and beamline staff at Diamond Light Source i04-1.

The experimental procedures for crystal preparation and data collection vary considerably between the datasets and are not covered in depth here; the principal summary data for each of the crystal forms analysed is shown in Table 4.1. All data was automatically processed at diamond with XDS (Kabsch 2010), pointless (Evans 2006) and aimless (Evans & Murshudov 2013). All datasets are single-compound soaked crystals. The analysis of the data covered here begins with the merged diffraction data.

4.1.1 Data preparation, processing and analysis

Free-R flags are transferred to each dataset from a reference dataset using *cad* (CCP4; Winn et al. 2011). Missing miller indices are then added with *uniqueify* (CCP4); missing free-R values are populated and missing observed structure factors are filled with *NA* values. Observed data set to *NA* causes the corresponding $2mF_o$-DF_c values to be set to DF_c values after refinement with REFMAC (to make maps comparable; Chapter 3). Datasets are refined with Dimple (CCP4) using a high-resolution ground-state structure.

The resulting refined structures were then processed and modelled with the PanDDA implementation using the default parameters as described in the previous chapter. Additional modelling of any missing ground-state atoms was guided by density of a ground-state dataset. Events which could not be confidently modelled were ignored.

Table 4.1. ***Summaries of the datasets in the analysed fragment screens.*** *R-free and R-work are calculated after refinement of the reference model with Dimple for each dataset.*

Protein	**JMJD2D**			**BAZ2B**		
Protein Name	**Lysine-specific demethylase 4D**			**Bromodomain adjacent to zinc finger domain 2B**		
Crystallisation Conditions	0.1M HEPES, pH 6.8-7.2, 0.2-0.25M ammonium sulphate, 25-29% PEG3350			0.1M MES, pH 6.0, 30% PEG600		
Cryo-Protectant (conc.)	Ethylene Glycol (25%)			Ethylene Glycol (25%)		
Solvent (conc.)	DMSO (10-30%)			Ethylene Glycol (25%)		
Soaking Time	50mins			1h		
Resolution Range (Å)	1.1-2.6			1.5-2.5		
Mean Resolution (SD) (Å)	1.45 (0.21)			1.79 (0.15)		
Space Group	P 43 21 2			C 2 2 21		
Mean Unit Cell Axes (Å)	71.42	82.17	96.57	58.03	71.42	150.41
Unit Cell Axes SD	0.29 %	2.02 %	2.31 %	1.51 %	0.29 %	0.25 %
Unit Cell Volume SD	0.80 %			3.03 %		
R-free quartiles	0.178	0.181	0.186	0.212	0.218	0.224
R-work quartiles	0.153	0.156	0.159	0.182	0.186	0.191
RMSD(bonds) quartiles (Å)	0.024	0.026	0.028	0.023	0.024	0.026
RMSD(angles) quartiles (Å)	2.31	2.44	2.52	1.92	2.00	2.08
NCBI Gene ID	55693			29994		
Domain Range	JmjN 18-64, JmjC 182-295			1871-1955		
Domain Category	JmjN, JmjC (Jumonji)			Bromodomain		

Protein	**SP100**			**BRD1**		
Protein Name	**SP100 nuclear antigen**			**Bromodomain containing 1**		
Crystallisation Conditions	0.1M MES, pH 6.1, 20% PEG20K			0.1M bis-tris, pH 7.0, 30% PEG3350		
Cryo-Protectant (conc.)	Ethylene Glycol (30%)			Ethylene Glycol (30%)		
Solvent (conc.)	Ethylene Glycol (30%)			Ethylene Glycol (30%)		
Soaking Time	12h			2-4h		
Resolution Range (Å)	1.3-2.7			1.5-3.6		
Mean Resolution (SD) (Å)	1.72 (0.22)			1.76 (0.33)		
Space Group	C 1 2 1			P 21 21 21		
Mean Unit Cell Axes (Å)	127.67	45.39	83.36	55.46	56.42	101.76
Unit Cell Axes SD	0.09 %	0.21 %	0.20 %	0.75 %	0.30 %	0.27 %
Unit Cell Volume SD	0.50 %			1.05 %		
R-free quartiles	0.203	0.207	0.213	0.215	0.223	0.238
R-work quartiles	0.169	0.173	0.176	0.181	0.187	0.199
RMSD(bonds) quartiles (Å)	0.019	0.021	0.024	0.019	0.021	0.023
RMSD(angles) quartiles (Å)	1.78	1.88	2.05	1.73	1.87	2.01
NCBI Gene ID	6672			23774		
Domain Range	PHD 704-747, Bromo 773-874			570-654		
Domain Category	PHD, Bromodomain			Bromodomain		

4.1.2 Assumptions for the application of PanDDA to fragment screening data

For the initial analysis of fragment screens, all datasets are considered as ground-state datasets. Since the binding of ligands is weak, rare and binding sites occur at multiple locations in the unit cell, averaging over multiple datasets is assumed to be a good approximation of the ground state, as any accidentally-included bound datasets will have minimal influence at a site when large numbers (>30) are averaged.

If the hit rate is high enough that averaging over multiple datasets is considered not to accurately reflect the ground-state, it is possible to run an iterative form of PanDDA, where identified "interesting" datasets are removed from the pool of ground-state datasets, and the analysis repeated. The hit-rates of the analyses presented here were considered low enough, and the event maps clear enough, that iterative processing of the datasets was not performed.

In hindsight, however, there are indications that this assumption was not correct: for the BRD1 fragment screening campaign (section 4.6), some ligands were identified in one NCS copy automatically, but then further identified in the second NCS copy by manual inspection of the Z-maps. Though this may be due in part to the method used to analyse and identify significant blobs in the Z-maps, some of the missed ligands are high occupancy (Table 4.7), and this indicates that the presence of binding ligands has artificially increased the derived s_m values in the binding sites, which then serves to suppress signal from binding ligands.

Reanalysis of the data with bound datasets removed may therefore increase the hit rate further through enhanced signal in the Z-maps; however, the clarity of the event maps attests to the robustness of the averaged ground-state map and the method in general.

4.2 Results

The summary of identified hits from both conventional analysis and after PanDDA analysis is shown in Table 4.2. The conventional analysis of the data was performed by several crystallographers at the SGC Oxford, using a combination of RHOFIT (Womack et al. 2010) and manual inspection of the difference density. Approaches vary within experiments, and no consistent protocol was followed.

The hit rates are significantly increased after application of the PanDDA approach, both in the case of an initially high hit-rate (BRD1), and where no hits were confidently identified (SP100). The following sections discuss the results of ligand identification in each fragment screen in more detail.

Table 4.2. ***Hit rates before and after the application of the PanDDA method.*** *Human identification was performed by a manual inspection of the binding sites of the protein using the standard crystallographic maps. An automated ligand fitting program (RHOFIT) was also applied to the BAZ2B and JMJD2D datasets to detect hits. A "site" is defined as a region of the protein that binds two fragments that do not interact heavily with crystal contacts.*

Protein		SP100	BAZ2B	JMJD2D	BRD1
Datasets		116	200	226	292
Resolution Range (Å)		1.3-2.7	1.5-2.5	1.1-2.6	1.5-3.6
Identified Hits	(*Human* / **PanDDA**)	*0* / **2**	*3* / **9**	*2* / **24**	*29* / **40**
Identified Hit Rate (%)	(*Human* / **PanDDA**)	*0* / **1.7**	*1.5* / **4.5**	*0.9* / **10.6**	*9.9* / **13.7**
Identified Sites	(*Human* / **PanDDA**)	*0* / **1**	*1* / **1**	*1* / **5**	*1* / **2**

4.2.1 Data Availability

All datasets for the four fragment screens have been uploaded to repositories on Zenodo (https://zenodo.org). All refined ligand-bound structures, Z-maps and event maps are uploaded, as well as all merged diffraction datasets, so that the PanDDA may be repeated by others. DOIs for the uploaded screens are shown in Table 4.3.

Table 4.3. ***DOIs for fragment screening datasets upload to Zenodo repositories.***

Protein	DOI	Protein	DOI
SP100	10.5281/zenodo.48771	**JMJD2D**	10.5281/zenodo.48770
BAZ2B	10.5281/zenodo.48768	**BRD1**	10.5281/zenodo.48769

4.3 Nuclear auto-antigen SP-100

The bromodomain of nuclear auto-antigen SP-100 (SP100) was screened against a selection of compounds from the Maybridge fragment library (www.maybridge.com), resulting in 116 datasets. From the initial difference-density based inspection, three tentative ligand models were identified, but with no confidence in the generated models: both the $2mF_o$-DF_c and the mF_o-DF_c difference density were unconvincing.

4.3.1 Rejection of mismodelled compounds

Upon application of the PanDDA method, none of the modelled sites showed any evidence of deviation from the ground-state; one model was a misinterpreted ground-state HEPES molecule. The PanDDA approach clarified the *absence* of binding signal.

4.3.2 Confident identification of weak binders

However, the PanDDA application also identified and enabled the modelling of two ligand species (three copies in total); one instance is shown in Figure 4.1. There is no evidence of the ligand in the standard maps, but the ligand is clearly visible in the PanDDA maps; the ligand is bound to both copies of the protein in the asymmetric unit.

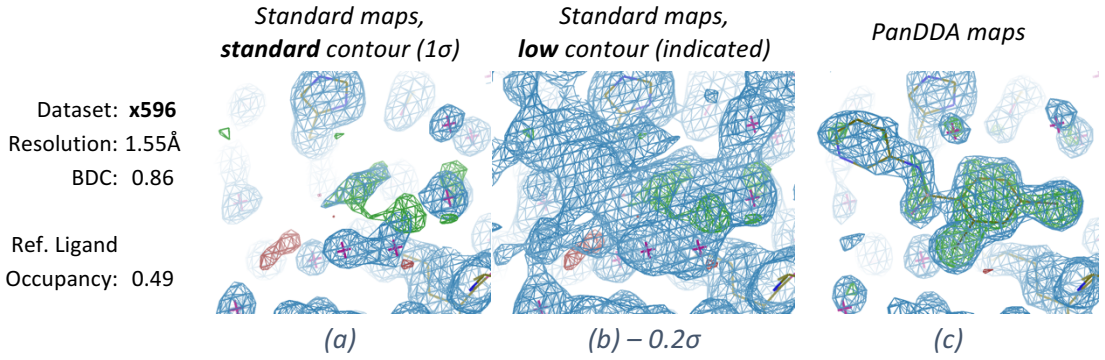


Figure 4.1. ***SP100: ligand identification example.*** *Dataset statistics are as labelled; the first two columns show the $2mF_o$–DF_c map (blue) contoured at the different levels indicated, and mF_o–DF_c map (green/red) contoured at ±3σ (both maps from Dimple). The third column shows the PanDDA event map (blue) contoured at 2σ, and Z-Map (green/red) contoured at ±4.*

Dataset x596: *Conventional maps show evidence for nothing but solvent. PanDDA maps show clear density for a weakly bound ligand.*

4.4 Bromodomain adjacent to zinc finger 2B

Bromodomain adjacent to zinc finger 2B (BAZ2B) was screened against fragments from the *Zenobia* library (www.zenobiafragments.com), resulting in 200 datasets. Originally, three ligands were identified in the main binding site (Table 4.2). After PanDDA, six further ligands were identified at this site; two examples are shown in Figure 4.2.

4.4.1 Unambiguous identification of mislabelled/misdispensed ligands

In dataset x538, the supposed ligand – the ligand that was thought to have been added to the crystal – does not match the density in the event map (Figure 4.2d-f, Figure 4.3). However, the ligand of a nearby dataset (dataset x536) has an excellent fit to the density, and contains a bromine which occupies the region of strong density in the event map. Applying Occam's razor, a dispensing error or a labelling error has likely occurred in the experiment. The presence of the bromine was confirmed by anomalous difference map (Figure 4.3d); the diffraction data were collected at Diamond, beamline i04-1, which is tuned to the K-edge of bromine.

4.4.2 Reordering of sidechains upon ligand binding

Further to the nine binding site ligands, the binding of a ligand in the crystal contacts of dataset x603 causes the reordering of several sidechains of the protein (Figure 4.4). These movements, though visible in the conventional crystallographic maps, are clearer in the PanDDA maps. The fact that this is a ligand-induced re-ordering, rather than a mismodelled ground-state sidechain, is further immediately evident from the Z-map, which shows signal at a very high contour level of ±5.

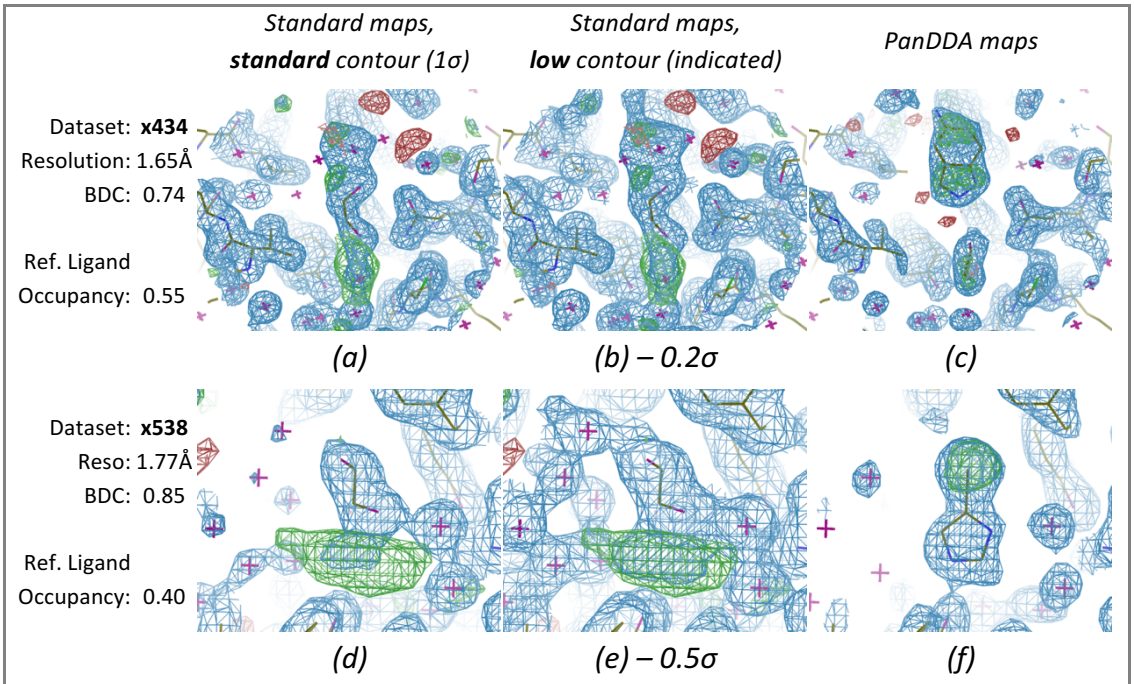


Figure 4.2. ***BAZ2B: ligand identification examples.*** *Maps and contour levels are as in Figure 4.1, except where indicated.*

Dataset x434: *PanDDA maps show clear density for the ligand as well as an ethylene glycol that displaces a conserved binding site water.*

Dataset x538: *PanDDA maps show clear density which does not however match the shape of the labelled ligand; the clarity of the event map enabled the identification of the modelled ligand by its chemical connectivity, and the presence of a bromine (subsequently confirmed by calculation of an anomalous difference map; Figure 4.3). The refined occupancy of 0.40 is for the non-bromine atoms only (bromine refined occupancy 0.24).*

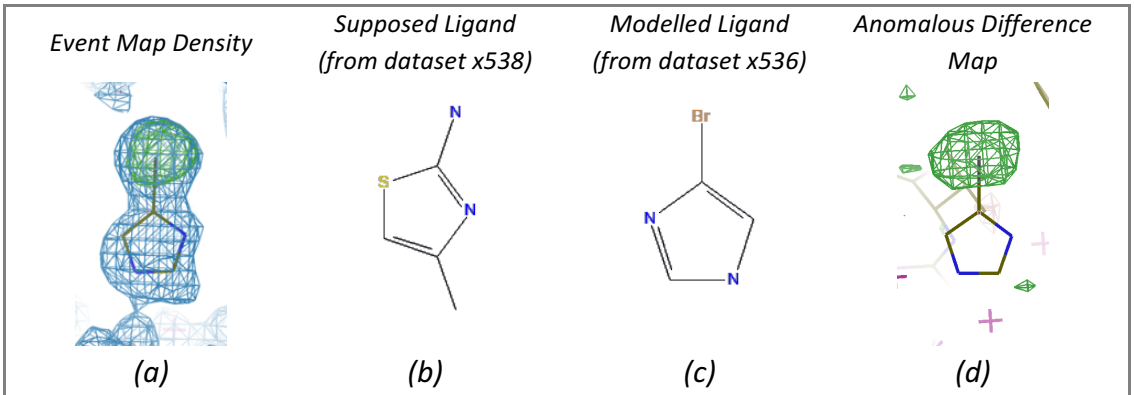


Figure 4.3. ***BAZ2B: "unexpected ligand" identification.*** *(a,b) The density in the event map for dataset x538, contoured at 2σ, clearly shows that the ligand thought to be in the crystal, shown in (b), is incorrect. The sulphur would need to be placed at the top of the image to occupy the region of strong density, leading to a poor fit to the density. (c) Instead, the ligand can be uniquely identified as being from a nearby crystal (the ligand from dataset x536), as modelled in (a). (d) The presence of the bromine in this ligand can be confirmed by anomalous difference map (contoured at ±3σ), through absorption at the K-edge (collected at DLS i04-1).*

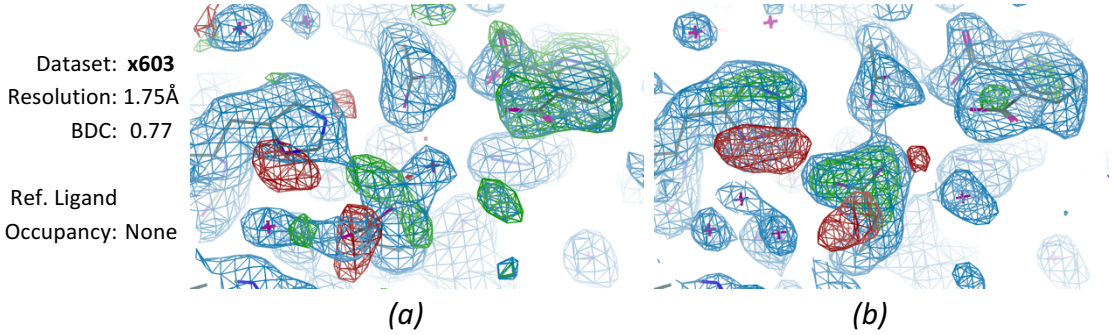


Figure 4.4. ***BAZ2B: sidechain movement example.*** *Maps and contour levels are as in Figure 4.1, except where indicated. The binding of the ligand (top right of images) in dataset x603 of BAZ2B (resolution 1.749Å) causes the reordering of the sidechains of nearby residues. The two largest movements are the histidine and the glutamate (highlighted by the Z-map). (a) Some evidence for the movement of the residues is seen in the standard maps, highlighted by the presence of difference density, but this could have arisen from a mismodelled ground-state sidechain. (b) The movement of the sidechains in the event map is clear and unambiguous, and the deviation from the ground-state is indicated strongly by the Z-map (contoured at ±5). This ligand was not refined as it binds in the crystal contacts of the protein.*

4.5 Lysine-specific demethylase 4D

Lysine-specific demethylase 4D (JMJD2D) is a lysine demethylase from the KDM4 family. This crystal form was soaked in the presence of the co-factor analogue N-oxalylglycine (NOG) tightly bound to the metal in the main binding site. No binders were observed to bind in main binding site from the initial analysis, and only two allosterically binding ligands were detected (Table 4.2). After PanDDA, 22 additional binders were identified, both in the main binding site and covering the surface of the protein; another instance of one of the previously-modelled binding ligands was also identified.

4.5.1 Opportunistic binding in the main binding site

Consistently, refinement of ground-state crystals indicates both metal and NOG are not present at full occupancy (refined occupancies of metal and NOG ≈90%), and one fragment evidently exploits this by binding to empty sites in the crystal; this low-occupancy event is clearly revealed by PanDDA (Figure 4.5).

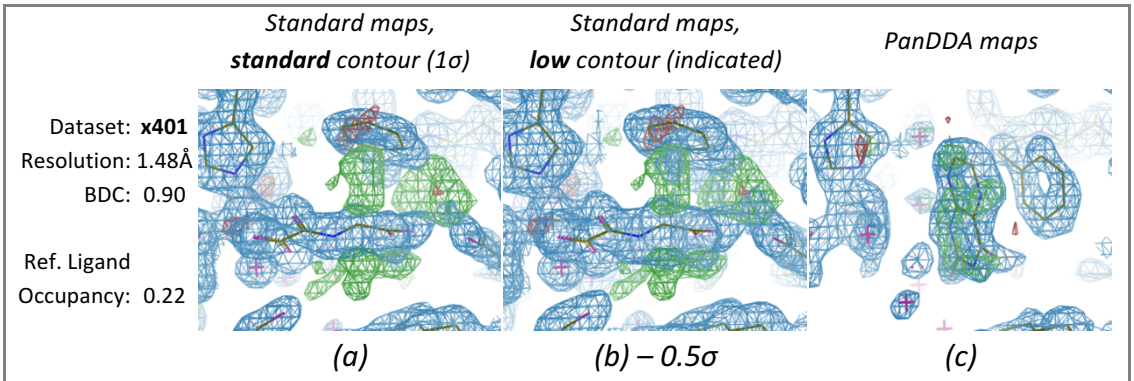


Figure 4.5. ***JMJD2D: opportunistic weak binding of fragment to unoccupied binding sites.*** *Maps and contour levels are as in Figure 4.1, except where indicated.*

Dataset x401: *Conventional maps are dominated by the NOG co-factor analogue bound in the majority fraction of the crystal, whereas event maps (BDC=0.9) and Z-maps unambiguously reveal both the ligand and the associated changes in protein conformations (movement of the phenylalanine). The difference density in (a) and (b) is not useful for model building. The alternate phenylalanine conformer is also visible in several other datasets, indicating variability between the datasets, and explaining the low Z-scores in this region.*

4.5.2 Unambiguous identification of atomic connectivity

Analysis of the JMJD2D datasets also revealed unexpected ligands that appeared in the "wrong" dataset, similarly to the BAZ2B example described above (section 4.4.1). In the example in Figure 4.6, it is unclear from the standard crystallographic maps: whether a ligand is bound at all; what the identity of the bound molecule is; what the pose is; and whether any bound ligand is present in multiple conformations.

The event map clearly reveals the atomic connectivity of the bound ligand, but the absolute identity of ligand cannot be determined. The ligand from another dataset can be modelled into the density, but the bound ligand could also be a reactant or a hydrolysed form of the supposed ligand (Figure 4.7).

Although the PanDDA maps cannot uniquely identify the ligand, they are able to unambiguously identify the atomic connectivity of the bound ligand, and show that it is bound in one conformation. Testing of the compounds and repetition of the experiment would be required to identify for certain which species of ligand is bound.

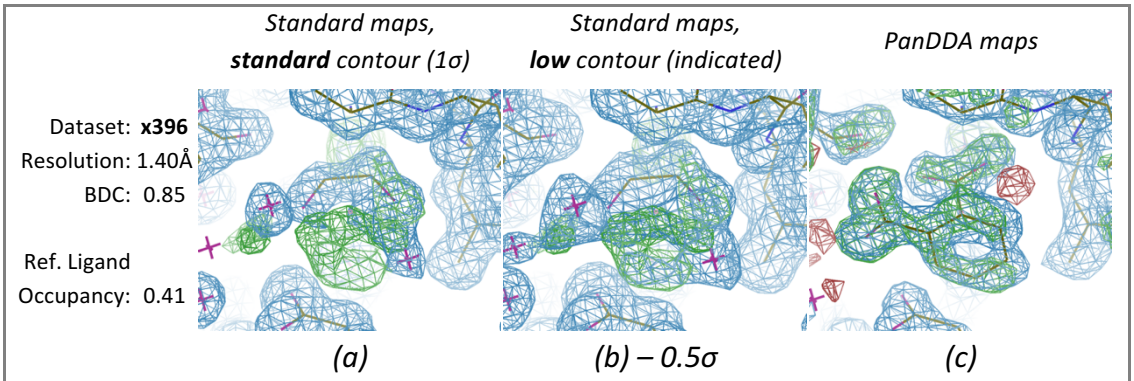


Figure 4.6. ***JMJD2D: "unexpected ligand" example.*** *Maps and contour levels are as in Figure 4.1, except where indicated.*

Dataset x396: *Binding and identity of the ligand are obscured by solvent in conventional maps, whereas PanDDA maps show clear density, which does not match the shape of the supposed ligand; the clarity of the event map enables the identification of the bound ligand's atomic connectivity.*

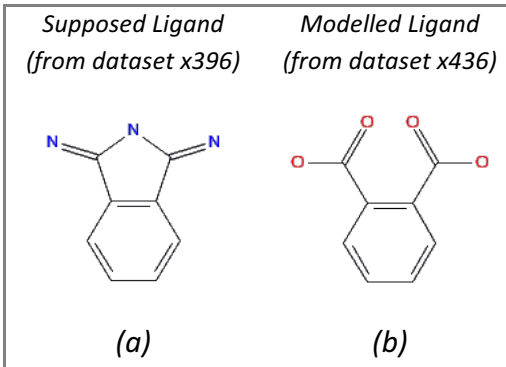


Figure 4.7. ***JMJD2D: Possible ligands bound in dataset x396.*** *(a) The soaked ligand in dataset x396 does not match the event density in Figure 4.6c. (b) This instead matches the connectivity of the ligand in x436 (as modelled and refined). Although this ligand fits the density and refines well, it is still possible that a hydrolysed form of the compound in (a) is bound instead.*

4.5.3 Identification of putative allosteric binders

The analysis of the JMJD2D fragment screen revealed many compounds bound on the surface of the protein and in the crystal contacts (in contact with symmetry-related copies of the protein). The distribution of identified non-crystal-contact fragments is shown in Figure 4.8, and some examples of the binding ligands are shown in Figure 4.9. However, many of these ligands make non-specific interactions with the protein, and moreover exist as singleton binding sites, making them of limited interest to FBLD. Two ligands are bound in the peptide binding groove (site A in Figure 4.8), one of which is shown in Figure 4.9a-c.

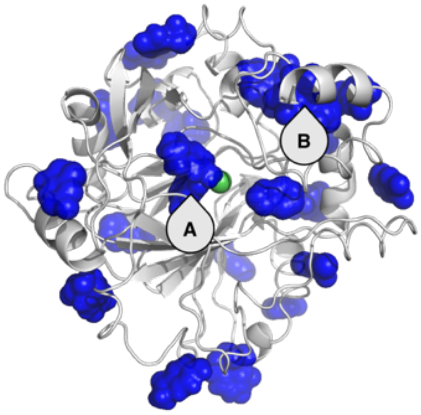


Figure 4.8. ***JMJD2D: fragments bind all over the surface of the protein.*** *The ground-state structure of the protein is shown in cartoon representation. Binding ligands not interacting with crystal contacts are shown as blue surface representation. The binding site metal (green sphere) is located at the centre of the image. Site A: Two ligands bind in and around the principal peptide-binding site, in addition to the example in Figure 4.5. Site B: Six ligands reorder the terminal helix of the protein (section 4.5.4).*

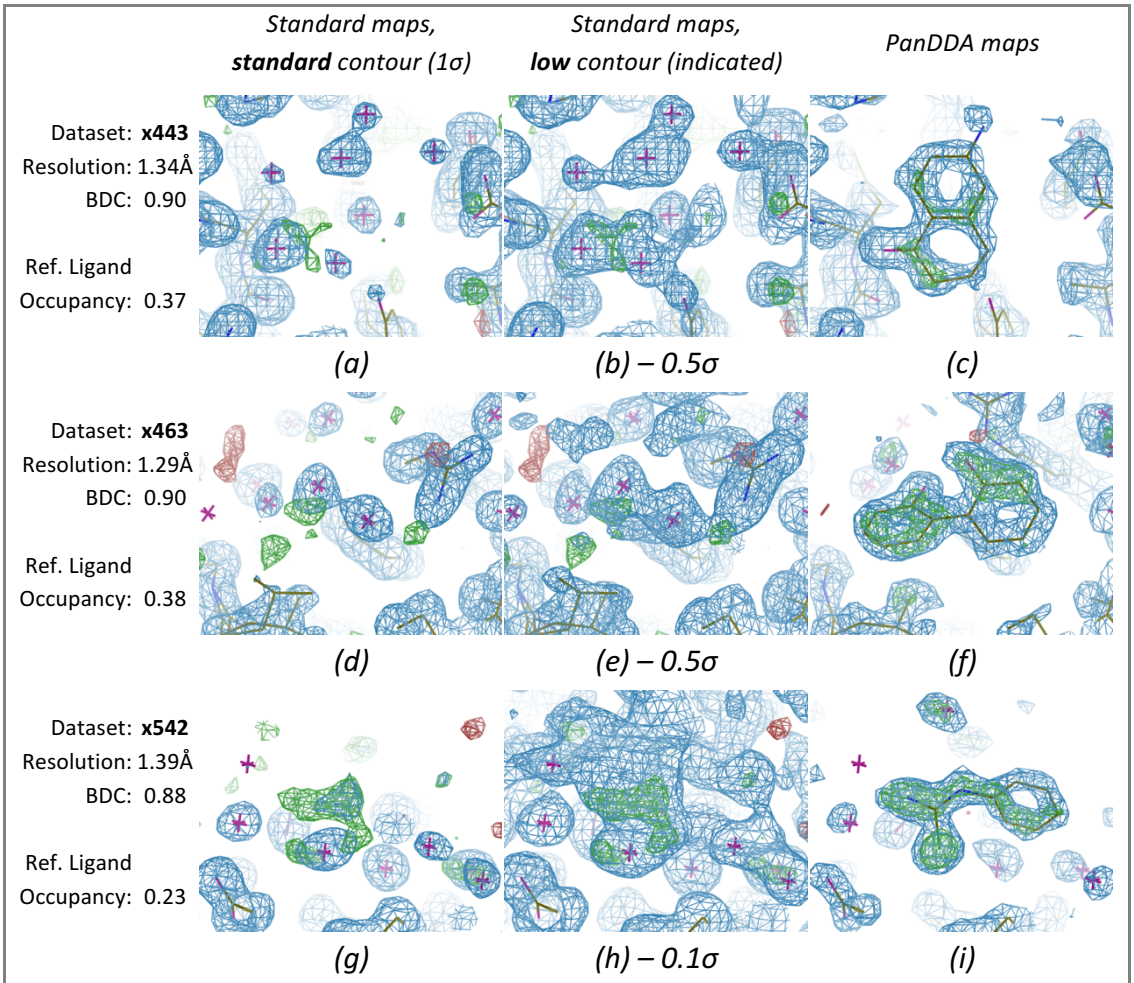


Figure 4.9. ***JMJD2D: detection of weak binders.*** *Maps and contour levels are as in Figure 4.1, except where indicated.*

Datasets x443, x463 and x542: *Conventional maps show evidence for nothing but solvent. However, the PanDDA maps show clear density for a weakly bound ligand.*

4.5.4 Binding-induced conformational changes

Multiple compounds are observed to reorder the terminal helix of JMJD2D (site B in Figure 4.8); two of the ligands are shown in Figure 4.10. The superposition of the two very different states makes the density completely uninterpretable. However, the PanDDA maps separate the modelling of the two states and enable the identification of the binding pose of the ligand and the bound conformation of the protein. The resulting atomic models for dataset x402 are shown in Figure 4.11; the logical and systematic labelling of the alternate conformers of the protein permit the straightforward extraction of the bound-state structure from the ensemble.

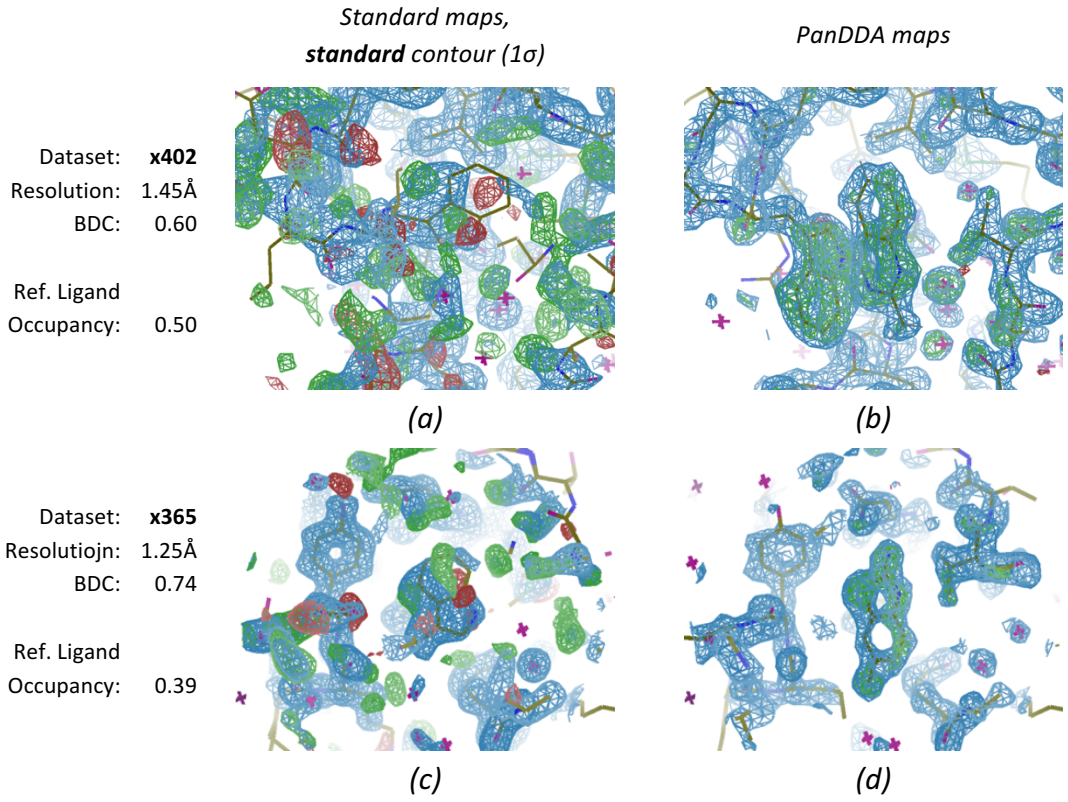


Figure 4.10. ***JMJD2D: detection of helix reordering.*** *Maps and contour levels are as in Figure 4.1, except where indicated.*

Datasets x402 & x365: *Conventional maps show the uninterpretable superposition of two very different helix states. PanDDA maps show clear density for a partial-occupancy ligand.*

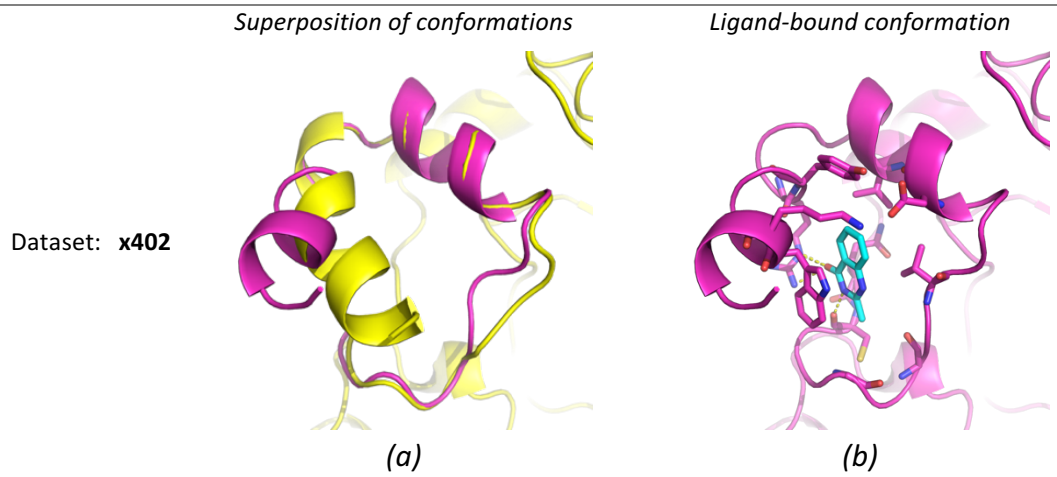


Figure 4.11. ***JMJD2D: binding-induced helix movement.*** *(a) The binding of the ligand causes a cryptic pocket to open under the terminal helix of the protein. The crystallographic dataset results in the superposition of the unbound (yellow) and bound (magenta) states. (b) The PanDDA maps enable the modelling of the bound conformation, and the systematic assignment of conformer IDs enables the simple extraction of the bound state (as shown). The easy extraction of bound states aids the use of the structure outside of the crystallographic experiment (e.g. for medicinal chemists).*

4.6 Bromodomain-containing protein 1

A fragment screen of the 3D fragment library against the bromodomain of Bromodomain-containing protein 1 (BRD1) resulted in 292 crystallographic datasets. Conventional difference-density based analysis revealed a high hit rate; two copies of the protein are contained within the ASU, and 29 hits were identified to bind in at least one of the binding sites, resulting in a hit-rate of 9.9% (Table 4.2).

4.6.1 Detection of further hits

On application of the PanDDA method, a further 13 binding fragments were identified, including those binding at an additional binding site on the side of the protein, which likely has no biological importance. Two examples of new PanDDA-identified ligands in the main binding site are shown in Figure 4.12.

One particularly exemplary example is dataset x050, where almost no density is seen that is not explained by the bound solvent from the reference datasets. This reason for

the complete lack of difference density is revealed in the event map, which shows that a large part of the ligand overlaps well with the bound ethylene glycol molecules. The ligand is only detected in the Z-map due to the part of the ligand that does not overlap with the solvent: a five-membered ring, which contains a sulphur (top-left of Figure 4.12c). The event map further shows clearly that this ring is present in two distinct conformations (Figure 4.13).

As discussed in section 4.1.2, the high hit rate may have led to the suppression of Z-scores at the principal binding sites, resulting in several binding ligands being missed in the automated analysis; these ligands were still identifiable in the Z-maps.

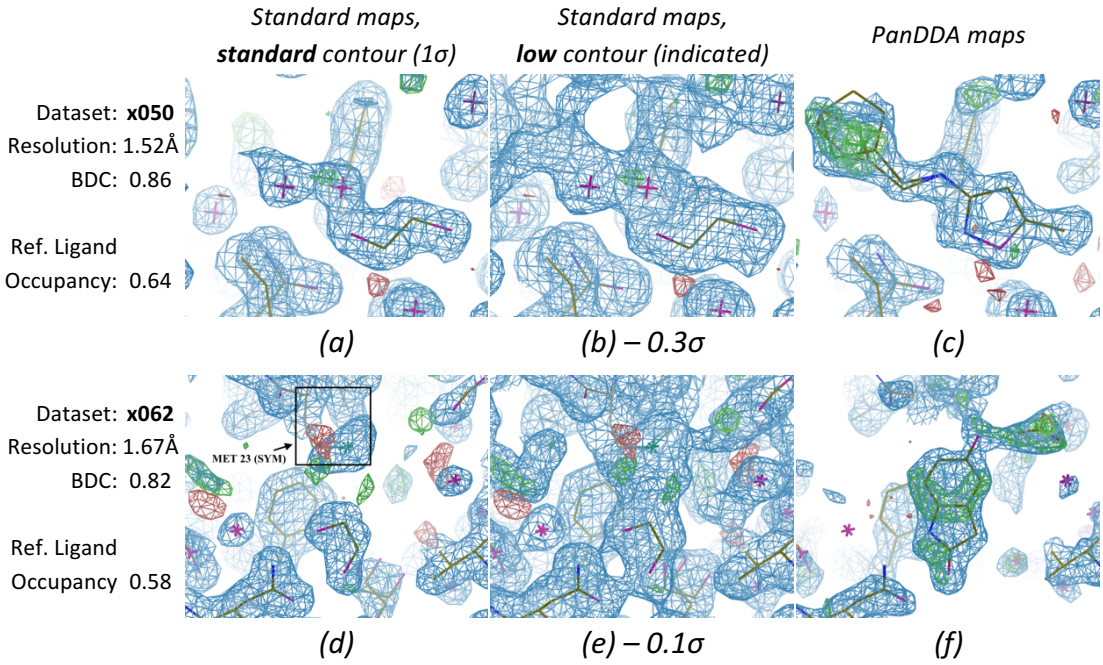


Figure 4.12. ***BRD1 examples: ligand identification.*** *Maps and contour levels are as in Figure 4.1, except where indicated.*

Dataset x050: *Conventional maps show no density in the binding site that is not accounted for by the pre-existing ground-state solvent model. PanDDA maps show clear density for the ligand, and show that the solvent-exposed ring, which contains a sulphur, is partially disordered (Figure 4.13).*

Dataset x062: *Conventional maps show poor and unconvincing evidence for a bound ligand. PanDDA maps show clear density for the ligand, and associated side-chain conformations. The ligand binding requires the movement of symmetry methionine (MET 23), indicated in (d), likely explaining the low occupancy and the poor quality of the event map.*

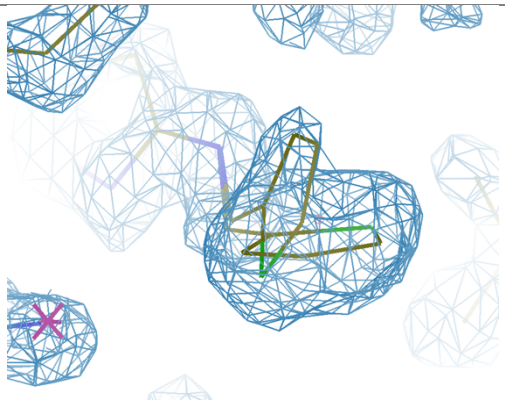


Figure 4.13. ***BRD1-x050: The five-membered ring of a ligand is clearly present in two conformations****. Even though the binding ligand creates no difference density in the standard crystallographic maps (as shown in Figure 4.12), the presence of two distinct conformations of the five-membered ring can be seen in the event map (blue, 2σ). The strength of the density allows the locations of the sulphur to be identified unambiguously. The refined occupancies of the two conformations are 0.44 (horizontal conformation) and 0.20 (angled conformation).*

4.6.2 Unambiguous identification of mislabelled/misdispensed ligands

As well as the identification of new ligands, the PanDDA method also revealed the presence of two more "unexpected compounds", as in previous sections. The compound for dataset x098 is found in the binding site of dataset x099, as well as the binding site of dataset x098, suggesting that the compound has diffused from one crystal drop to the other during soaking. Dataset x099 has two different species of compound bound, as the recorded ligand for dataset x099 is also bound allosterically to the side of the protein; PanDDA maps make these events clear and easily interpretable.

4.6.3 Binding of follow-up compounds

The PanDDA characterisation of a crystal form can be re-used to analyse any subsequent datasets of the same crystal form. This was applied in the case of several chemically elaborated fragments (follow-ups) following the initial screen. Of the 25 follow-up compounds screened, three binders were identified and modelled manually. After the application of the PanDDA method, eight binding ligands were identified. Disorder of sections of the ligand was common, but was frequently interpretable as for

the example shown in Figure 4.13. This shows the application of the PanDDA method beyond the initial fragment screen, and further that the PanDDA method has potential use beyond the identification of "weak" fragment binding.

4.7 Quality of the refined models

Although the modelling of the ligand is often clear in the PanDDA event maps, we still require that refinement of the ligands results in a good model as a further step of validation. The validation methods from the previous chapter were applied to all of the refined models from the fragment screens; all validation results are included below, ordered by the refined occupancy of the ligand (Table 4.4-Table 4.7). All the identified ligands score well by these validation metrics; for example, no modelled ligand has an RSCC below 0.7, even at low occupancy.

*Table 4.4. **SP100 ligand validation scores.** Rows are sorted by final refined occupancy. The Residue column annotates different ligands in the same dataset. RMSDs marked by asterisks are large because ligands were manually re-modelled after initial placement. RMSD is missing where ligands had atoms added or removed, i.e. through alternative conformations.*

Dataset	Residue	Res. (Å)	1-BDC	Occ.	RSCC	RSZD	B Ratio	RMSD (Å)
x601	F 1	1.57	0.13	0.38	0.82	0.3	1.26	4.99*
x596	F 2	1.55	0.11	0.40	0.86	1.4	1.66	0.43
x596	F 1	1.55	0.14	0.49	0.87	0.8	1.42	0.34

*Table 4.5. **BAZ2B ligand validation scores.** Details are as described in the legend of Table 4.4. Dataset x538 reports the occupancy of all ligand atoms except the bromine, which is likely lower due to radiation damage.*

Dataset	Residue	Res. (Å)	1-BDC	Occ.	RSCC	RSZD	B Ratio	RMSD (Å)
x538	C 1	1.77	0.15	0.40	0.96	1.7	1.05	0.27
x434	C 1	1.65	0.26	0.55	0.94	1.1	1.32	0.24
x559	C 1	1.78	0.23	0.55	0.91	1.4	1.14	0.30
x529	C 1	1.78	0.20	0.61	0.97	1.2	1.07	0.22
x492	C 1	1.78	0.24	0.64	0.97	1.8	1.16	0.17
x509	D 1	1.92	0.38	0.76	0.97	1.7	1.18	0.44
x481	C 1	1.65	0.34	0.77	0.95	0.9	1.26	0.16
x575	C 1	1.83	0.28	0.94	0.99	2.2	1.06	0.16
x583	C 1	1.80	0.30	0.94	0.98	1.1	1.21	0.33

Table 4.6. ***JMJD2D ligand validation scores.*** *Details are described in the legend of Table 4.4.*

Dataset	Residue	Res. (Å)	1-BDC	Occ.	RSCC	RSZD	B Ratio	RMSD (Å)
x336	F 1	1.34	0.11	0.22	0.88	1.8	1.30	0.15
x401	F 1	1.48	0.08	0.22	0.87	1.7	1.45	0.12
x542	F 3	1.39	0.12	0.23	0.87	0.2	2.03	0.35
x543	F 1	1.40	0.10	0.25	0.87	0.0	2.30	0.42
x486	F 3	1.36	0.15	0.27	0.86	0.1	1.42	0.54
x443	F 2	1.34	0.08	0.37	0.91	2.2	2.54	0.22
x463	F 1	1.29	0.10	0.38	0.81	0.3	1.68	0.45
x365	F 1	1.25	0.18	0.39	0.88	1.2	1.33	0.16
x376	F 1	1.45	0.22	0.40	0.90	1.8	1.25	0.31
x396	F 1	1.40	0.15	0.41	0.91	0.7	1.70	-
x443	F 1	1.34	0.10	0.41	0.77	1.7	1.88	0.20
x542	F 2	1.39	0.13	0.41	0.82	1.5	2.07	0.35
x396	F 2	1.40	0.15	0.43	0.91	0.5	1.43	-
x486	F 2	1.36	0.18	0.43	0.83	0.5	2.07	0.32
x639	F 1	1.29	0.17	0.44	0.95	0.2	1.28	1.04*
x393	F 1	1.74	0.25	0.45	0.80	2.0	0.99	0.12
x486	F 1	1.36	0.19	0.46	0.93	2.3	1.19	0.20
x639	F 2	1.29	0.15	0.46	0.87	0.0	1.60	0.39
x637	F 1	1.40	0.23	0.47	0.94	2.1	1.50	0.19
x402	F 1	1.45	0.20	0.50	0.75	0.5	1.22	0.08
x555	F 1	1.60	0.17	0.50	0.91	0.6	1.54	0.54
x386	F 1	1.27	0.17	0.52	0.92	2.1	0.96	0.99
x486	F 4	1.36	0.23	0.52	0.87	1.0	1.91	0.21
x568	F 1	1.97	0.35	0.52	0.83	1.0	1.15	0.27
x623	F 1	1.14	0.22	0.53	0.91	0.0	1.63	0.16
x637	F 4	1.40	0.22	0.53	0.94	1.0	1.34	-
x637	F 2	1.40	0.25	0.55	0.97	1.3	1.28	0.17
x637	F 3	1.40	0.20	0.55	0.98	0.8	0.95	1.07*
x494	F 1	1.35	0.18	0.56	0.94	0.1	1.08	0.14
x542	F 1	1.39	0.12	0.61	0.97	1.4	1.72	0.32
x611	F 1	1.15	0.17	0.61	0.87	0.7	2.15	0.30
x402	F 2	1.45	0.17	0.62	0.75	0.5	2.33	0.19
x620	F 1	1.25	0.14	0.62	0.78	0.2	2.49	0.23
x393	F 2	1.74	0.26	0.63	0.89	0.0	1.71	0.14
x395	F 1	1.43	0.36	0.66	0.91	1.1	0.72	0.13
x378	F 1	1.24	0.30	0.74	0.94	1.2	1.01	0.14
x392	F 1	1.35	0.30	0.98	1.00	0.0	0.86	0.10

Table 4.7. ***BRD1 ligand validation scores.*** *Details are described in the legend of Table 4.4. The bromine of the ligand in dataset x097 has a different occupancy to the rest of the ligand. Ligands with no 1-BDC value were indicated by the Z-map but missed in the automated analysis; these ligands were typically bound at both NCS copies, and identified automatically as binding to one, but not the other. This suggests that the high hit-rate may have negatively impacted the statistical maps, and that an iterative analysis approach should have been used.*

Dataset	Residue	Res. (Å)	1-BDC	Occ.	RSCC	RSZD	B Ratio	RMSD (Å)
x038	E 1	1.47	0.10	0.29	0.79	1.9	1.20	2.99*
x047	E 1	1.56	0.14	0.38	0.76	0.1	1.08	0.38
x099	E 1	1.52	0.22	0.45	0.76	1.2	1.42	0.56
x044	E 1	1.48	0.19	0.48	0.79	0.0	1.35	-
x271	E 2	1.58	-	0.49	0.88	0.3	1.41	0.35
x069	E 1	1.49	0.23	0.55	0.79	0.1	1.41	0.37
x092	E 1	1.62	0.21	0.55	0.78	0.1	1.53	0.50
x082	E 2	1.50	-	0.57	0.86	0.1	1.23	-
x062	E 1	1.67	0.18	0.58	0.95	0.7	1.15	0.20
x160	E 1	1.77	0.44	0.58	0.92	0.1	0.99	0.45
x081	E 1	1.61	0.12	0.59	0.86	1.3	1.41	0.48
x271	E 1	1.58	0.26	0.59	0.92	0.6	1.30	0.43
x225	E 1	1.56	0.24	0.60	0.88	2.6	1.16	0.42
x237	E 2	1.62	0.28	0.62	0.80	0.3	1.36	0.39
x099	E 3	1.52	-	0.63	0.89	0.1	1.47	-
x223	E 1	1.62	0.34	0.63	0.91	1.8	1.03	0.22
x270	E 2	1.97	0.61	0.63	0.90	1.7	1.07	-
x050	E 1	1.52	0.14	0.64	0.83	0.0	1.73	0.44
x099	E 2	1.52	0.19	0.66	0.90	0.2	1.41	-
x334	E 1	1.61	0.27	0.67	0.91	0.4	1.24	0.31
x295	E 2	1.77	0.27	0.68	0.94	0.0	1.19	0.26
x083	E 2	1.50	0.26	0.69	0.91	0.1	1.30	0.29
x160	E 2	1.77	0.44	0.69	0.91	0.6	1.17	0.39
x295	E 1	1.77	0.23	0.70	0.91	0.1	1.57	0.41
x066	E 1	1.70	0.20	0.71	0.94	0.9	1.45	0.33
x225	E 2	1.56	0.27	0.71	0.93	0.1	1.52	0.91
x080	E 1	1.44	0.24	0.72	0.91	0.6	1.30	1.16*
x270	E 1	1.97	0.59	0.73	0.94	2.0	0.91	0.45
x261	E 1	1.58	0.40	0.74	0.89	1.0	1.17	0.23
x186	E 1	2.37	-	0.77	0.94	0.7	1.05	0.40
x049	E 1	1.46	0.30	0.78	0.96	1.6	1.18	0.15
x298	E 1	1.75	0.26	0.78	0.94	1.0	1.31	0.35
x274	E 2	1.75	0.36	0.80	0.95	0.2	1.21	0.21
x167	E 2	1.61	-	0.82	0.96	0.2	1.02	0.32
x310	E 1	1.76	0.53	0.84	0.93	1.9	0.98	0.40
x136	E 1	2.27	0.40	0.86	0.94	1.5	1.05	0.47
x084	E 2	2.12	0.57	0.87	0.96	0.9	1.17	0.27
x310	E 2	1.76	0.62	0.88	0.91	0.2	1.05	0.38
x186	E 2	2.37	0.41	0.89	0.95	1.0	1.32	0.49
x167	E 1	1.61	0.37	0.90	0.90	0.3	1.19	0.45

x274	E 1	1.75	0.36	0.91	0.96	1.7	0.99	0.15
x325	E 1	1.50	0.29	0.92	0.97	0.1	1.12	-
x083	E 1	1.50	0.27	0.95	0.94	1.8	1.17	0.30
x084	E 1	2.12	0.53	0.95	0.96	2.1	0.99	0.40
x082	E 1	1.50	0.35	0.97	0.94	2.2	1.12	0.36
x093	E 1	1.78	0.32	0.97	0.97	1.3	0.98	0.14
x237	E 1	1.62	0.29	0.97	0.95	1.3	0.97	0.16
x284	E 1	1.43	0.28	0.98	0.95	0.1	1.37	0.25
x083	E 3	1.50	-	0.99	0.96	0.4	1.38	0.11
x249	E 1	1.52	0.42	0.99	0.95	0.5	1.21	1.23*
x093	E 2	1.78	0.35	1.00	0.97	1.1	1.16	0.11
x097	E 1	1.48	0.30	1.00	0.96	0.6	1.12	0.17
x098	E 1	1.56	0.28	1.00	0.93	2.1	1.77	0.16
x098	E 2	1.56	0.31	1.00	0.97	0.4	1.06	0.09
x136	E 2	2.27	0.47	1.00	0.92	1.2	1.15	1.10*
x223	E 2	1.62	0.42	1.00	0.96	2.0	1.30	0.16
x242	E 1	1.54	0.40	1.00	0.94	0.7	0.92	0.50
x242	E 2	1.54	0.41	1.00	0.96	0.8	1.22	0.22
x249	E 2	1.52	0.46	1.00	0.96	0.3	0.95	0.11
x258	E 1	1.50	0.37	1.00	0.93	0.3	1.27	0.30
x258	E 2	1.50	-	1.00	0.94	1.2	1.48	0.22
x261	E 2	1.58	-	1.00	0.95	0.3	1.33	0.24
x276	E 1	1.63	0.31	1.00	0.97	1.6	1.14	0.12
x292	E 1	1.57	0.45	1.00	0.97	0.3	1.00	-
x292	E 2	1.57	0.47	1.00	0.97	0.1	0.97	-

4.8 Discussion

In this chapter, I have shown that the PanDDA method greatly increases the amount of structural information derived from fragment screening experiments by enabling the confident detection of partially-occupied ligands that do not express significant amounts of difference density; these weak ligands can further only be modelled once the superposed ground-state solvent has been removed in the event map. Ligands with occupancies as low as 22% have been detected (e.g. Figure 4.5). Beyond the identification of weakly-bound ligands, examples presented in this chapter show that mislabelled/misdispensed/degraded compounds (section 4.4.1 & 4.5.2), binding-induced sidechain- and loop-reordering (section 4.4.2 & 4.5.4) and discrete

conformational disorder (section 4.6.1) can all be identified through the clarity of the event map. Strongly bound ligands are also trivially identified.

Furthermore, I have shown that bound ligands can be identified in the crystallographic data without the requirement for extensive refinement to provide "optimal phases" (as in Schiebel et al. 2016): dimple performs only a short refinement, with no re-building, and yet we are able to identify weakly-bound ligands. The power of the PanDDA method arises through the harnessing of information from multiple datasets; however, it is still likely that better refinement protocols than Dimple would lead to improved signal identification.

4.8.1 Precision of the PanDDA events output

Here, I briefly discuss the efficiency that the PanDDA method allows in the analysis of the identified hits. For the four datasets shown here, the PanDDA analysis provide an acceptable precision (the fraction of identified events that resulted in a modelled ligand) of between 17-56%; the number of modelled ligands compared to the number of PanDDA-identified events are listed in Table 4.8. The PanDDA method reduces the search of the whole unit cell of all datasets down to ~100 events. This compares favourably to the human analysis, especially for the SP100 datasets: the human analysis took more than two days to identify three incorrect ligands, whereas the inspection of the PanDDA results took approximately 10 minutes to identify the three binders. The PanDDA approach means that the crystallographer spends most time modelling, rather than trying to assess the presence of the ligand in the data.

Table 4.8. ***Precision of PanDDA event identification.*** *The fraction of the identified events that resulted in bound ligands is shown for the four fragment screening datasets.*

Protein	Total Datasets	Modelled Ligands	Identified Events	Percentage (%)
SP100	116	3	18	17
BAZ2B	200	9	37	24
JMJD2D	226	37	145	26
BRD1	292	65	116	56

Furthermore, the number of false positives as shown by Table 4.8 is an overestimate as the PanDDA method identifies “changed-state signal” (crystallographic states that are unique to a specific dataset), rather than binding ligands. For instance, in the BAZ2B datasets, multiple events were caused by the interaction of a metal with several sidechains on the protein surface (this metal is likely a contaminant introduced with the ligand to the crystal). It is correct for the PanDDA implementation to have identified this as interesting, even though it is not a bound ligand. Similarly, disordered parts of the protein that spontaneously order in random datasets constitutes “real” crystallographic signal, although this information is not useful/interesting in most cases – it is a crystallographic artefact/noise.

One piece of future work that is required is to determine the optimal ranking and classification of the output list of identified events, and the tuning of the default PanDDA parameters. This will require many fragment screens to have been processed by PanDDA to train a scoring system. Automated ligand-fitting approaches may also be utilised to fit the identified ligands into the event density. Most importantly, utilisation of more robust refinement protocols, such as those available in autoBUSTER (Bricogne et al. 2011; Smart et al. 2012) will greatly improve the quality of the output structures.

4.8.2 Interpretability and clarity of events

Although many events are easily interpretable – where it is possible to identify the cause of the "changed-state" (even if it is not a binding ligand) – some events remain uninterpretable. In some of these cases, it may be because the "ground state" portion of the putative changed-state crystal is not represented by the average ground state of the other datasets; subtracting the averaged ground state in these cases will generate an event map that does not represent the changed state.

It may therefore be necessary to investigate further processing of the diffraction data prior to the comparison of the maps, such as blurring the density so that the same average B-factor is presented for each dataset: subtraction of a "low B-factor" map from a "high B-factor" map will clearly introduce noise into the analysis. This is the topic of future work.

Lastly, there is the possibility for the development of an iterative event map procedure, where the refined ground- and changed-state occupancies are used to re-compute the event map. Using these refined values as an improved estimate of BDC should lead to better interpretability of the event map.

4.9 Chapter Summary

In this chapter I have shown that the PanDDA approach can identify obscured changed-state signal in crystallographic datasets. I have further shown that the resulting ligands – modelled correctly – result in high-quality models. In the next Chapter I compare the PanDDA approach to the current state-of-the-art difference-density-based ligand detection method in Schiebel, Krimmer, et al. (2016), and discuss the ramifications of these results in the context of current crystallographic modelling approaches.

Chapter 5

Further application of the PanDDA method:

Challenging current crystallographic conventions

"The conventional way isn't always the right way."

The analysis of multiple fragment screening datasets in the previous chapter raises several points which merit further in-depth discussion.

Firstly, I discuss the ramifications of the results from the last chapter regarding conventional crystallographic approaches concerning the identification of bound ligands: that a complete model with "high-quality phases" enables the sensitive detection of bound ligands (section 5.1). I re-analyse a fragment screening campaign processed with a state-of-the-art model-refinement pipeline and show that the conventional difference-density-based approach misses many bound ligands (section 5.1.2). I also re-analyse the BAZ2B dataset with purposefully-degraded phases and show it is still possible to identify bound ligands (section 5.1.3). Both tests confirm that harnessing information from multiple datasets vastly improves signal detection.

Secondly I question the conventional crystallographic modelling approach whereby ligands are modelled as the sole crystallographic state (section 5.2). I show, across a range of ligand occupancies, that the inclusion of a superposed solvent model in every

case leads to an improved ligand model; I propose that a superposed ground-state model should *always* be included by default in crystallographic ligand modelling.

Lastly, I discuss several observations about the results from the PanDDA analysis in the previous chapter, such as the relationship between the validation metric scores and the occupancy of the ligands (section 5.3). I also discuss the relationship between the background-density correction factor and the refined occupancy of the ligand.

5.1 Comparison with the state-of-the-art and the effects of phase errors

"A multi-state superposition won't disappear with better phases."

First and foremost, it is clear from the results of the previous chapter that obstacles to ligand identification and modelling are not solely caused by errors in the phases close to the end of refinement; very clear density for the bound ligand can be obtained in the PanDDA event maps that is obscured in a multi-state superposition in the conventional maps. This is in contrast to the widely-accepted conventional ligand-modelling paradigm, which states that models must be complete and refined extensively for ligands to be identified (e.g. Schiebel, Krimmer, *et al*. 2016). Instead I have shown that phase quality is not an obstacle to signal detection when information from orthogonal datasets is used to enhance signal.

In the 364 crystallographic datasets published by Schiebel *et al.* (2016), the authors describe how ligands only became detectable after extensive refinement and optimisation of the model. I claim that the signal is still present in the maps even without this extensive refinement, and that the extensive refinement simply brings the

noise level in the difference map down to a level where the differences can be seen. All the flaws of difference-map-based detection methods remain: the density obtained in the difference map is not guaranteed resemble any bound ligand, and the presence of difference density that "matches the shape of the ligand" does not mean that a ligand is bound at that location (it may be a fortuitous superposition of bound solvent molecules).

5.1.1 Method for re-analysis of the data with sub-optimal phases

To test the assertion that phases are not the limiting factor to ligand identification, I analysed the Schiebel *et al.* datasets with PanDDA, but using the Dimple pipeline (CCP4) to generate the refined structures. Dimple is a "quick-and-dirty" refinement pipeline by comparison to the pipeline used by Schiebel *et al.* (currently unpublished): Dimple simply consists of a rigid body refinement followed by cycles of restrained refinement with REFMAC. No re-modelling or anisotropic B-factor refinement are performed.

Several datasets showed significant non-isomorphism and variability in spacegroup; datasets were grouped by spacegroup and then further clustered using average-linkage clustering on LCV (Foadi et al. 2013) between datasets, with a cutoff of 20%. This is a generous non-isomorphism cutoff, but removes datasets which may require representation by a different atomic model; variable input structures are not supported in the PanDDA implementation. Clustering resulted in the rejection of 16 datasets; one fragment was identified to bind in these 16 datasets in the original analysis.

The phase quality of the remaining 348 Dimple-processed datasets is degraded relative to the originally-refined datasets (Figure 5.1). The originally-refined R-free and R-work values are consistently lower than the Dimple-refined values (R-values calculated using

phenix.model_vs_data): the average difference is 1.85% for R-free and 2.85% for R-work. There is a slight increase in overfitting for the originally-refined datasets; the average R-factor gap is 1.00% larger in the originally-refined datasets (Figure 5.1c). The Dimple-refined datasets were analysed with PanDDA using the default parameters.

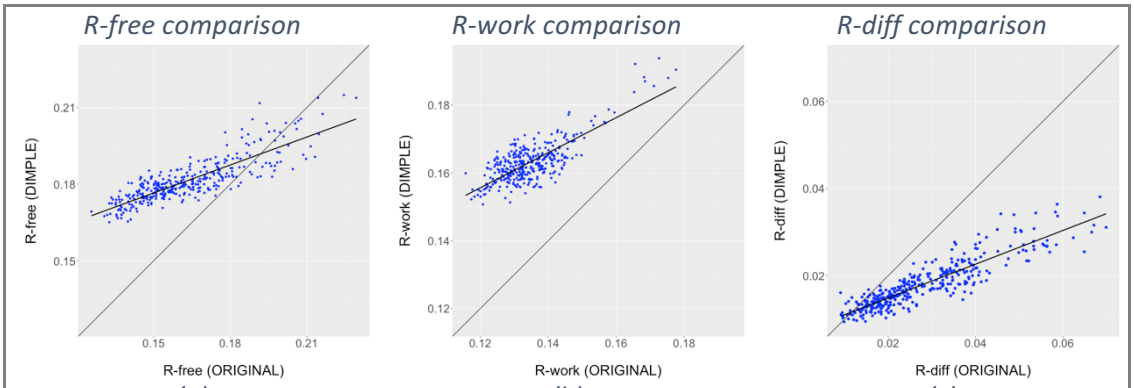


Figure 5.1. ***R-free, R-work and the R-diff for original and dimple refinements.*** *Comparison of the (a) R-free, (b) R-work and (c) R-diff (R-free – R-work) for all datasets. In most cases, the original published refinement results in a smaller R-free, but also to a larger R-diff.*

5.1.2 Results from re-analysis of the Schiebel dataset

The following is a preliminary analysis of the data; ligands were fitted but not refined. Furthermore, not all the Schiebel *et al.* structures had been released in the PDB at the time of writing (only the structures from automated refinement were published); comparisons with the publication were regenerated from the manuscript figures.

Overall, more binding ligands were identified using the PanDDA analysis of the dimple-refined data than were identified in the original analysis by Schiebel *et al.* (Table 5.1 and Table 5.2; full details can be found in Appendix D).

The ligands identified by PanDDA are divided into two classes: high-confidence models and medium-confidence models. High-confidence models provide an excellent match to the event map and modelling was straightforward. Medium-confidence models were those where the density called into questions at least a part of the ligand, for instance

where the ligand reacted in the crystal or had chemically degraded; this introduces some uncertainty into exactly which chemical species is bound, reducing the confidence in modelled atoms. In addition to this, some medium-confidence ligands were modelled into event density that indicated a different chemical species was introduced to the crystal than was thought (Figure 5.2); this is distinct to where the introduced compound reacted or was degraded. Other cases merely displayed unconvincing density.

Broadly, I confirm the binding of the ligands from the original analysis by Schiebel et al, though the PanDDA event maps result in different models for some binders. Several of the ligands in the original Schiebel analysis are only partially modelled; density for some of the atoms was not seen in the difference maps. The PanDDA maps generally show density for these missing sections, allowing them to be modelled, and further show that they are not disordered. Density for the azepane ring in fragment 17 was not seen in the original map, but the density for the ring is clear in the event map (Figure 5.3); a similar situation is seen for fragment 189 (Figure 5.4).

The PanDDA analysis also reveals upwards of 28 new bound ligands (depending on whether medium-confidence ligands are included); two high-confidence examples – one of which contains both a bromine and a sulphur – are shown in Figure 5.5 & Figure 5.6. One ligand identified by Schiebel *et al.* was not identified by the PanDDA method; this ligand was marked with Z-density (Figure 5.7) but fell below the threshold to be detected by the current un-optimised blob-finding algorithm used to process the Z-map in the PanDDA implementation.

Table 5.1. ***Comparison of the PanDDA analysis and the published analysis of the Schiebel datasets: high-confidence PanDDA ligands only.*** *Summary statistics for binders identified by PanDDA with "high confidence"; high confidence ligands were clearly identified and easily modelled into the event map. All numbers are shown only for the 348 analysed datasets, to permit direct comparison.*

Analysis	PanDDA (High Confidence)	Schiebel
Total datasets with bound ligands	94	70
Total number of bound ligands	106	85
Number of common datasets	66	
Number of unique datasets	28	4

Table 5.2. ***Comparison of the PanDDA analysis and the published analysis of the Schiebel datasets: including medium confidence PanDDA ligands.*** *Listed numbers are as described in Table 5.1. Summary statistics for binders identified by PanDDA including "medium confidence" ligands; medium confidence ligands either had unclear density for parts of the ligand or required that a reacted/degraded form of the soaked ligand was bound. The three ligands identified by in the original Schiebel analysis and the PanDDA high+medium confidence set, but not in the PanDDA high-confidence set, are fragments 34, 66, 218; fragment 48 was not identified in the PanDDA analysis, although it's binding is supported by manual inspection of the Z-map (Figure 5.7).*

Analysis	PanDDA (High+Medium Confidence)	Schiebel
Total datasets with bound ligands	107	70
Total number of bound ligands	122	85
Number of common datasets	69	
Number of unique datasets	38	1

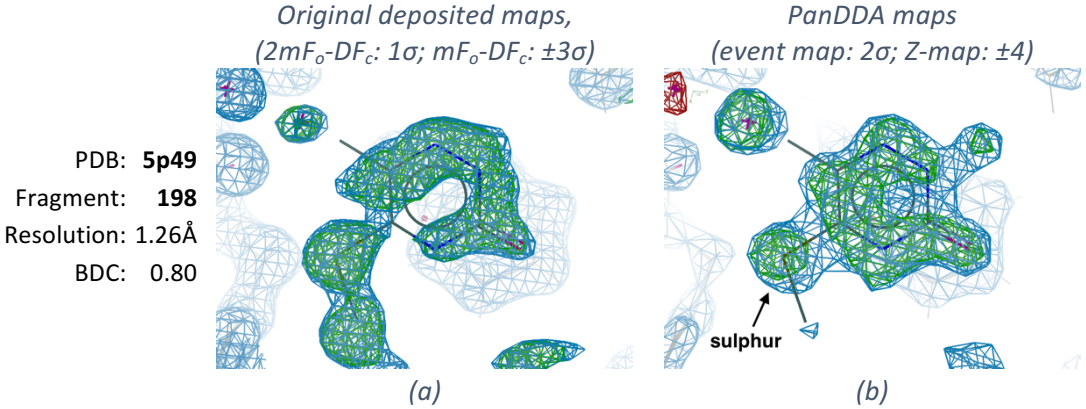


Figure 5.2. ***PanDDA maps indicate that the bound ligand is not the soaked ligand.*** *(a) The standard maps do not conclusively show that a ligand is bound; an amorphous blob of density is seen. (b) The PanDDA maps show that the atomic connectivity of the bound ligand is different to that of the assigned ligand, since the strength of the density unambiguously shows the position of the sulphur (indicated).*

PDB: **5oz6**
Fragment: **17**
Resolution: 1.55Å
BDC: 0.75

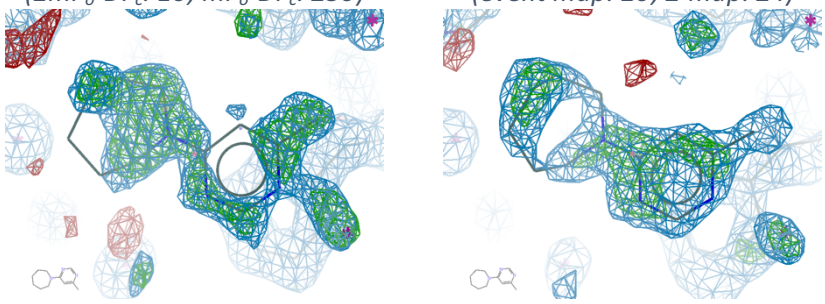


Figure 5.3. ***The PanDDA event map reveals density for the azepane ring of fragment 17.*** *(a) The density in the original deposited maps does not show overly convincing density for the bound ligand. Density for the azepane ring is missing; this ring was not modelled in the original publication. (b) The PanDDA maps reveal clear density for the binding ligand and the previously-unseen ring. The density for the water molecule in the bottom-right is also clarified.*

PDB: **5ozk**
Fragment: **31**
Resolution: 1.33Å
BDC: 0.86

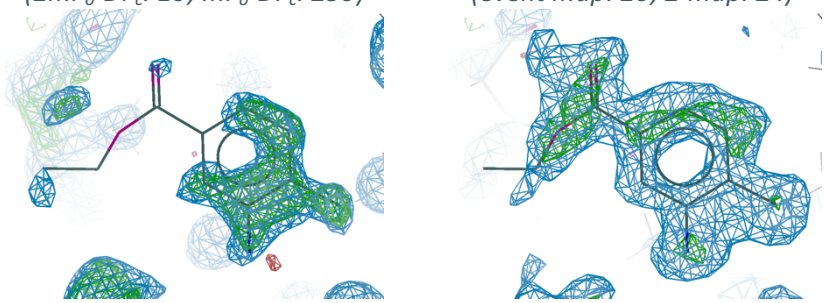


Figure 5.4. ***The PanDDA event maps reveal more density for bound fragments.*** *(a) Standard maps shows only the density for part of the ligand, and only part of the ligand was originally modelled. (b) The PanDDA event maps reveal more of the ligand, at a higher contour level, except for the terminal group (top-left), which is potentially disordered.*

PDB: **5p47**
Fragment: **196**
Resolution: 1.18Å
BDC: 0.87

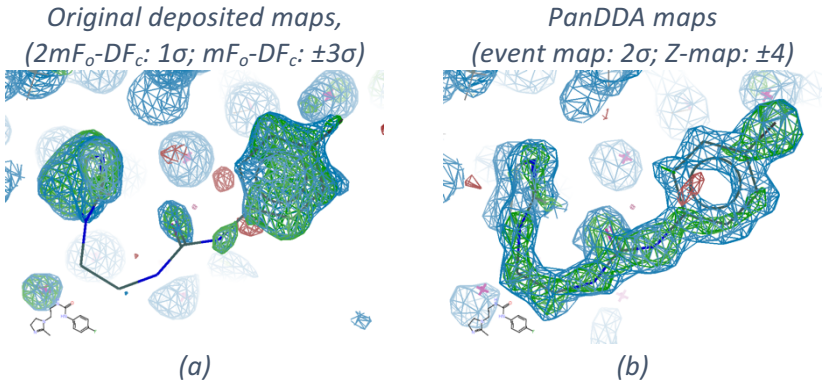


Figure 5.5. ***PanDDA event maps identify previously-unidentified bound fragments.*** *(a) The density in the standard maps shows only the density for part of the ligand, and this ligand was undetected/unmodelled. (b) The PanDDA maps identify the ligand strongly and present clear density for it.*

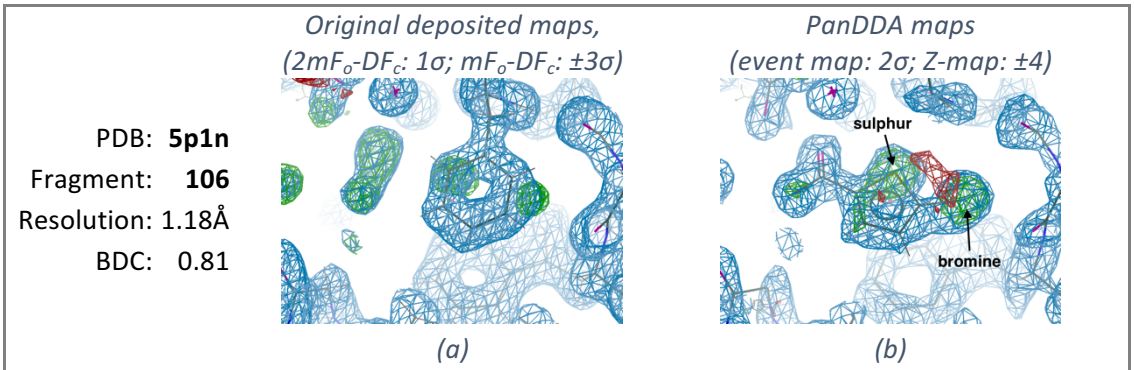


Figure 5.6. ***PanDDA event maps identify fragments containing heavy atoms.*** *A binding ligand displaces a ground-state phenylalanine conformation and displays minimal difference density. (a) Minimal evidence can be seen for the ligand even though it contains both a bromine and a sulphur atom. (b) The ligand pose is clearly revealed by the PanDDA maps; the sulphur and bromine are clearly defined.*

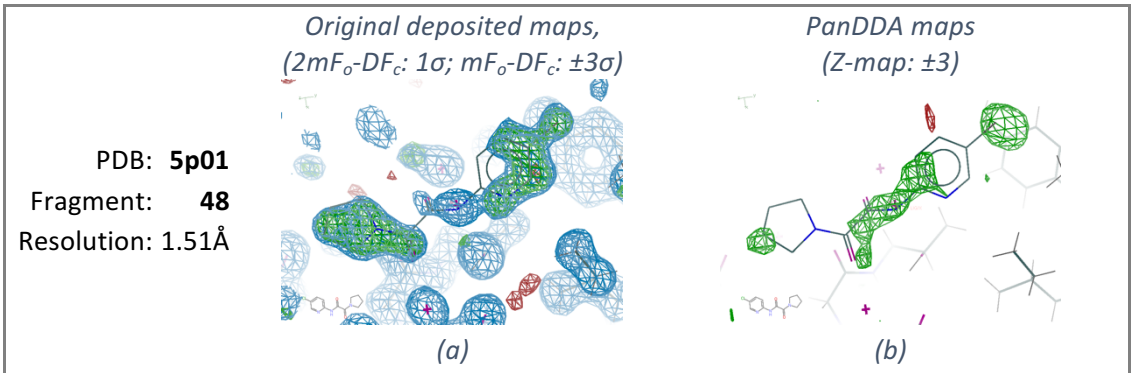


Figure 5.7. ***A bound ligand missed by the PanDDA implementation.*** *(a) Evidence for the bound ligand can be seen in the difference map, though the shape of the density clearly indicates that a superposed solvent state is present. (b) Although the ligand was not detected by the PanDDA method, Z-density confirms the presence of a changed-state.*

5.1.3 Ligand detection with significantly degraded model phases

To further demonstrate that phase quality is not the main obstacle to ligand detection, but merely a requirement of the difference-density approach, I purposefully degraded the protein model for the BAZ2B dataset (in regions distant to the binding site) and reanalysed the data with dimple and PanDDA. Degradation of the model induced an average 30° phase difference into the structure supplied to dimple (calculated with cphasematch between degraded and original structure); this degradation of the input model increases the R-free by approximately 10% in the refined datasets (Figure 5.8).

Even with these degraded phases, six of the nine bound ligands could still be identified clearly in the event map density (eight ligands were correctly identified by Z-maps, but the density for two ligands was not convincing enough to merit a model). Though the conventional maps are severely degraded such that no indication of a bound ligand can be seen, the PanDDA maps can still show clear evidence of binding (Figure 5.9).

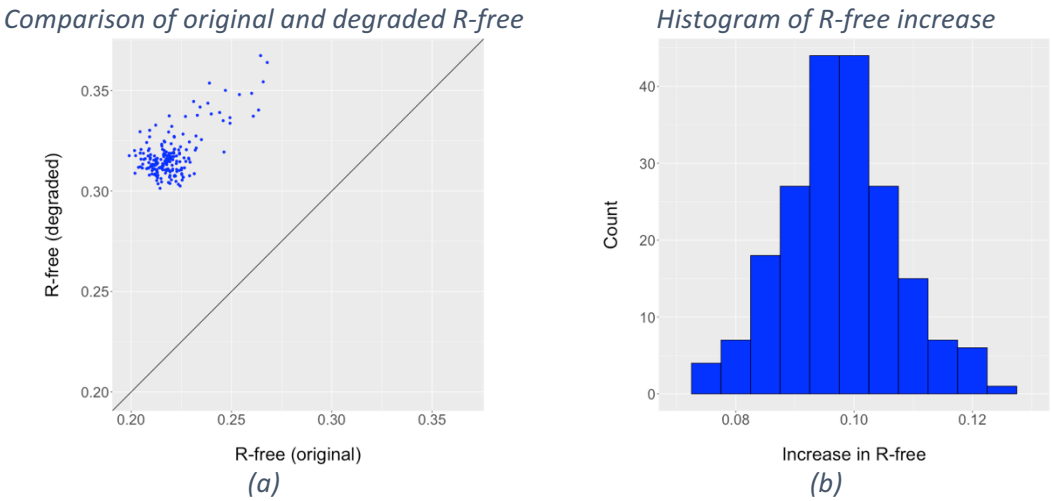


Figure 5.8. ***R-free comparison for the original and degraded dimple outputs.*** *(a) R-free comparison of the dimple output models for the original and the degraded protein model. (b) Histogram of the R-free differences for original and degraded refinements. The R-free of each dataset is increased by an average of 10%.*

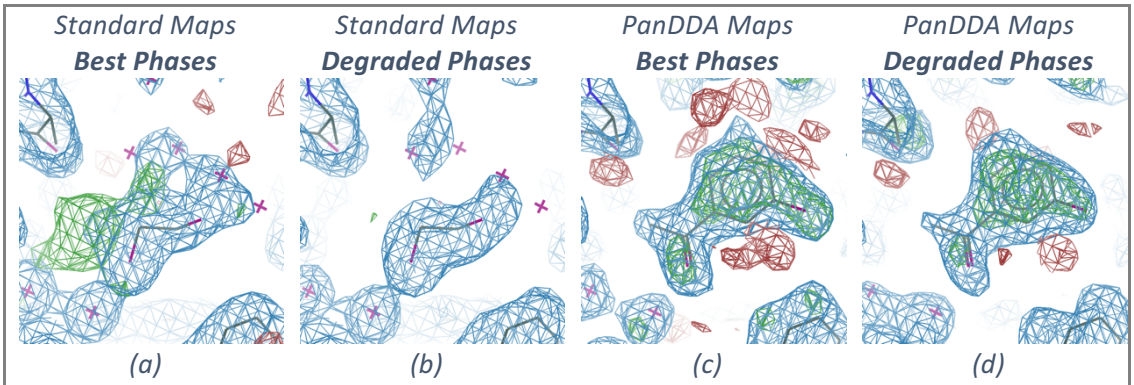


Figure 5.9. ***Weak ligand identification remains straightforward when phases are degraded.*** *A weak hit (occupancy: 0.64) in dataset x492: (a,b) 1.78Å $2mF_o$-DF_c (blue, 1σ) and mF_o-DF_c (green/red, ±3σ). (c,d) 1.79Å event (blue, 2σ) and Z-maps (green/red, ±3). R_{work}/R_{free} are 0.18/0.21 and 0.30/0.32 for best and degraded phases respectively. Degradation of phases removes all evidence of the ligand from conventional maps at standard contour levels. Event map BDC for best and degraded phases are 0.77 and 0.73 respectively, and although the quality of the event density for the ligand is reduced, ligand identification is no more difficult.*

5.1.4 Discussion

Clearly, phase quality need not be an obstacle to ligand-identification; more successful ligand identification can be performed using the PanDDA method with sub-optimal phases than can be seen in the difference map with optimal phases (section 5.1.2), and even with severely degraded phases, ligand identification is still possible (section 5.1.3).

These results therefore upend a long-held tenet in macromolecular crystallographic model building: that to visualise subtle features requires optimal phase estimates and thus a model as complete and globally error-free as possible. Phase-quality has long been the holy grail of macromolecular crystallography; these results suggest that in a large proportion of cases, use of multiple datasets could lead to more robust and sensitive interpretation of experimental data without perfect phases.

Conscientiously observed, the conventional modelling approach further places a heavy time burden on the analysing scientist as it demands multiple iterations of modelling for each dataset: a *fit-refine-assess* approach. The Schiebel et al. analysis required the input of eight crystallographers to analyse the 364 datasets; in comparison, using the PanDDA approach, only one crystallographer is needed to process, assess and model the data, with one further to provide an independent validation.

5.2 Representation of multiple crystal states in modelling

"Of course the solvent's there; no ligand is truly unitary occupancy."

Another reality that has become clear through the modelling of the partial-occupancy ligands identified in the previous chapter is that it is imperative to represent the unbound "ground state" of the crystal in refinement. Without the superposed solvent

model, low-occupancy ligands would not refine correctly, and would poorly explain the observed crystallographic density for the whole crystal. For this reason, multi-state ensembles are automatically part of the PanDDA implementation.

Outside of the experiments covered in this thesis, ligands will often – and likely almost invariably – bind at sub-unitary occupancy. However, it is standard practice not to model a superposition of multiple states, but instead to model only the ligand-bound conformation (and furthermore normally at unitary occupancy); this is commonly observed in the PDB (Figure 5.10) (Berman et al. 2000; Berman et al. 2003).

Occupancy refinement of ligands is likely avoided due to well-known interdependencies, instabilities and ambiguities that can occur in the simultaneous refinement of both B-factors and occupancies: improvements in crystallographic model fit can equally well be achieved by reducing occupancy or increasing B-factors (Bhat 1989). When ligands are modelled at full occupancy, any resulting error is absorbed by inflating the refined B-factors. One is led to conclude that occupancy refinement is only deemed necessary when difference density appears over the ligand model, an impression corroborated by multiple conversations in online discussion fora such as *ccp4bb* and *ResearchGate*.

If occupancy refinement of the ligand-bound state is performed without a superposed solvent model, this implicitly implies that the rest of the crystal is either represented by vacuum – which is highly unlikely – or by bulk solvent, depending on the refinement program used. Close to the surface of the protein, it is unlikely that the solvent is truly represented by a bulk solvent model; this is especially true of binding sites, where solvent and buffer molecules bind in an ordered fashion at high occupancy, as in the

examples presented in section 5.2.2. The absence of a superposed model is a glaring modelling omission, and here I will show that inclusion of the superposed unbound state leads to a more complete model of crystal, and to a better ligand model.

I posit that ligands will – in the general case – always be better modelled with explicit representation of the superposed ground state, determined from a ground-state crystal of the protein. Inclusion of the ground-state allows the occupancy of the superposed states to be constrained in refinement, and enforces prior knowledge of the crystal, leading to a Bayesian approach of multi-state ligand modelling.

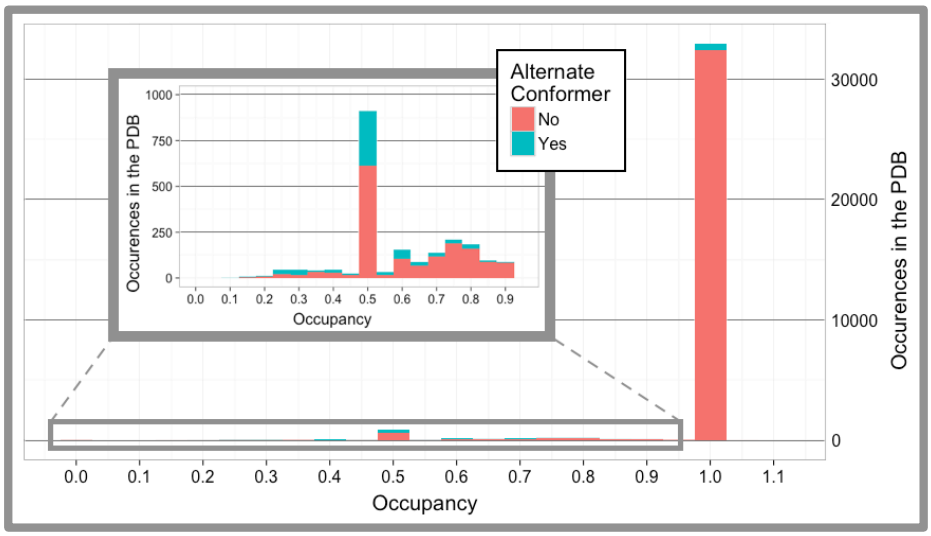


Figure 5.10. ***Most ligands in the PDB are modelled at unitary occupancy, and many partial occupancy ligands are not modelled with an alternate state.*** *Histogram of all ligand occupancies in the PDB (red: no conformer ID, blue: modelled with a conformer ID). Only the first instance of each ligand type from each PDB structure was used; following this all ligands with fewer than 5 non-hydrogen atoms and more than 50 instances were removed to avoid bias towards common molecules. Where alternate conformations of ligands are present, the total occupancy is used. The large majority of ligands are modelled at unitary occupancy (32396, 92.1%). A smaller number have non-unitary occupancies but no alternate conformer identifier (1640, 4.7%). The remainder are modelled using alternate conformers (1122, 3.2%), of which 548 are ligands with alternate conformers that sum to unitary occupancy. Therefore, a maximum of 574 (1.6%) are modelled with a superposed solvent state (as this requires an alternate conformer and sub-unitary occupancy). Worryingly, there are also ten instances of ligands with more than 100% occupancy. These modelling statistics are unlikely to represent the true situation in crystals, where ligands will rarely bind at near-full occupancy; ligands will always have a superposed solvent model where present at partial occupancy.*

5.2.1 Method for assessing the effect of inclusion of the ground-state

To demonstrate the improvement of ligand models through inclusion of the superposed ground state, I present four examples from the analysis in Chapter 4 that cover a range of ligand occupancies. I begin with the ensemble model from the PanDDA analysis, where the modelled bound state and the input ground state are assigned alternate conformers and merged (see Chapter 3). The ground state is determined in an orthogonal (unbound) reference dataset, and the bound state is determined by the PanDDA event maps (Figure 5.11); here I show the effect of the inclusion/omission of the ground state.

Three models of the crystal containing ligands are refined and compared: a ligand-state-only model; a high-quality ensemble model; and a degraded-phase ensemble model. A solvent-state-only model is also refined for completeness. The degraded phase model is once again included so that the effect of inclusion/omission of the ground-state can be compared with reference to the quality of the overall phases.

The ligand-state-only model for refinement is obtained by removing the ground state from the ensemble and setting the ligand occupancy to 0.95. The solvent-state-only model is similarly generated by removing the ligand-bound state and setting the solvent occupancy to 1.0 (this simulates the normal modelling case, where a single-state solvent model's occupancy would not typically be refined). Degraded-phase models are created by distorting the ensemble model in regions distant from the ligand binding site, thereby introducing global phase error. Induced mean model phase difference relative to the full ensemble model is in the range of 24-35° (as calculated by cphasematch; CCP4) for the examples in section 5.2.2.

All models are refined with phenix.refine (version 1.9-1682, Afonine et al. 2012), using the default parameters, against crystallographic data from before a ligand was placed, to prevent phase bias. Ligand occupancy is refined for all models; for the ensemble models, the occupancies of superposed states are constrained to sum to unity.

The refined ligand models are compared using the set of validation metrics defined in Chapter 3: RSCC, RSZD, RSZO, B-factor ratio & RMSD. The observed RSZO is directly related to the occupancy of the ligand, since RSZO is calculated by taking the average of the density over the model and dividing by the noise in the map. To allow the comparison of models at different occupancies in the same dataset, we divide RSZO by the occupancy of the residue to give a normalised value (RSZO/OCC).

Furthermore, RSZD and RSZO are less informative when analysing models with poor phases, because they are dependent on the quality of the model phases. Both RSZD and RSZO are derived with the assumption of near-convergence phases, and use an estimation of the noise in the maps to calculate quality criteria for residues. Lower RSZD would normally indicate a better model, but this is not the case here: when the quality of the phases is reduced, the noise in the maps increases, and therefore decreases both RSZD and RSZO, regardless of whether the model has changed.

To allow visual comparison of the models, validation metrics are displayed visually as radar plots, where the “better” the metric value, the closer it is to the centre of the plot. The axes of the radar plot are scaled such that the “best” value is plotted at the centre of the plot and the “worst” value is plotted at the extreme of the axis.

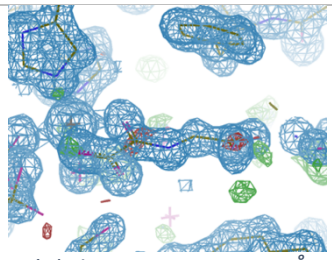
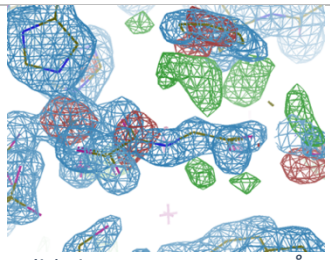
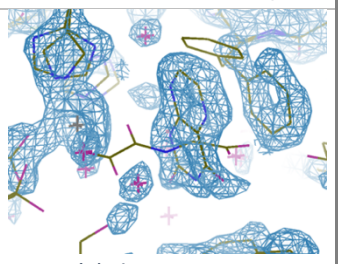
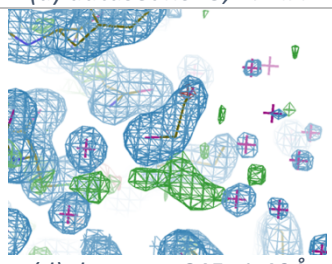
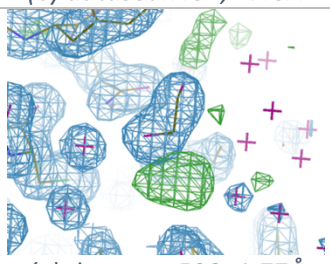
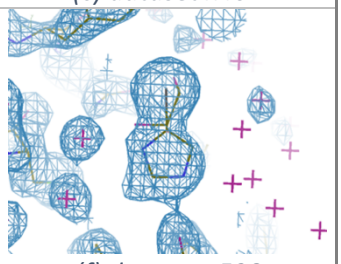
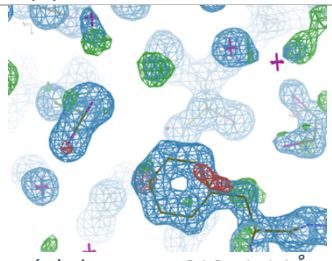
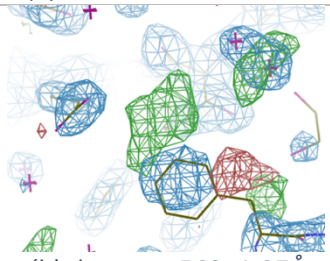
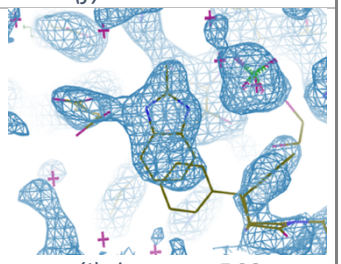
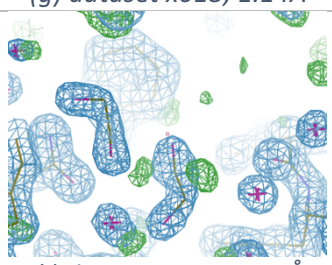
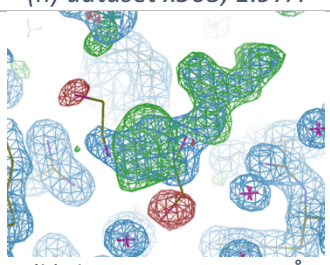
Protein	Ground-state only (reference dataset)	Ground-state only (ligand-bound dataset)	Ligand-bound state (PanDDA event map)
JMJD2D	 (a) dataset x618; 1.14Å	 (b) dataset x401; 1.48Å	 (c) dataset x401
BAZ2B	 (d) dataset x645; 1.49Å	 (e) dataset x538; 1.77Å	 (f) dataset x538
JMJD2D	 (g) dataset x618; 1.14Å	 (h) dataset x568; 1.97Å	 (i) dataset x568
BRD1	 (j) dataset x108; 1.38Å	 (k) dataset x049; 1.46Å	

Figure 5.11. ***Determination of the different states for the example datasets.*** *First two columns: $2mF_O$-DF_C maps contoured at 1.5σ (blue) and mF_O-DF_C maps contoured at ±3σ (green/red). Last column: PanDDA event maps (blue) contoured at (c,f) 2σ or (i) 1σ. Resolutions are as indicated. First column: A reference dataset provides the ground-state model of the crystal. Centre column: The ground-state refined into a ligand-bound dataset leaves (generally uninterpretable) residual density for a superposed state. Last column: The PanDDA event map provides clear density for the ligand-bound model of the crystal (the superposed ground-state model is shown for reference). (a-c) Dataset JMJD2D-x401. (d-f) Dataset BAZ2B-x538. (g-i) Dataset JMJD2D-x568. (j,k) Dataset BRD1-x049: the event map is not shown since it is not required.*

5.2.2 Results from comparing modelling approaches

I now present several cases where the inclusion of a complementary solvent model leads to a better description of the crystal, and thereby a higher-quality ligand model. I applied the process described in the previous section to four models from Chapter 4; details of the models are shown in Table 5.3-Table 5.6.

Table 5.3. ***Crystallographic parameters and ligand model scores for JMJD2D-x401.***

Model	Mean Phase Diff. (°)	R-work / R-free	Occ	RSCC	RSZD	RSZO/ OCC	B-factor Ratio	RMSD (Å)
Solvent Only	2.92	0.129 / 0.171	-	-	-	-	-	-
Ligand Only	9.97	0.147 / 0.195	0.79	0.44	4.3	0.76	3.54	0.26
Ensemble	-	0.127 / 0.171	0.26	0.87	0.5	5.38	1.47	0.01
Degraded Ensemble	24.17	0.241 / 0.290	0.27	0.77	1.0	4.81	1.40	0.05

Table 5.4. ***Crystallographic parameters and ligand model scores for BAZ2B-x538.*** *The occupancy of the bromine was refined separately; occupancy shown in the table is for the other atoms of the ligand.*

Model	Mean Phase Diff. (°)	R-work / R-free	Occ	RSCC	RSZD	RSZO/ OCC	B-factor Ratio	RMSD (Å)
Solvent Only	2.95	0.183 / 0.217	-	-	-	-	-	-
Ligand Only	4.15	0.184 / 0.215	0.68	0.92	3.6	1.62	1.41	0.20
Ensemble	-	0.182 / 0.216	0.41	0.96	1.6	3.41	1.04	0.02
Degraded Ensemble	31.06	0.311 / 0.363	0.29	0.90	0.1	1.72	1.18	0.70

Table 5.5. ***Crystallographic parameters and ligand model scores for JMJD2D-x568.***

Model	Mean Phase Diff. (°)	R-work / R-free	Occ	RSCC	RSZD	RSZO/ OCC	B-factor Ratio	RMSD (Å)
Solvent Only	3.86	0.157 / 0.219	-	-	-	-	-	-
Ligand Only	4.38	0.164 / 0.222	0.80	0.78	3.40	2.75	1.23	0.16
Ensemble	-	0.159 / 0.220	0.51	0.83	1.20	4.31	1.13	0.03
Degraded Ensemble	28.48	0.274 / 0.332	0.59	0.71	0.50	2.54	1.19	0.17

Table 5.6. ***Crystallographic parameters and ligand model scores for BRD1-x049.***

Model	Mean Phase Diff. (°)	R-work / R-free	Occ	RSCC	RSZD	RSZO/ OCC	B-factor Ratio	RMSD (Å)
Solvent Only	4.22	0.186 / 0.216	-	-	-	-	-	-
Ligand Only	2.20	0.183 / 0.213	0.89	0.95	2.00	4.38	1.26	0.03
Ensemble	-	0.182 / 0.212	0.84	0.96	1.60	5.95	1.20	0.01
Degraded Ensemble	34.46	0.341 / 0.380	0.77	0.91	0.10	3.77	1.14	0.07

JMJD2D-x401: Binding of the ligand across a bound substrate mimetic

To demonstrate the process of modelling both states, I first present an example where a strongly bound substrate mimetic is superposed with a weakly-bound soaked ligand, and an ensemble is clearly necessary. N-oxalylglycine (NOG) is tightly bound at high occupancy (~90%) in the ground-state crystal form of JMJD2D, as shown in the reference dataset (Figure 5.11a). A soaked ligand binds across this substrate mimetic in a small fraction of the crystal, as shown in the PanDDA event map (Figure 5.11c). The superposition of the two states leads to a good model, with negligible amounts of difference density remaining (Figure 5.12b).

As expected, refinement of the ligand without the superposed NOG results in a poor-quality model (Figure 5.12a), because a large fraction of the crystal is locally unrepresented; refinement of the ensemble results in the better model for the ligand, scoring well across all five metrics. On the radar validation plot (Figure 5.13a) this is shown as the ensemble-model line (green) being entirely contained within the ligand-only line (red) – the closer the line is to the centre of the plot, the better the model. Optimal refinement of the ligand requires the superposed ground-state conformation to be present in refinement.

The degraded protein model (Figure 5.12c) has a 24° average phase difference to the high-quality ensemble model, increasing the R-free from 17% to 29%. However, the model of the ligand is not significantly degraded and still scores well on all five model validation metrics, although worse than the ensemble model with high-quality phases. In this case, the correctness of the local model is more important than the convergence of the global phases.

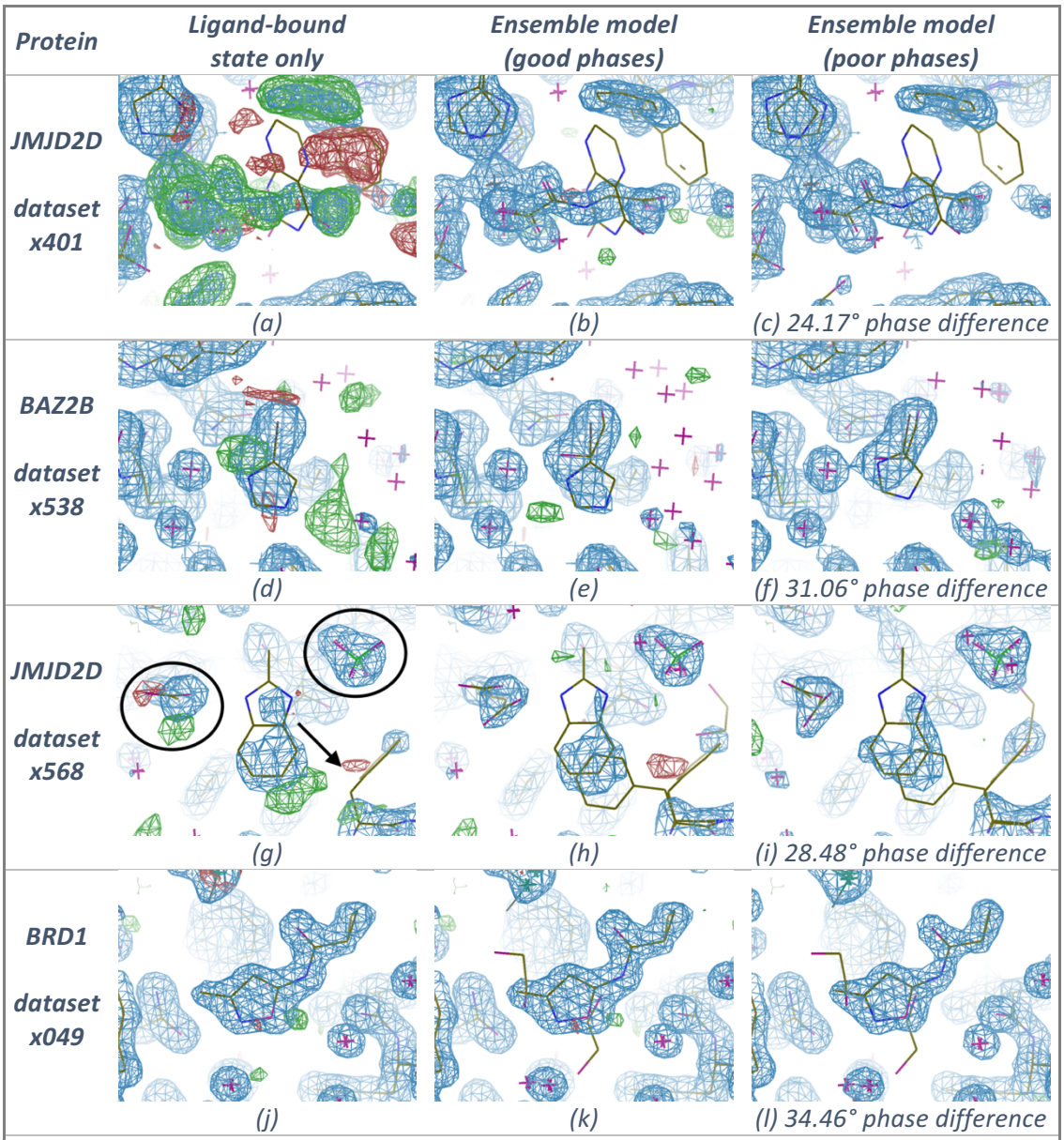


Figure 5.12. ***Ensemble models consistently leave less residual difference density than ligand-only models.*** *All images: $2mF_O$-DF_C maps contoured at 1.5σ (blue) and mF_O-DF_C maps contoured at ±3σ (green/red). First column: Refinement with the ligand model only. Centre column: Refinement of the crystal as an ensemble of states. Last column: Refinement of the crystal as an ensemble of states with a degraded protein model (phase difference as indicated, relative to the ensemble model). (a,d,g) Modelling the ligand but removing the ground-state leads to difference density for the absent state, and in (d) the ligand moves into density for the ground-state. (j) Removing the ground-state for a high occupancy ligand (refined value 0.83 when refined as ensemble) does not lead to discernible difference density. (b,e,h,k) Refinement of ensemble models explain all of the observed density, and ligands do not move from the fitted pose (confirmed by the validation plots in Figure 5.13). (c,f,i,l) Refining with degraded phases leads to only minor visual differences, except in (f) where the ligand moves relative to the fitted pose.*

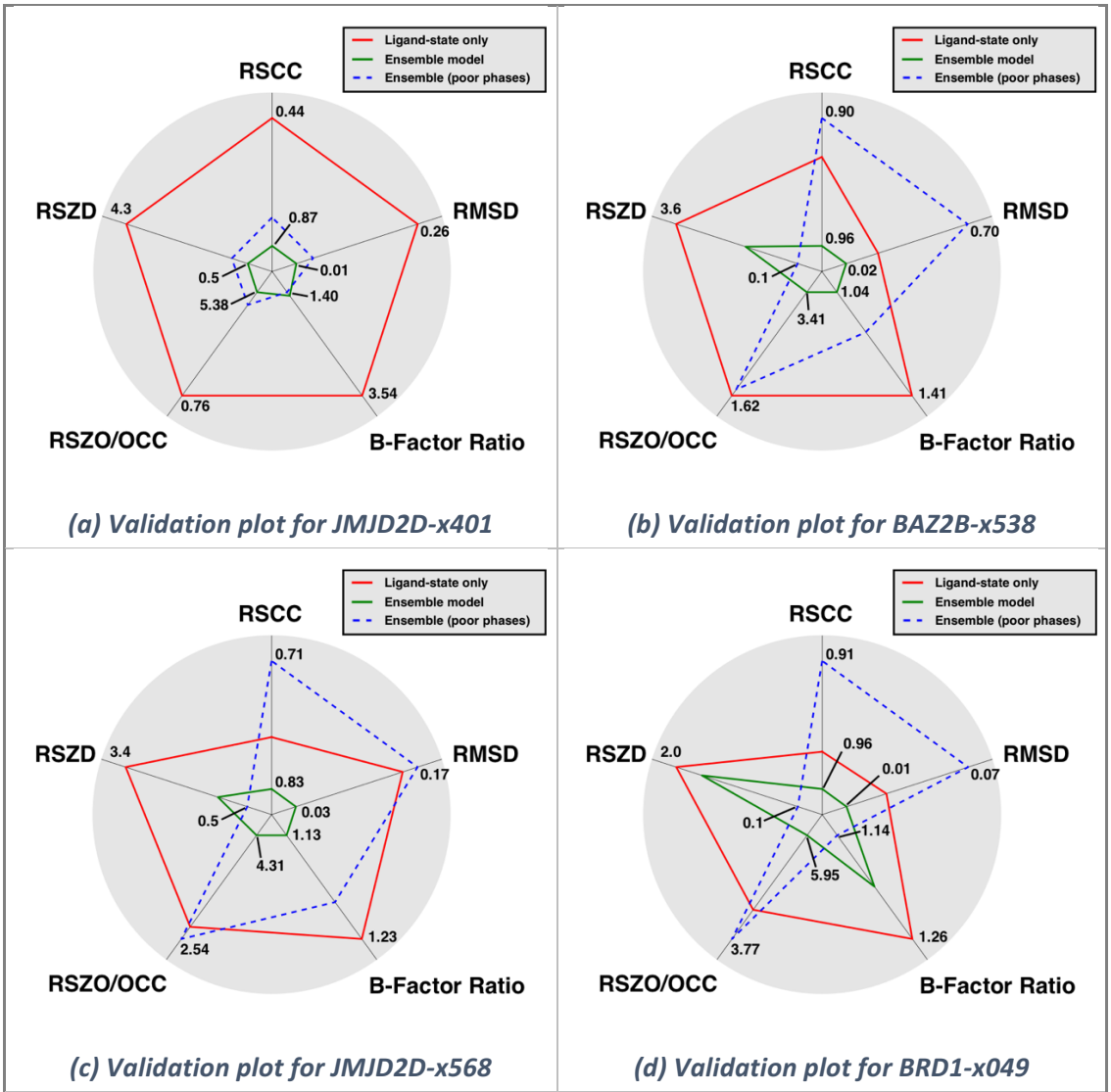


Figure 5.13. ***Validation plots for the different modelling approaches****: axes are not fixed, but have been scaled relative to minima and maxima of the plotted values, and only the minimum and maximum values are marked on the axes; for all model scores refer to Table 5.3-Table 5.6. (a) Plots for Figure 5.12a-c. The plot confirms the visual inspection of the electron density; the ligand scores are improved across all metrics when refined as an ensemble, relative to the ligand modelled alone. The absence of the superposed substrate model has a greater effect on the ligand model than the degradation of the protein model phases. (b) Plots for Figure 5.12d-f. The ensemble model provides the best model for the ligand. The RSZD is decreased in the degraded-phase model for reasons explained in the main text, and is not related to an improved model. (c) Plots for Figure 5.12g-i. Once more, the model statistics are improved with the addition of a superposed solvent model, with the caveat that the lower RSZD for degraded phases is not indicative of an improved model. (d) Plots for Figure 5.12j-l. The inclusion of the solvent model still increases the quality of the model compared to when it is omitted, albeit marginally. The degraded phase model has lower B-factor ratios than either of the other two models due to a decrease in the B-factors of the ligand and a corresponding drop in occupancy.*

BAZ2B-x538: Binding of a ligand in place of a solvent molecule

In BAZ2B-x538 (discussed in Chapter 4), an ethylene glycol is bound in a semi-ordered fashion, with a superposed ligand, to the asparagine in the binding site. The solvent model derived from a reference dataset is not optimal, and some difference density remains even when a ligand is not present (Figure 5.11d). Refinement with the ground-state model in the ligand-bound dataset does not lead to significant additional difference density, as the refined solvent model masks the presence of the ligand's bromine (Figure 5.11e).

The PanDDA map, however, shows clear evidence for the ligand (Figure 5.11f), as discussed in Chapter 4. Refinement with only the bound state causes the ligand atoms to be pulled into the density for the ethylene glycol, and difference density remains (Figure 5.12d). Refinement of the ensemble leads to a good model (Figure 5.12e), with all density well-explained, and no movement of the ligand from the fitted pose.

Refinement of the degraded-phase model (Figure 5.12f) also causes the ligand to move relative to the fitted position. In this case, the absence of the superposed model and the quality of the model phases are both important for the quality of the final ligand model; this is reflected by the validation metrics (Figure 5.13b).

It is noteworthy that the RSCC of the ligand in all models is greater than 0.9, showing that whilst a large RSCC is necessary for a good model, it is not sufficient: it does not account for difference density. As explained in section 5.2.1, the RSZD of 0.1 for the degraded-phase ligand model, which would normally indicate a very good model, is affected by the degraded phases.

JMJD2D-x568: A binding ligand overlaps with alternate conformations of a sidechain

Another ligand in a JMJD2D dataset binds along with a sulphate to a putative allosteric site. Refinement with the ground-state conformation leaves residual unmodelled difference density (Figure 5.11g,h). The pose and identity of the ligand is clearly revealed in the PanDDA event map (Figure 5.11i), revealing the re-ordering of two sidechains and that the ligand is superposed on the ground-state conformation of the phenylalanine.

Upon inspection of the refined ensemble model (Figure 5.12h), it was suggested by another experienced crystallographer that the ground-state conformation should be deleted and the ligand-bound state refined as the sole conformation. This recommendation supports my observation that the pervading convention – to generate only a single conformation of the crystal wherever possible – dominates even in the face of clear evidence that multiple states are present. The density in the area of overlap between the ligand and the phenylalanine is significantly stronger than over the rest of either residue, and difference density is present when either state is refined separately (Figure 5.11h, Figure 5.12g). The residual density from the ligand-state-only model might further tempt a crystallographer to move the model down and right by ~1Å (as indicated by the arrow in Figure 5.12g), although this causes clashes with the C_β of the phenylalanine and adversely affect the interactions that the ligand makes with the aspartate and the sulphate (marked with ovals in Figure 5.12g). All evidence points towards the presence of multiple states in the data, and therefore these multiple states should be present in the model.

The phase degradation in Figure 5.12i (mean phase difference to ensemble model 28.48°) degrades the ligand model RSZO and the B-factor ratio to a similar level as the omission of the ground state model, and significantly degrades the RSCC (Figure 5.13c). Again, a decrease in RSZD is observed with the decrease in phase quality. The ensemble model provides the best interpretation of the experimental data.

BRD1-x049: Traces of the ground state remain, even for a high occupancy ligand

One ligand screened against the bromodomain of BRD1 binds strongly in the principal binding site (Figure 5.11j,k), with a refined occupancy of 83-89% (multi-state and ligand-only refined occupancies respectively). In the reverse case of JMJD2D-x401, the ligand occupancy is much higher than the ground-state occupancy, and this ligand would conventionally be modelled at unitary occupancy.

Once more, inclusion of the ground-state solvent improves the model quality, although in this case only marginally (Figure 5.12j,k & Figure 5.13d). Even for this strong binder, visual traces of the ground-state model remain: contouring the $2mF_o$-DF_c map to zero rmsd shows density for ground-state solvent (Figure 5.14). That density for the EDO molecules is not seen at typical contour levels (e.g. in Figure 5.12k) again exemplifies a problem with visualising multi-state superpositions: weak features may be hidden when viewed in superposition with another state (cf. the examples of weak ligands in Chapter 4). Though contouring maps to low levels is not normally productive for modelling, due to the appearance of multiple overlapping molecular states that can impede interpretation, here the density may be used as weak evidence (allowing for the possibility of phase bias) in support of our Bayesian approach: density is present in regions where we expect it to be.

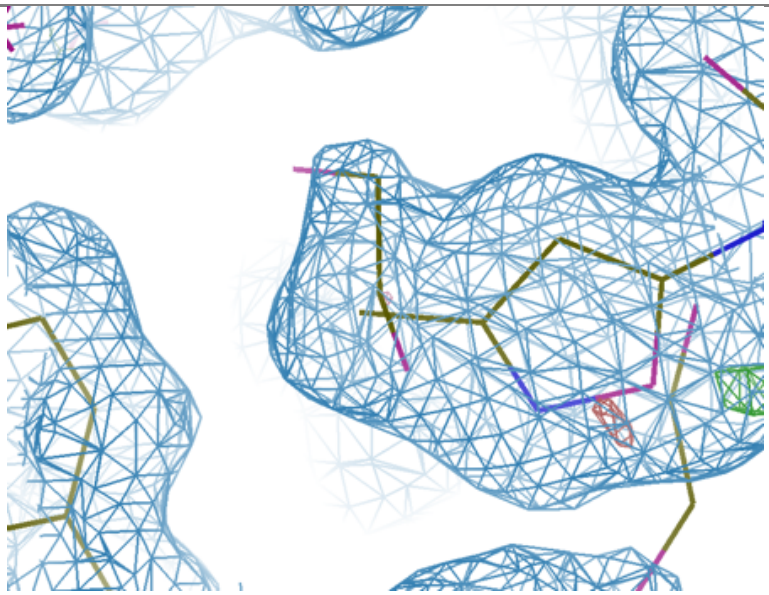


Figure 5.14. ***Weak evidence for the ground-state model is still visible in refined maps****. No evidence is seen that would support the removal of the superposed solvent model. $2mF_o$-DF_c map (blue) contoured at 0.0σ, mF_o-DF_c map (green/red) contoured at ±3σ.*

Phase degradation does not lead to a significant change in the model (Figure 5.12l), but degrades the RSCC, RMSD and the RSZO more than the absence of the solvent model, with a decrease in RSZD as previously (Figure 5.13d). Here the B-factor ratio is seen to be lower for the phase-degraded model than for the other models, due to a decrease in the B-factors of the ligand by two, and a corresponding decrease in the occupancy to 0.77; this behaviour demonstrates the ambiguity that can be observed in simultaneous refinement of B-factors and occupancies.

5.2.3 Discussion

The examples I have presented here show that there is consistent evidence that ground-state molecules are superposed in the experimental data on top of binding ligands across a range of non-unitary occupancies. I have also shown that the inclusion of a superposed ground-state model, obtained from a reference dataset, improves the quality of obtained ligand models in all cases.

In the case of some weak ligands, the ground-state model is crucial for the refinement of the protein/ligand complex (JMJD2D-x401); in another case, it acts simply to remove "extraneous" difference density that could be interpreted by an over-zealous modeller as being caused by a ligand in multiple conformations (BAZ2B-x538). The modelling approach can affect the interpretation of inter-molecular interactions (JMJD2D-x568), and in the case of high occupancy, a superposed ground state marginally improves the ligand model, alongside providing a complete model of the crystal (BRD1-x049).

With the current increase in popularity of experiments such as fragment screening by crystallography amongst academic groups, the PDB is set to see a sharp increase in structures that contain binders with considerably less than unitary occupancy (e.g. Schiebel et al. 2016). I have shown that the models of such partial-occupancy ligands benefit from the inclusion of a superposed ground-state; from these results, I propose that a new standard modelling convention is adopted, where bound ligands are modelled as a superposition of states *wherever possible*. *We should assume that the ground-state is present in a ligand-bound crystal until it is proven absent*; this is contrary to the current convention, which appears to assume the opposite.

Experimentally, determination of the ground state is no extra burden, as an unbound reference dataset is normally already available when soaking experiments are performed. Computationally, however, this approach will require the implementation and availability of tools that allow the trivial generation of ensembles from multiple single-state models; the PanDDA implementation goes some way towards achieving this, and tools such as autoBUSTER also implement many of the required features.

Performed correctly, the addition of a superposed ground-state model allows no further degrees of freedom for the crystallographer, as the ground-state model is solely determined in an orthogonal reference dataset. However, the addition of atoms to the model – and the refinement of occupancies – introduces more parameters in refinement; additional restraints will be required to prevent local overfitting of the model to the data. Yet, utilisation of prior knowledge in the modelling process should in general lead to higher quality models and thereby crystallographic phases, and should ultimately contribute to closing the R-factor gap (Holton et al. 2014).

From a pragmatic perspective, I propose that the ground state should only be removed from the ensemble model if the occupancy of the refined ground-state conformer is $\lessapprox$10%: in this case the benefit of the ground-state model in stabilising the ligand model parameters is likely negligible, and outweighed by the negative consequences of allowing additional parameters in refinement, which may allow overfitting.

5.2.4 Problems in refinement of the ensembles

Correct parameterisation of the ensemble model can lead to complicated models and refinement constraints that are currently not supported by some refinement programs (REFMAC, Murshudov et al. 2011; phenix.refine, Afonine et al. 2012): in some cases, not shown here, I have found that refinement of multiple conformer models permitted occupancies for amino acids that summed to greater than unity. Furthermore, workarounds are needed during model building due to restrictions of conformer models (Figure 5.15). Further work is required to generate occupancy and structural restraints that allow complex ensemble refinement in the general modelling case, without permitting unphysical atomic models; many of these features are available in

autoBUSTER (Smart et al. 2012), and utilisation of this tool in solving these problems is the subject of future work. Procedural generation of ensembles and the corresponding restraints will be critical to the uptake of this multi-state modelling approach.

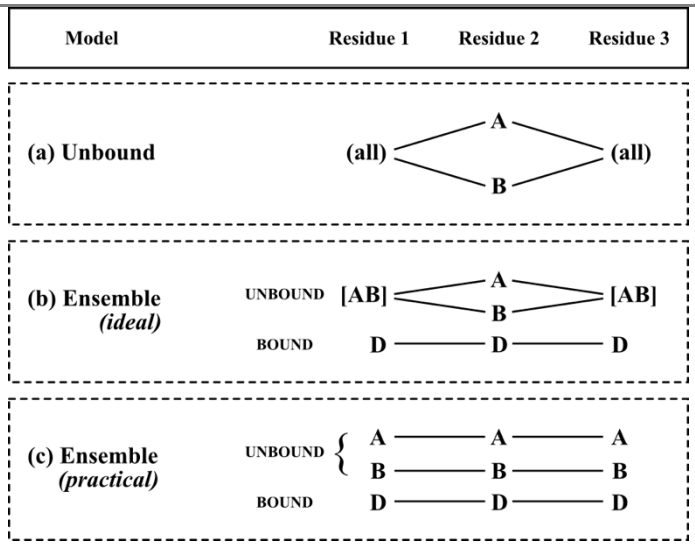


Figure 5.15. ***Limitations of alternate conformers require workarounds in modelling****. Conformers in crystallographic models do not support branching of conformers – where alternate sub-conformations can be added to existing alternate conformations – so workarounds are required where complex alternate conformations are present. (a) The unbound structure contains a residue in alternate conformations. These are assigned to conformers A and B. (b) In the ideal case, a new conformer (D) could be introduced without editing the existing structure. However, it is not possible to create the [AB] conformer (conformers must be either A or B). If the [AB] conformation is labelled as A, then the B-conformation of residue 2 becomes disconnected in refinement. (c) The workaround is to enumerate all conformations of [AB] for each of the surrounding residues; this creates multiple conformations A and B of residue 1 and residue 3.*

Lastly, I have investigated the impact of phase degradation on ligand model quality, compared to the effect of local modelling. I conclude that the modelling of local ground-state atoms is generally far more important than convergence of the global model, especially as global errors in typical modelling situations are likely to be much less than the 24-35° phase error introduced here. “Tweaking” of sidechain conformations and water molecules in distant regions in the model to improve phases is likely not of importance if the binding of a ligand is the feature of interest. However, the modelling of the environment around and “under” any ligand is conversely of great importance.

5.3 The quality of low-occupancy ligand models

Multiple papers have been published recently that discuss the quality of ligand models in the PDB (e.g. Weichenberger et al. 2013; Deller & Rupp 2015). The conventional cutoff used for a "good" ligand model is a minimum correlation of 0.7 with the experimental electron density; however, some ligands can express lower correlations, and still be considered acceptable models (Weichenberger et al. 2013).

As discussed in the previous section, most ligands in the PDB are not modelled with a superposed unbound state; the observed RSCC is thus lower than would be observed were the crystal modelled correctly (Figure 5.13). The results from the previous sections and Chapter 4 lead me to conclude that a better modelling paradigm allows stricter limits to be imposed on the cutoff for a poor model: better models lead to more reliable validation metrics.

In this section, I plot the validation metrics for all models from Chapter 4 against the occupancy of the ligand (section 5.3.1). The majority of the ligand models are high quality, regardless of the occupancy of the ligand. We conclude that the "weakness" (occupancy) of a ligand is not a good reason for it to be poorly modelled; there is no reason for ligands at high resolution to express poor validation scores, if modelling is performed correctly.

Furthermore, the occupancy of the ligand correlates well with the background-density correction (BDC) factor used to create the event maps (section 5.3.2). This validates the approach used in Chapter 3 & 4, and offers a possibility to investigate phase bias in future investigations.

5.3.1 Relationship between occupancy and validation metrics

To investigate the relationship between the quality of ligand models and the occupancy of the ligand, the five standard validation metrics and the real-space R (RSR) are plotted against the refined occupancy of the ligand for all the ligands identified in Chapter 4 (Figure 5.16); the RSR is calculated using EDSTATS as with the other density scores.

The RSCC decreases (and the RSR increases) with a decrease in the ligand occupancy (Figure 5.16a,b). This agrees with the observations in Tickle (2012) of a relationship between the B-factor and the RSCC. Both decreases in RSCC are likely due to decreases in the signal-to-noise ratio; for reduced occupancy, less density for a feature increases the relative amount of noise, and similarly for the blurring with the B-factor. A second possible effect is due to the increased ground-state occupancy as the occupancy of the ligand decreases. Where the ground-state is formed of multiple disordered solvent molecules, it is difficult to represent the ground state with an atomic model; this poor modelling of the ground-state solvent may lead to residual unmodelled density that will degrade the validation metrics of the superposed ligand model.

The RSZD and the normalised RSZO, on the other hand, show no strong relationship with the occupancy of the ligand (Figure 5.16c,d); this supports our normalisation of the RSZO by the occupancy of the ligand, and confirms the argument in Tickle (2012) that these metrics are robust measures of model quality and not correlated to model parameters.

The B-factor ratio of the refined ligands shows an increase for low-occupancy ligands (Figure 5.16e), indicating that the B-factors of the ligands are less stable at lower

occupancies, in agreement with the known instability of occupancy and B-factor refinement. The RMSD, however, shows no correlation with occupancy (Figure 5.16f).

The use of the multiple validation metrics has emphasised that low-occupancy ligands do not need to result in poor-quality models: where signal is confidently identified, and modelled correctly, resulting models are of high quality.

5.3.2 Relationship between occupancy and background-density correction factor

The refined occupancy of the ligand relates well to the background density correction (BDC) factor used in the construction of the event map (Figure 5.17); this confirms that the contrast maximisation approach described in Chapter 3 is indeed extracting information in the event map. This is despite the instabilities that were observed in the joint refinement of occupancies and B-factors, as discussed previously.

The occupancy of the ligand is consistently higher than the event map pseudo-occupancy (1-BDC); since the algorithm is a contrast-maximisation approach, event map density for changes appears somewhat stronger than density for unchanged atoms (typically, the surrounding protein). The overestimation of BDC indicates that the event maps hereby may lead to false negatives, manifesting as uninterpretable density.

The overestimation of BDC further combines with any phase effects due to the absence of the ligand from the model to create the relationship shown in Figure 5.17. An improved method for the estimation of BDC – confirmed on a set of simulated datasets, where the correct BDC is known in advance – may allow multi-dataset experiments to become a way of studying phase effects in model building.

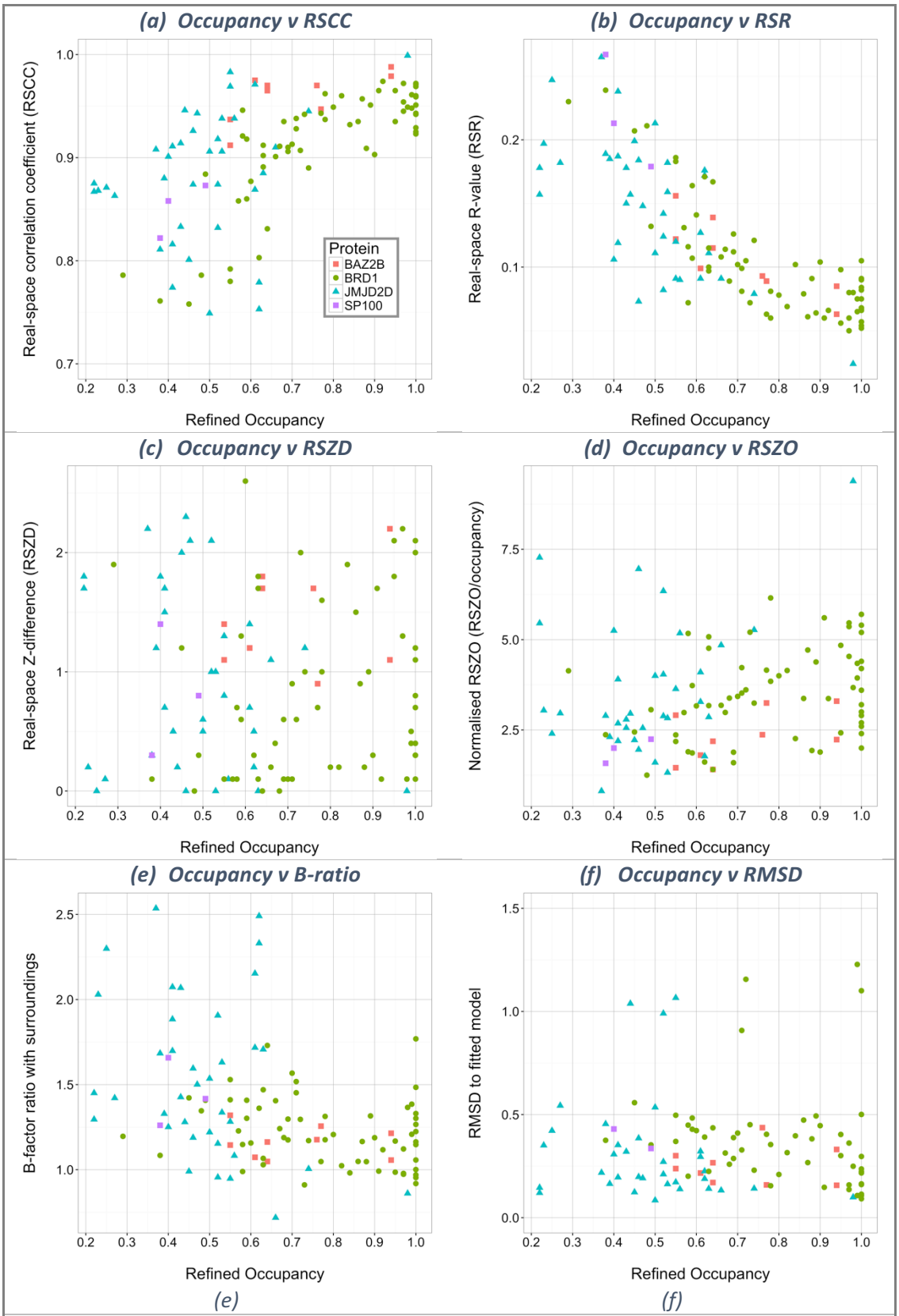


Figure 5.16. ***The relationship between occupancy and validation metrics.*** *The validation scores for all five validation metrics (plus RSR) are plotted against the refined occupancy for all of the refined ligands from Chapter 4. The correlations between the validation metrics and the refined occupancies are (a) 0.62,(b) -0.79, (c) -0.00, (d) 0.22, (e) -0.43 and (f) -0.20.*

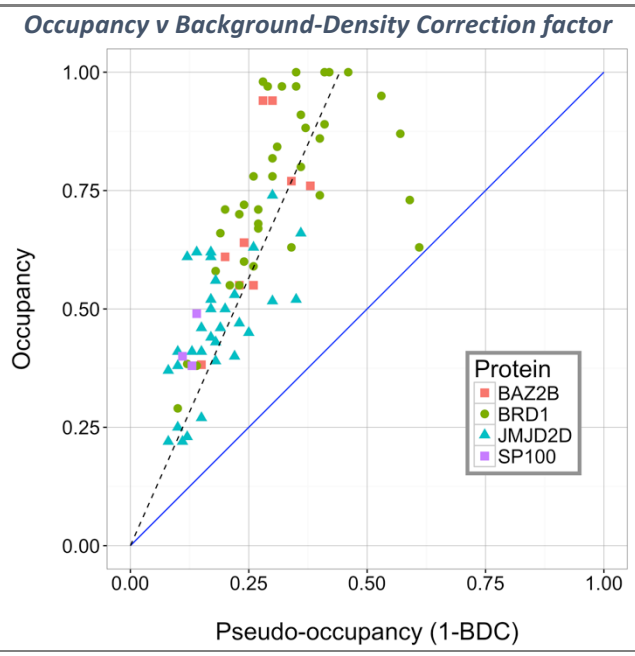


Figure 5.17. ***Relationship between occupancy and the Background-Density Correction (BDC) factor of the event maps.*** *The pseudo-occupancy of the event map (1-BDC) relates well to the refined ligand occupancy for the four fragment screens from Chapter 4. The linear fit (dashed line) has a slope of 2.262; the linear model was constrained to pass through the origin. The line of identity (blue line) is also shown.*

5.3.3 Discussion

The relationship between the refined occupancy of the ligand and the validation metrics shows that so-called "weak" ligands do not necessarily lead to poor models. Treated correctly, there is no reason for a partial-occupancy ligand (at the resolution range covered here) to be modelled badly, or score badly during validation.

Stable simultaneous refinement of the occupancy and B-factors remains problematic, resulting in large refined B-factors, especially for very low-occupancy ligands, and where the solvent model is not easily reduced to an atomic model; further work is required, involving the use of external restraints to constrain to a reference structure in refinement (e.g. Smart et al. 2012; Nicholls et al. 2012; Headd et al. 2012), and the use autoBUSTER (Bricogne et al. 2011), which implements occupancy refinement robustly.

In light of the relationship in Figure 5.17, the occupancy of modelled ligands is now set to twice 1-BDC in the PanDDA implementation. However, this simply serves as an initialisation for occupancy refinement, as the currently-estimated BDC is not a physical parameter, nor a refined parameter, and must not be over-interpreted.

5.4 Chapter Summary

From the re-analysis of the Schiebel *et al.* datasets in section 5.1, I conclude that difference-density-based ligand-detection methods are not suited to the detection of partially-occupied ligands; even though I confirmed all the binders of Schiebel *et al.*, the difference density as evidence of binding is not wholly convincing, especially when compared to the clarity of the event maps. Schiebel *et al.* expertly interpret the data, yet the difference-density-based identification approach leads to incomplete models of many ligands, affecting the interpretation of the resulting models: ligands may be thought to be disordered, where in fact they are ordered at low occupancy.

The PanDDA approach makes the conventional and problematic *fit-refine-assess* approach both practically and theoretically unnecessary; a single local modelling step fully validates an interpretation, even when the model retains problems elsewhere, leading instead to an *assess-fit-refine* procedure. The only validation that is required is to ensure that the model matches the density, that the model parameters refine well, and that the ligand model does not move under refinement: the five validation metrics used throughout this thesis measure all of these characteristics.

Modelling of the superposed solvent state is crucial for the refinement of low-occupancy ligands, and further leads to higher quality models for high-occupancy ligands. Lastly, though there is some work to be done to ensure the stable refinement of low-occupancy ligands, there is no evidence that low-occupancy ligand models should be of poor quality; where there is good evidence for the bound ligand, this is reflected in a high-quality model by all metrics.

Chapter 6

The PanDEMIC method: A multi-dataset approach to detecting conformational heterogeneity and structural variability

---

"There must be enough information here to build a better model."

In the previous chapters, I have concentrated on identifying states that are unique to a particular dataset (i.e. binding ligands). Through the PanDDA method, I have shown that multi-dataset methods are a promising way to deconvolute a superposition of multiple crystal states; the analysis of multiple datasets allows the sensitive detection of weak perturbations. In this chapter, I extend the analysis to the detection of crystallographic states which are *variably* present in *all* datasets of a particular crystal form of a protein: rotameric states and conformationally heterogeneous regions.

Multiple papers have shown that conventionally-performed cryocooling of crystals distorts the crystallographic heterogeneity of rotamer conformations relative to their room-temperature distributions (e.g. Keedy et al. 2014; Fraser et al. 2011). These perturbations are considered an unfavourable artefact of a cryogenic-temperature crystallographic experiment; the biologically relevant sub-states of the protein may be quenched at cryogenic-temperatures and thus be undetectable in the crystallographic data (Fischer et al. 2015).

In this chapter, I outline an approach that investigates variation that is observed in crystallographic datasets of the same crystal form; I present preliminary data and outline an approach to utilise these variations for the improved determination and modelling of heterogeneity in crystallographic data.

6.1 A novel hypothesis and the PanDEMIC paradigm

I begin with two observations. Firstly, the perturbation caused by cryo-cooling results in the quenching of some sidechain rotamers: the occupancy of particular sidechain conformations is reduced or increased in the crystallographic data, relative to a room temperature dataset (Keedy et al. 2014). Secondly, the variation between equivalent cryo-cooled crystals is significant (Fraser et al. 2011); the process of cooling a crystal is fundamentally stochastic, and the rate of cooling depends on factors such as the size of the crystal and e.g. the speed at which it is plunged into liquid nitrogen.

The combination of these two observations leads me to hypothesise that the stochasticity of the cryo-cooling process will lead to variable amounts of conformational quenching in different crystals, resulting in randomly-perturbed occupancies for each sidechain rotamer across the datasets.

The occupancy of a sidechain rotamer is directly reflected in the strength of the density at the location of the rotamer; analysis of the variation in the electron density maps should reveal regions with variable rotamer occupancies, allowing the presence of structural heterogeneity to be detected. Single-conformer rotamers cannot be quenched, therefore variation will only reveal where multiple conformations of the residue are present in the crystal.

Through analysis of the electron density variation, using the s_m map from the PanDDA method (section 3.5), and comparison with all possible rotamers, methods will be developed in the future to determine which rotamers are present in the crystal. Accurate identification of alternate sidechain conformations will lead to improved multi-state models of the protein, permitting a better understanding of biological function, as well as better crystallographic model phases, ultimately leading to a closing of some of the R-factor gap (Holton et al. 2014).

I call this potential approach PanDEMIC (Pan-Dataset Ensemble Modelling of Isomorphous Crystals); by simultaneously analysing multiple perturbed datasets, there is the potential for all datasets to be better modelled.

6.2 Preliminary detection of rotameric heterogeneity

In this section, I analyse the structures and electron density maps of the BAZ2B dataset from Chapter 4 to look for evidence of the hypothesised variable rotameric heterogeneity described above. All datasets were collected at cryogenic temperatures.

6.2.1 Evidence for variable occupancy of sidechain rotamers

To look for evidence of the variable quenching of conformers, I refined the occupancies of sidechains with multiple conformers in all fragment screening datasets of BAZ2B. The Dimple-refined structures (CCP4; Winn et al. 2011) were re-refined using phenix.refine (Afonine et al. 2012, version 1.9-1682) with default parameters; the occupancy of sidechains are refined where multiple conformers are present. Allowing for the instability of occupancy refinement, as discussed in previous chapters, the occupancies of individual sidechain conformations are indeed observed to vary over the 200 datasets; for some residues, the occupancy range is more than 20% (Figure 6.1).

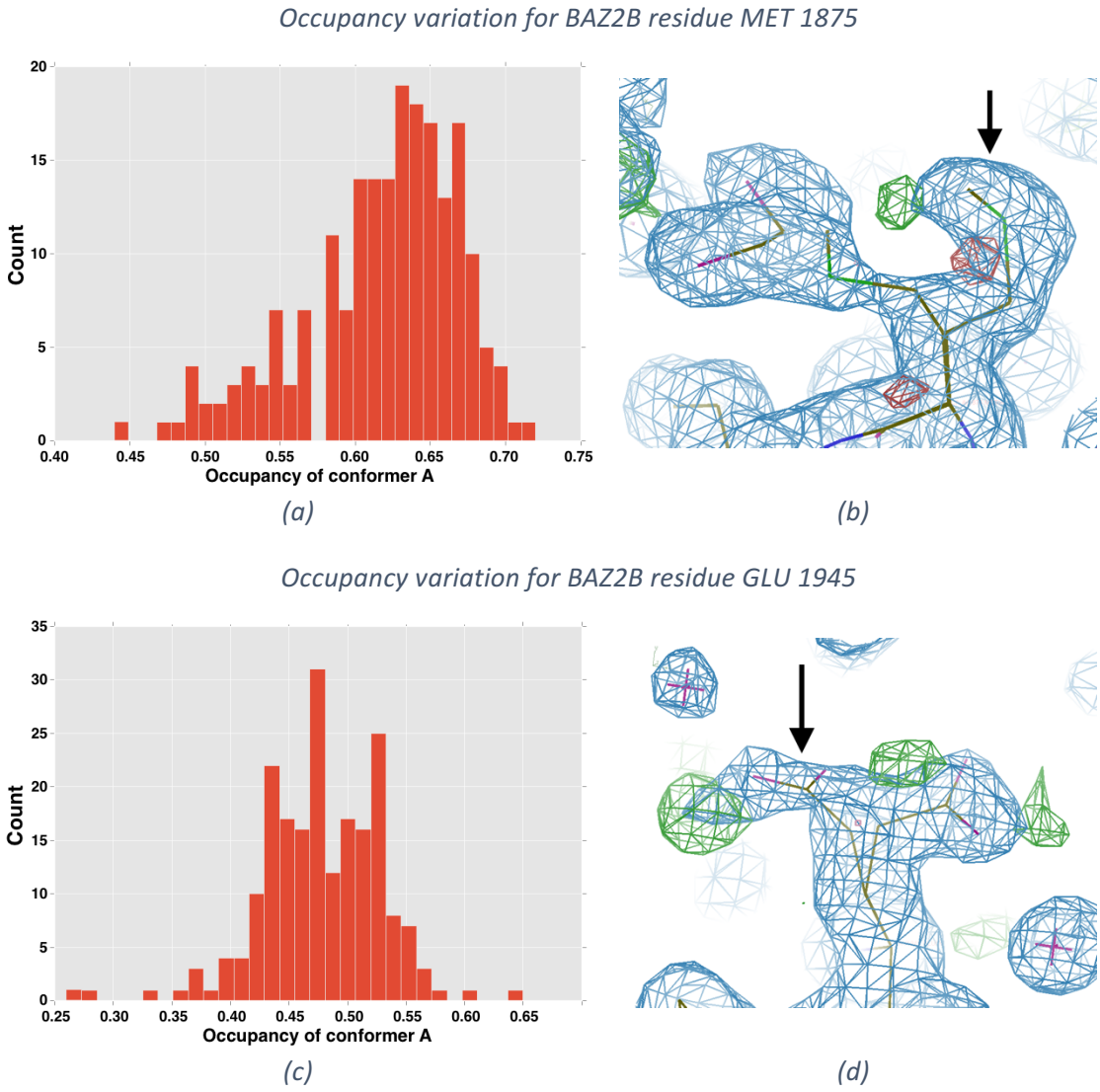


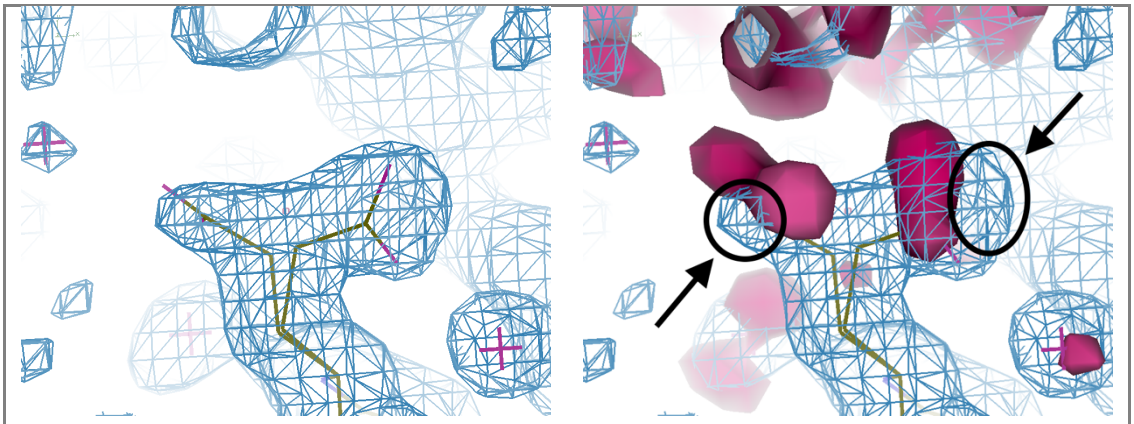
Figure 6.1. ***Occupancy variation of modelled alternate conformers for BAZ2B.*** *(a,c) Occupancy distributions for two residues after re-refinement with phenix.refine. (b,d) Images of electron density for residues are shown for dataset x481 (resolution 1.65Å); $2mF_o$-DF_c maps (blue mesh) are contoured at 0.5σ, and mF_o-DF_c maps (green/red mesh) are contoured at ±3σ.* ***Residue 1875 (MET):*** *(a) The occupancy distribution for conformer A of residue 1875 (indicated with an arrow in (b)); this residue is modelled in two conformations. Conformer A is on average the most populated conformer, with a mode of ≈63%, and a range of ≈45-70%. (b) The refined multi-conformer model (conformer A, occupancy 0.69). Conformer B of the methionine overlaps with an ethylene glycol molecule (assigned to conformer A). The varying occupancy of the conformation is confirmed by visual analysis; the sulphur of the minor conformation (conf. B) is clearly defined in some datasets and less so in others (not shown).* ***Residue 1945 (GLU):*** *(c) The occupancy distribution for conformer A of residue 1945 (indicated with an arrow in (d)), also modelled in two conformations. Conformer A is marginally less common on average, with a modal occupancy of ≈47%, and a range of ≈25-65%. (d) The refined multi-conformer model (conformer A, occupancy 0.47). The presence of difference electron density shows that the alternate conformations are not optimally modelled; the superposed solvent model corresponding to each state is missing from the model.*

6.2.2 Detection of sidechain conformational heterogeneity

To establish whether density variation can indicate conformational heterogeneity, I utilise the electron density variation analysis resulting from the PanDDA method. The s_m map from PanDDA produces an estimate for the error-corrected variation in the density at a point, across all datasets. A manual inspection of the s_m map of the BAZ2B analysis reveals several residues where heterogeneity of the sidechains was recognisable in the variation map, both where the conformational heterogeneity had been previously modelled (Figure 6.2), and where it had not (Figure 6.3). The contour level used in all figures, 0.25σ, though manually selected, reflects a very large change across the datasets, relative to the background variation (Figure 3.7b, page 90).

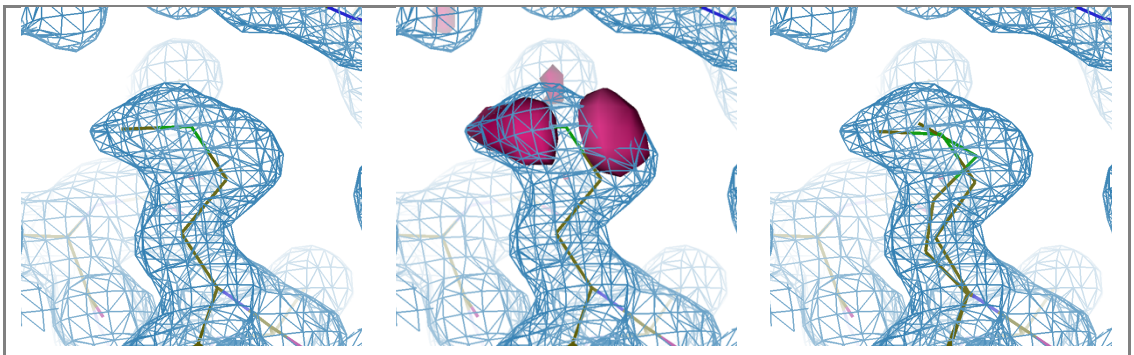
Datasets containing bound fragments may be systematically different to unbound datasets, and may affect the results. However, datasets with fragments bound were not removed from the analysis; fragments only affect small regions of the structure, and the number of binders is small compared to the total number of datasets used for analysis at a certain resolution.

The adjusted-variation map also identified a sidechain for which no density was observed in the mean $2mF_o$-DF_c map, even at low contour levels (Figure 6.4). The lack of visible electron density in Figure 6.4c does not imply the absence of the supposed rotameric state: the 0σ level of the $2mF_o$-DF_c map does not correspond to the noise level of the electron density map, and iso-contouring at any fixed value is not guaranteed to make a particular state visible.



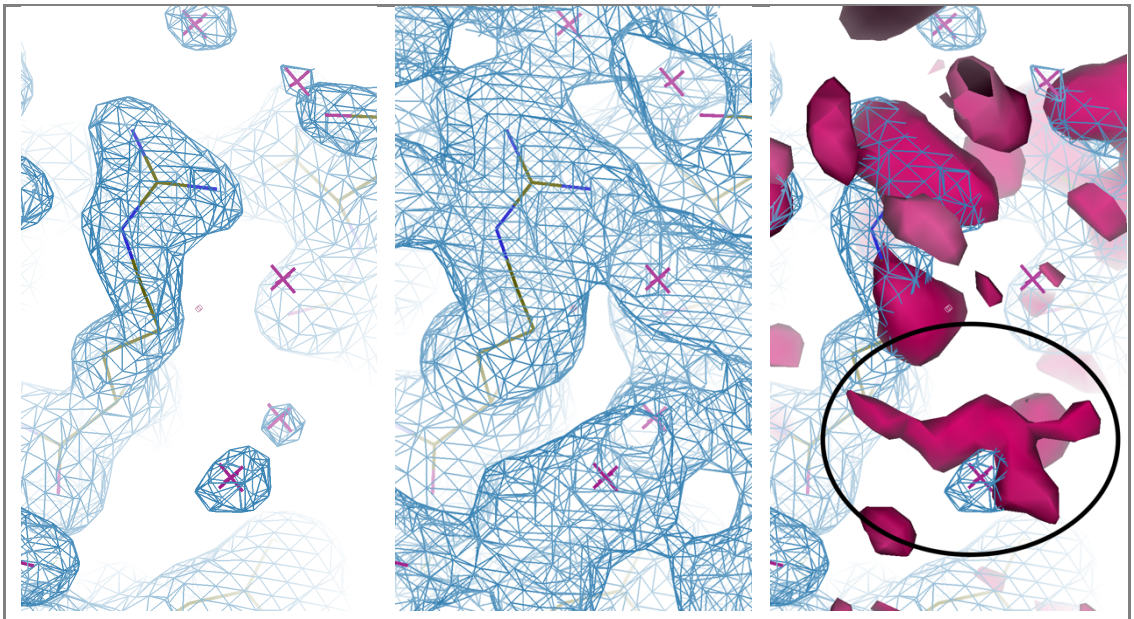
(a) (b)

Figure 6.2. ***Identification of modelled sidechain heterogeneity.*** *Electron density maps in the region of a solvent-exposed glutamate of BAZ2B (chain A, residue 1945), as in Figure 6.1d. All images: averaged $2mF_o$-DF_c map from 48 datasets at 1.7Å (blue mesh, 1σ). (a) The sidechain is modelled in two conformations and provides a reasonable description of the density (sub-optimality is discussed in Figure 6.1d). (b) The adjusted variation (s_m) map from PanDDA (hot pink blobs, 0.25σ) identifies the region as being heterogeneous, highlighting the positions of the carboxylate groups of the sidechain. The variation map allows us to identify which regions of the electron density (blue mesh) likely correspond to the unmodelled solvent molecules (indicated with arrows); the weaker, more diffuse, density of partially-disordered solvent molecules may explain the absence of visible solvent density variation in the variation map.*



(a) (b) (c)

Figure 6.3. ***Identification of unmodelled sidechain heterogeneity.*** *Electron density maps in the region of buried methionine of BAZ2B (chain A, residue 1880). All images: averaged $2mF_o$-DF_c map (blue mesh, 1σ, details as in Figure 6.2). (a) The sidechain was originally modelled in a single conformation. Minimal difference density was exhibited in the mF_o-DF_c difference maps at 3σ (not shown). (b) The adjusted variation (s_m) map from PanDDA (hot pink blobs, 0.25σ) identifies regions of variation, indicating the presence of conformational heterogeneity. (c) Proposed multi-conformer model for the residue. Regions of high density in the $2mF_o$-DF_c map and the s_m map were used to place the methionine's sulphur and the s_m map further indicates the direction of the methyl group.*



(a) (b) (c)

Figure 6.4. ***Identification of low-occupancy rotamers for which there is no evidence in the averaged electron density map.*** *Electron density maps in the region of solvent-exposed arginine of BAZ2B (chain A, residue 1952). (a) A single sidechain conformation is seen in the averaged $2mF_o$-DF_c density (blue mesh, 1.5σ, details as in Figure 6.2) with no evidence of alternate conformations. (b) The conventional $2mF_o$-DF_c density at very low contour levels (0σ) still shows no clear evidence of alternate conformations. (c) The adjusted variation (s_m) map (hot pink blobs, 0.25σ) clearly indicates the presence of a low-occupancy "rare" conformer, as indicated by the black oval. Unclear variation is observed over the main conformation of the arginine, possibly due to varying thermal disorder of the sidechain.*

6.3 Preliminary detection of backbone structural variation

The variability in crystals may not only affect local heterogeneity of the protein sidechains, but also the crystal-averaged conformation of the protein backbone; the local alignment of the protein structures in the PanDDA approach provide some indications that this variation is present (see Chapter 3; page 81).

I analysed the phi-psi angle variations in the BAZ2B datasets to investigate the presence of structural variability, with the aim of determining the "crystallographic flexibility" of the protein. The variation in the backbone angles is an alignment-independent measure of structural variation; I therefore calculate a measure of conformational variation as

$\mathrm{rms}\left(\mathrm{rmsd}(\boldsymbol{\phi}), \mathrm{rmsd}(\boldsymbol{\psi})\right)$, where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are the phi and psi angles, respectively, for a particular residue for all datasets – this results in a single value measuring the backbone variation of each residue. The angle variation across all 200 datasets of BAZ2B is shown in Figure 6.5. The background angle variation, caused by uncertainty in the data and refinement, is approximately 3° (25%, 50% and 75% quantiles are 2.09°, 2.37°, and 3.06° respectively).

The variation of residues 1947 and 1948 is significantly higher than the rest of the structure. This local region is ordered in some datasets and disordered in others, resulting in a continuous distribution of refined backbone conformations (Figure 6.6). This variation represents the extreme of the structural variation in the data, and is reflected by density in the adjusted-variation map (Figure 6.7). Smaller structural variations are also present for other residues, chiefly in the disordered extended N-terminal "tail" of BAZ2B (indicated with arrow in Figure 6.5, and shown between the indicated points (1) & (2) in Figure 6.8b).

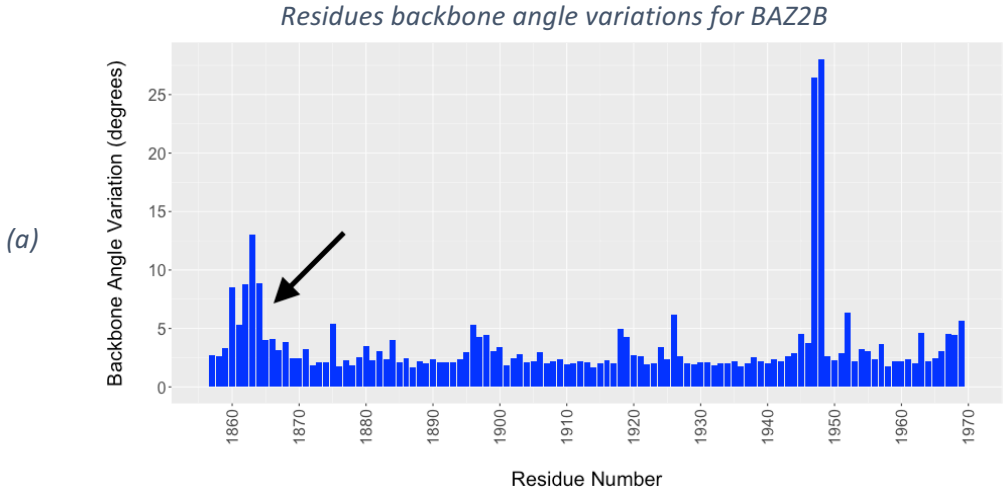
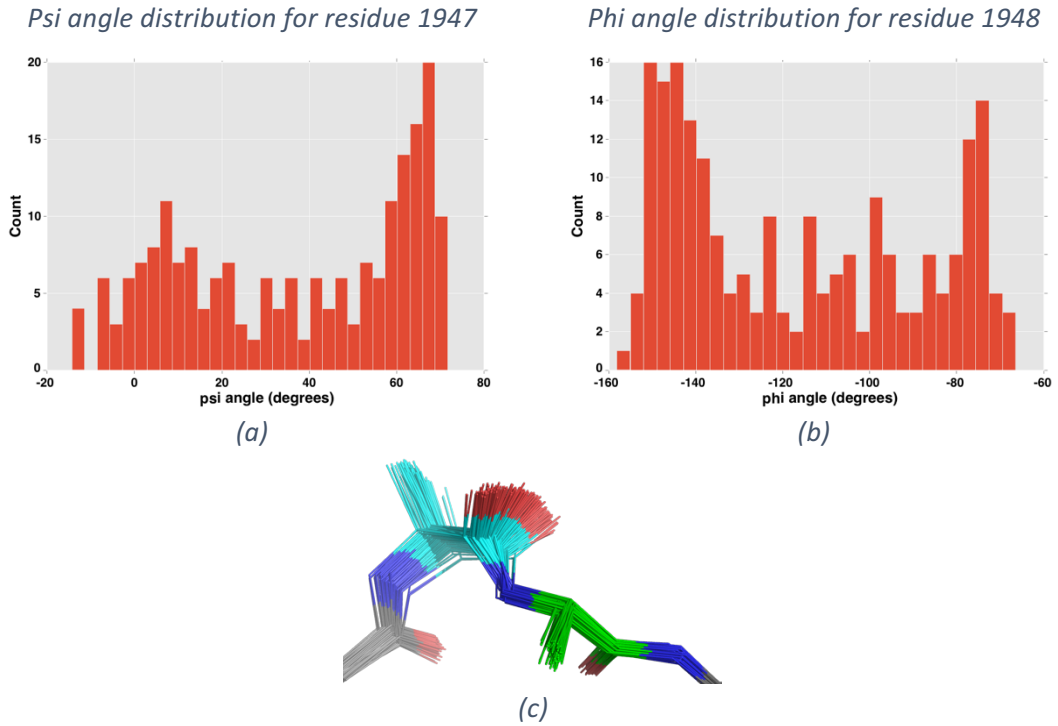


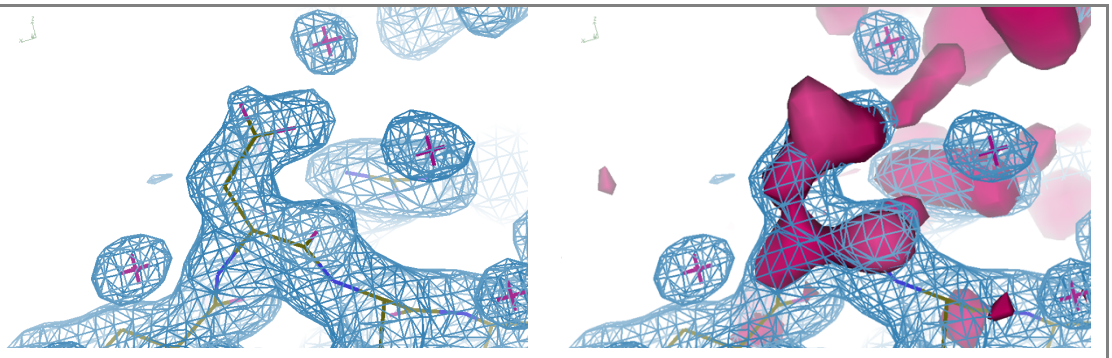
Figure 6.5. ***Variation in the residue backbone angles for BAZ2B.*** The angle variation is measured as the rms(rmsd(phi), rmsd(psi)) for each residue. Residues 1947-48 are the most variable residues, with varying degrees of crystallographic order across the datasets. The N-terminal "tail" of BAZ2B is also structurally variable (indicated with arrow). Smaller variations are also observed in loop regions of the protein.



(a) (b)

(c)

Figure 6.6. ***Distributions of residue backbone angles for residue 1947 and 1948.*** *The distributions of backbone angles for residues (a) 1947 (aspartate) and (b) 1948 (serine) for the 200 BAZ2B datasets span a large range. (c) Backbone of the globally aligned structures of all 200 BAZ2B datasets in the region of residue 1947 (carbons, cyan) and residue 1948 (carbons, green). All other carbons are coloured in grey, and other atoms are coloured by element. The structure of the residue is poorly conserved between datasets; the refined pose of the residue is highly variable.*



(a) (b)

Figure 6.7. ***Some density variation is indicative of variability but not heterogeneity.*** *Electron density maps in the region of a solvent-exposed aspartate of BAZ2B (chain A, residue 1947). (a,b) averaged $2mF_o$-DF_c map from 48 datasets at 1.7Å (blue mesh, 1σ). (b) adjusted variation (s_m) map from PanDDA (hot pink blobs, 0.25σ). The density for the sidechain shows no clear signs of disorder, and the conformation appears well-modelled. However, the variation over the residue indicates that the density varies between datasets.*

6.3.1 Relationship between structural variability and crystallographic order

The large variability of disordered residues (e.g. Figure 6.6 & Figure 6.7) raises the question of whether the structural variation in the models is simply indicative of crystallographic order in the crystal; disordered residues may be more susceptible to the effects of cryo-cooling, which would be reflected in the refined atomic model.

Direct comparison of B-factors between crystallographic datasets is not possible, since the average disorder varies between crystals. I overcome the issues of comparison by normalising all atomic B-factors, B_{atom}, in the model to Z-scores, using the B-factors of the backbone for the whole protein:

$$B_{atom}^{Z} = \frac{B_{atom} - mean(\boldsymbol{B}_{\boldsymbol{backbone}})}{rmsd(\boldsymbol{B}_{\boldsymbol{backbone}})}, \qquad 6.1$$

where $\boldsymbol{B}_{\boldsymbol{backbone}}$ is the vector of B-factors for all backbone atoms in the model. I normalise by the mean and rmsd of the backbone B-factors since we are comparing the backbone atoms only. After normalisation, a mean-$\mathrm{B_z}$ score is calculated for the backbone of each residue, measuring the average thermal motion or static disorder of the residue, relative to rest of the backbone of the protein. These mean-$\mathrm{B_z}$ scores are then averaged for each residue across the datasets, generating a mean(mean-$\mathrm{B_z}$) for each residue. The relationship between the mean(mean-$\mathrm{B_z}$) scores and backbone angle variation, $\mathrm{rms}\big(\mathrm{rmsd}(\boldsymbol{\phi}), \mathrm{rmsd}(\boldsymbol{\psi})\big)$, is shown in Figure 6.8.

There is a clear relationship between the crystallographic order of a residue and the structural variation over datasets, but there are several residues which demonstrate larger-than-expected variation, indicating significant possible effects of cryo-cooling-induced structural variation; these variable residues are predominantly in the loops of the protein, and make crystal contacts to other symmetry-copies of the protein.

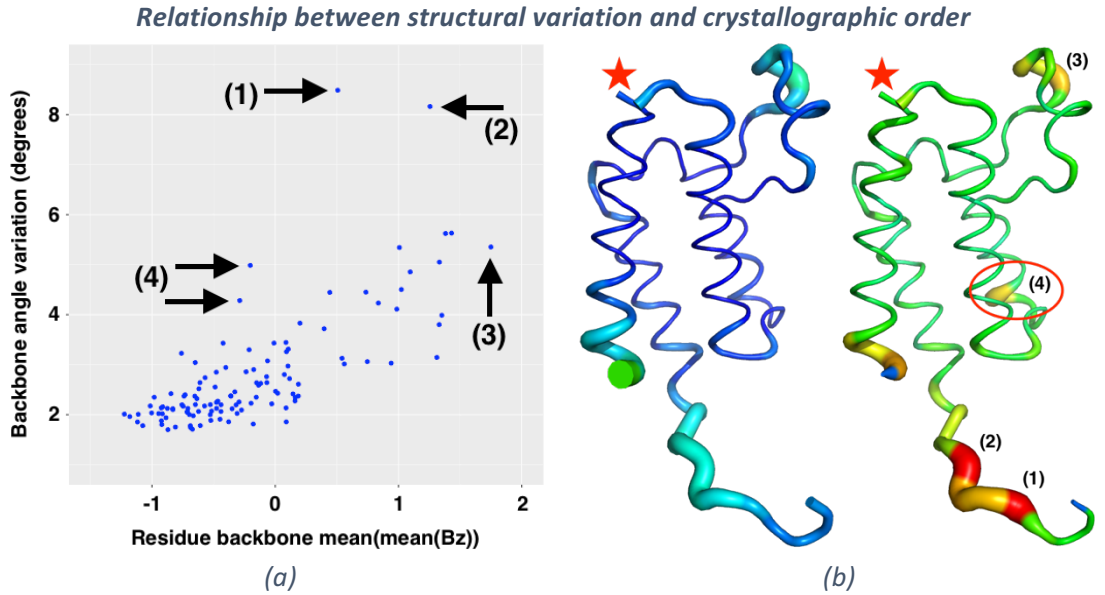


Figure 6.8. ***Relationship between crystallographic order and backbone structural variation.*** *(a) Scatter plot of mean residue backbone B_z scores, a measure of crystallographic order, and rms-rmsd backbone angle variation, both as defined in the text. (b) B-factor putty of mean residue backbone B_z scores (left) and backbone angle variation (right), generated with Pymol. Four regions are marked in both (a) and (b); numbering (1-4) increases along the amino acid sequence. There is a general increase in the structural variation for disordered residues, with some highly variable and highly disordered residues, e.g. (3) (residue 1896, lysine). However, there are also regions which are disordered, e.g. (1) (residue 1860, lysine) & (2) (residue 1863, lysine), but where the structural variation indicates that there are "hinge-points" demarcating a structurally conserved region from a structurally variable region. Finally, there are regions with relatively low backbone B-factors, but which are structurally variable over the datasets, e.g. (4) (residue 1918, leucine, and residue 1919, serine). The highly variable region of residues 1947-1948 are marked with the red star but not shown (mean residue backbone B_z scores for these residues are 1.34 and 0.17 respectively).*

6.4 Chapter Summary and Discussion

The preliminary results presented in section 6.2 confirm my hypothesis that the variation in electron density strength across the datasets can be used to detect conformational heterogeneity, in the case of the BAZ2B dataset. These results, combined with the results in section 6.3, and in continuation of the work in previous chapters, further outline that analysis of multiple datasets can reveal more information about the protein structure than the analysis of a single dataset in isolation.

In future work, I intend to develop methods to utilise the variation in the occupancy of sidechains to better model the conformational heterogeneity of the protein. Identifying the rotamers that are consistent with the density variation across the datasets – using techniques such as principal component analysis (PCA) and independent component analysis (ICA) – will allow multiple-conformer ensemble models to be constructed that better explain the crystallographic data. More robust methods for structure determination will permit much more confident interpretation of crystallographic data, with a reduced reliance on human-focussed density interpretation. Subtle biologically-relevant states may thereby be detectable.

The occupancy of the identified rotamers may further be estimable from such a density variation analysis, thereby reducing the number of parameters in refinement, and reducing the ambiguity of B-factor and occupancy refinement. Better determination of the protein model may further lead to the generation of reliable solvent-omit maps, which could lead to better modelling of the solvent molecules associated with each rotameric state.

Chapter 7

Discussion and Future Work

X-ray crystallography has the power to determine very subtle structural changes in proteins, such as ligand binding. However, current methods are time-consuming, susceptible to human error, and moreover provide generally unsatisfying visual evidence of binding.

7.1 The identification of crystallographic signal

The development and application of the PanDDA method has emphasised that the conventional difference-density-based approach to ligand identification has fundamental flaws for sub-unitary occupancy ligands: difference maps neither provide an objective measure of changed-state signal, nor are sensitive enough to detect partial-occupancy binding, nor present easily-interpretable evidence of binding.

Instead, the PanDDA Z-map presents new possibilities for objective changed-state signal identification and the event map represents a new type of OMIT map that can be used for ligand identification and modelling. Through the PanDDA approach, the sensitivity, objectivity and clarity of ligand identification are all improved.

These results suggest that a new standard approach to ligand-binding and other changed-state studies is in order – that ground-state datasets are routinely collected and used to identify signal in changed-state datasets. Control experiments are a keystone of the scientific method, where a negative control is used to determine

whether signal is truly present in a sample of interest. Yet this standard is not currently applied regularly to ligand identification in crystallography, instead relying on subjective human interpretation of a single dataset (except in the case of isomorphous difference methods, which are not commonly utilised). Since ground-state datasets are almost universally obtainable for ligand-binding studies, there is no experimental reason that changed-state identification in crystallography should not adopt these new methods and become a quantifiable and more statistically robust science.

More rigorous statistical standards for the detection of crystallographic signal will both reduce the frequency of instances in the literature that deal with the unfounded misinterpretation of crystallographic data (e.g. Stanfield et al. 2016), and increase the amount of high-quality structural information available, even from weak crystallographic binders.

The current barrier to the uptake of this approach is the requirement for 30 ground-state datasets to be available. However, this is the required number of datasets for full convergence of the density-variation characterisation, which is only used to suppress noise in the analysis, rather than enhance signal. Instead it is the convergence of the average ground-state map that determines the minimum number of datasets required for a robust analysis.

The convergence of the average map, however, requires further characterisation and corrections to compensate for the variation of the crystallographic order between datasets – the variation in the average/global B-factor – which is overlooked in the current implementation. Methods to compensate for this variation through appropriate

map sharpening and blurring, and the subsequent characterisation of the minimum number of ground-state datasets required, are the subject of future work.

Furthermore, the current method of identifying an appropriate BDC factor – contrast maximisation – is sub-optimal. Determining a more robust method to separate the superposed crystallographic states – such as using the refined model occupancies – will permit the effects of phase bias to be truly extracted from the data (as discussed in section 5.3). Further work is also required on the identification of signal in the Z-map; currently a simple blob-finding method is applied. The success of the current crude implementation attests to the quality of the Z-map, but a more statistically robust method of identifying regions of interest is a key area of further development.

7.2 The modelling of weak crystallographic features

Systematic omissions in modelling such as the failure to model the superposed unbound state are further areas where the conventional modelling approach in crystallography needs to be overhauled. There is once more no obstacle to the adoption of this modelling approach – as observations of the ground state should be available in almost all cases – though flexible and robust methods will be required to make it accessible to the non-expert crystallographic user. The use of multi-state ensembles is crucial for the reliability of refinement and validation of weak features and will likely prove important for improving the overall quality of crystallographic models.

7.3 The utility of weak crystallographic signal

The most common challenge levelled at the weak crystallographic signal detected here is that weakly-binding ligands are uninteresting. However, there are few reasons to believe that low-occupancy in the crystal will always reflect the situation *in vivo*.

Crystallisation conditions rarely reflect biologically-relevant conditions, as binding fragments commonly compete with typical crystallisation components – such as ethylene glycol – that are present at very high concentrations in the crystallisation solution. This competition occurs in dozens of the ligands from the analysis in Chapter 4, and will intrinsically suppress the occupancy of the ligand in the crystal. Furthermore, the binding of the ligand can be impeded by the crystallisation-required conformations of the protein; this is evidenced by the multiple cases where the ligand binds strongly to one copy of protein in the ASU, but weakly to another. Lastly, the potential for weakly-binding fragments to result in possible leads has been demonstrated (Schiebel, Radeva, et al. 2016), though further work is required to show this generally.

7.4 The future of multi-dataset approaches in crystallography

Finally, I submit that a qualitative shift in approaches to generating crystallographic models is now due. It is clear that near-convergence $2mF_o$-DF_c density is necessary *but not sufficient* to completely model a crystal. The PanDDA method addresses one class of experiments, those involving induced local changes, but all problems of uninterpretable density, and indeed some of the R-factor gap, should be addressable by analogous map deconvolution methods.

Multi-dataset experiments are no longer difficult. Instead, what will be key is establishing methods for systematically perturbing poorly ordered or heterogeneous regions, rigorous algorithms for reconstructing and visualising discrete states, and subsequent model validation. Preliminary discussion of the PanDEMIC approach in the last chapter lays out initial evidence that this is possible; from the simultaneous analysis of a set of crystals, all crystals may be better modelled.

Thesis Summary

It is my hope that the contributions of this thesis have extended the limits of ligand identification in X-ray crystallography, enabling a level of signal detection and confidence in modelling that was not previously possible, allowing crystallographers to spend more time modelling, and less time guessing.

Future work will continue to focus on the development of multi-dataset electron density analysis methods for X-ray crystallography. Watch this space.

Appendix A

X-ray Crystallography

A.1 Crystal lattices and crystal symmetry

The stacking of identical blocks in a space-filling regular arrangement forms a crystal lattice. Crystal lattices are defined by a set of basis vectors; all combinations of these basis vectors, with different integer multipliers, defines the infinite set of regularly repeating points of the lattice. There are seven different types of lattice systems, providing 14 different types of *Bravais lattice* in total (Figure A.1). The basis vectors define a set of translational symmetry operations of the crystal; the crystal is invariant under coordinate transformations by these vectors.

Further symmetry can be added to crystals by combining the symmetry operations of the lattice with a *point group symmetry*, which consists of a set of reflectional and rotational symmetry operations. Point groups are defined as a combination of symmetry operations which act to leave a single point invariant (the origin). There are an infinite number of possible point groups, but only 32 that are compatible with the possible crystal lattices (as the lattice itself must be invariant under any combination of a point group symmetry and the lattice translational symmetry); these define the so-called *crystallographic point groups*. The combination of the point groups with the lattice symmetry allows two new types of symmetry operation to be defined: screw axes, a combination of a translation and a rotation; and glide planes, a combination of a translation and a reflection.

Combination of the Bravais lattices, crystallographic point groups, screw axes and glide planes creates the set of 230 possible *space groups*. A space group defines all of the symmetry operations present in a crystal. Protein molecules are chiral, and so the centrosymmetric space groups – those that are invariant upon coordinate inversion – are not possible in protein crystals; the only possible symmetry operations are rotations and screw axes. The chirality restriction leaves 65 possible space groups for protein crystals, the most common of which is $P2_12_12_1$ (No. 19), present in 22.6% of the structures in the PDB (on 2016-08-09).

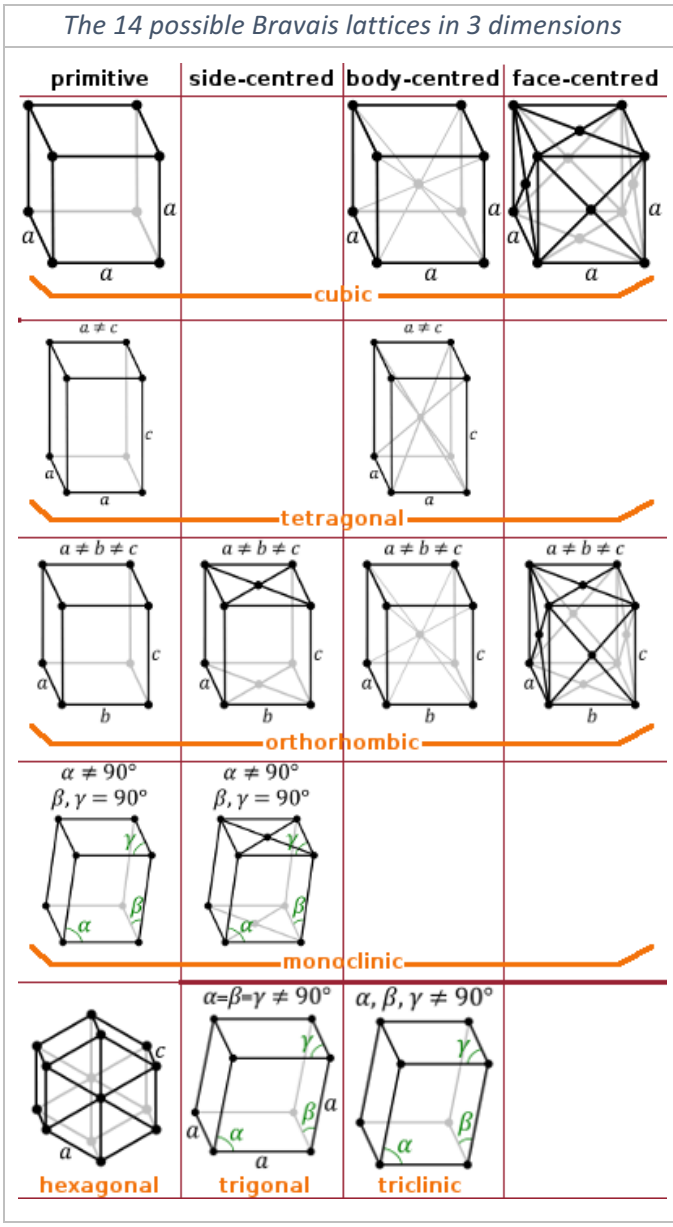


Figure A.1. ***The 14 possible Bravais lattices in three dimensions.*** *Lattices are generated by combinations of three basis lattice vectors. Positioning of a "primitive" unit cell at each lattice point creates a space-filling arrangement. However, it is frequently easier to compose the crystal from larger "conventional" unit cells (as shown), which comprise integer numbers of primitive unit cells.*

A.2 Theory of diffraction

In the general case, the total amplitude of the scattered light measured at the point $\boldsymbol{r}$, at time t, due to electron density $\rho(\boldsymbol{x})$ at points $\boldsymbol{x}$, from an initial wavevector $\boldsymbol{k_i}$, can be written as

$$\mathrm{A}(\boldsymbol{r}) = e^{-i\omega t} f(\theta) \int_{crystal} \rho(\boldsymbol{x}) e^{i(\boldsymbol{k_f}\cdot(\boldsymbol{r}-\boldsymbol{x})+\boldsymbol{k_i}\cdot\boldsymbol{x})}\, d^3\boldsymbol{x}, \qquad \text{A.1}$$

where the integral extends over all of the crystal volume. ω is the angular frequency of the incident and the elastically-scattered wave, $\boldsymbol{k_f}$ is the scattered wavevector, $f(\theta)$ contains terms which vary as a function of the angle between the incident and the scattered wavevectors, as derived from Thomson scattering, and the final phase term encodes the distance travelled by the electron through the scattering event (up to an arbitrary constant common to all electrons in the crystal). Rewriting this, we can obtain

$$\mathrm{A}(\boldsymbol{r}) = e^{i(\boldsymbol{k_f}\cdot\boldsymbol{r}-\omega t)} f(\theta) \int_{crystal} \rho(\boldsymbol{x}) e^{i(-\boldsymbol{k_f}\cdot\boldsymbol{x}+\boldsymbol{k_i}\cdot\boldsymbol{x})}\, d^3\boldsymbol{x}. \qquad \text{A.2}$$

We now define the wavevector $\boldsymbol{q} = \boldsymbol{k_f} - \boldsymbol{k_i}$. The relationship between the wavevectors and the scattering angle is shown in Figure A.2.

Relationship between incident and scattered wavevectors

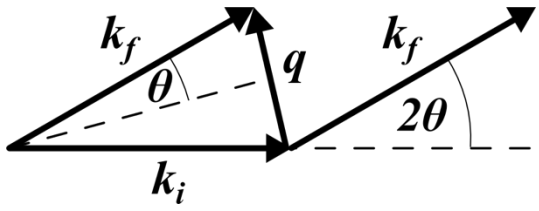


Figure A.2. ***The relationship between the incident, k_i, and scattered, k_f, wavevectors, and the scattering angle, 2θ.*** *In elastic scattering, the magnitude of the scattered wavevector is conserved. The scattering angle is defined as 2θ, rather than θ, for convenience in future calculations. The wavevector* ***q*** *is defined as the difference between the scattered and incident wavevectors.*

In the Fraunhofer diffraction (far-field) limit, the term $\boldsymbol{k_f} \cdot \boldsymbol{r}$ can be approximated as kR where k is the magnitude of the wavevectors of the incident and scattered photons and R is the distance from the crystal to the point of observation $\boldsymbol{r}$. This simplifies equation A.2 to

$$\mathrm{A}(\boldsymbol{q}) = e^{i(kR-\omega t)} f(\theta) \int_{crystal} \rho(\boldsymbol{x}) e^{-i(\boldsymbol{q}\cdot\boldsymbol{x})}\, d^3\boldsymbol{x}, \qquad \text{A.3}$$

now written as a function of $\boldsymbol{q}$**.** We can use the fact that the crystal is periodic, such that

$$\rho(\boldsymbol{x}) = \rho(\boldsymbol{R_i} + \boldsymbol{x}') = \rho(\boldsymbol{x}'), \qquad \text{A.4}$$

where $\boldsymbol{R_i}$ is a lattice vector, composed of linear combinations of the lattice basis vectors, and $\boldsymbol{x}'$ is a vector in the unit cell; this allows equation A.3 to be rewritten as

$$\mathrm{A}(\boldsymbol{q}) = e^{i(kR-\omega t)} f(\theta) \sum_i e^{-i(\boldsymbol{q}\cdot\boldsymbol{R_i})} \int_{unit\ cell} \rho(\boldsymbol{x}') e^{-i(\boldsymbol{q}\cdot\boldsymbol{x}')}\, d^3\boldsymbol{x}', \qquad \text{A.5}$$

where the the sum over i is over all of the unit cells in the crystal. For most vectors $\boldsymbol{q}$, the sum is over a random collection of phases; for a large number of atoms, this sum will be zero. The exception is the case where $\boldsymbol{q} \cdot \boldsymbol{R_i} = 2n\pi$. This is the case when $\boldsymbol{q}$ is a combination of the reciprocal set of vectors to the base vectors of the crystal lattice. For base vectors of the lattice, $\boldsymbol{a_j}$, the reciprocal set of vectors, $\boldsymbol{a_j}'$**,** are defined by

$$\boldsymbol{a_j} \cdot \boldsymbol{a_j'} = 1 \quad \text{and} \quad \boldsymbol{a_j} \cdot \boldsymbol{a_k'} = 0, \text{ for j≠k.} \qquad \text{A.6}$$

The set of reciprocal base vectors, $\boldsymbol{a_j}'$, define the *reciprocal lattice.* For the crystal lattice points, $\boldsymbol{R_i} = \mu\boldsymbol{a_1} + \gamma\boldsymbol{a_2} + \upsilon\boldsymbol{a_3}$, for $\boldsymbol{q} = 2\pi(h\boldsymbol{a_1'} + k\boldsymbol{a_2'} + l\boldsymbol{a_3'})$, where h, k, l, μ, γ and υ are all integers, we have $\boldsymbol{q} \cdot \boldsymbol{R_i} = 2\pi(h\mu + k\gamma + l\upsilon) = 2n\pi$, as required. From Figure A.2,

$$|\boldsymbol{q}| = 2\left|\boldsymbol{k_f}\right|\sin(\theta) = 2|\boldsymbol{k_i}|\sin(\theta) = \frac{4\pi}{\lambda}\sin(\theta), \qquad \text{A.7}$$

and so, identifying the distance, $d = 2\pi/|\boldsymbol{q}|$, we arrive at a condition for diffraction spots of

$$\lambda = 2dsin(\theta), \tag{A.8}$$

which is commonly known as *Bragg's Law*. Although not shown explicitly here, non-negligible diffraction amplitudes appear only when the difference in the distance travelled by photons scattering from electrons at positions separated by a lattice vector is equal to multiples of one wavelength. When this condition is fulfilled, the scattered photons are *in-phase* and lead to a non-zero amplitude when their contributions are summed – constructive interference. At all other points, photons interfere randomly, and for many photons this leads to a zero sum – destructive interference.

As shown in equation A.5, the measured diffraction amplitudes are related to the electron density of the unit cell by a Fourier transform. Using the measured diffraction amplitudes, we can apply a reverse Fourier transform and so regenerate the electron density in the crystal. Refactoring equation A.5, we define the structure factor

$$\mathrm{F}(\boldsymbol{h}) = \int_{unit\ cell} \rho(\boldsymbol{x}') e^{-2\pi i(\boldsymbol{h}\cdot\boldsymbol{x}')}\, d^3\boldsymbol{x}', \tag{A.9}$$

which is derived from the complex diffraction amplitude, $\mathrm{A}(\boldsymbol{h})$, and where

$$\boldsymbol{h} = h\boldsymbol{a}_1' + k\boldsymbol{a}_2' + l\boldsymbol{a}_3' \tag{A.10}$$

is a combination of the reciprocal lattice vectors defined in equation A.6. Writing $\rho(\boldsymbol{x}')$ as a Fourier series in F, where

$$\rho(\boldsymbol{x}') = \sum_{\boldsymbol{h}'} c(\boldsymbol{h}')\mathrm{F}(\boldsymbol{h}') e^{2\pi i(\boldsymbol{h}'\cdot\boldsymbol{x}')}, \tag{A.11}$$

and substituting this into equation A.9, we obtain

$$\mathrm{F}(\boldsymbol{h}) = \int_{unit\ cell} \sum_{\boldsymbol{h}'} c(\boldsymbol{h}')\mathrm{F}(\boldsymbol{h}') e^{2\pi i(\boldsymbol{h}'\cdot\boldsymbol{x}')}\, e^{-2\pi i(\boldsymbol{h}\cdot\boldsymbol{x}')}\, d^3\boldsymbol{x}'. \tag{A.12}$$

This can be rewritten as

$$\mathrm{F}(\boldsymbol{h}) = \sum_{\boldsymbol{h}'} c(\boldsymbol{h}')\mathrm{F}(\boldsymbol{h}') \int_{unit\ cell} e^{2\pi i(\boldsymbol{h}'-\boldsymbol{h})\cdot\boldsymbol{x}'}\, d^3\boldsymbol{x}'. \tag{A.13}$$

The integral over the exponential is only non-zero when $\boldsymbol{h}' = \boldsymbol{h}$, so eq. A.13 becomes

$$\mathrm{F}(\boldsymbol{h}) = c(\boldsymbol{h})\mathrm{F}(\boldsymbol{h}) \int_{unit\ cell} d^3\boldsymbol{x}' = c(\boldsymbol{h})\mathrm{F}(\boldsymbol{h})V_{unit\ cell}, \quad \text{A.14}$$

where $V_{unit\ cell}$ is the volume of the real-space unit cell. Identifying $c(\boldsymbol{h}) = 1/V_{unit\ cell}$, and substituting into eq. A.11, we arrive at

$$\rho(\boldsymbol{x}) = \frac{1}{V_{unit\ cell}} \sum_{\boldsymbol{h}} \mathrm{F}(\boldsymbol{h}) e^{2\pi i(\boldsymbol{h}\cdot\boldsymbol{x})}, \quad \text{A.15}$$

which allows us to calculate the electron density for the unit cell from the diffraction data. From this electron density, we can derive an atomic model of the molecules in the crystal, and so determine their structure.

A.3 Phase bias and phase probabilities

The following description of the process of calculating minimally-biased crystallographic structure factors is based in part on a lecture series by Randy Read (available at http://www-structmed.cimr.cam.ac.uk/course.html) and Read (2006). Several images by Kevin Cowtan are also used.

The problem of phase bias is illustrated in Figure A.3. Phase bias is a fundamental property of model-phased maps, as shown in Figure A.4.

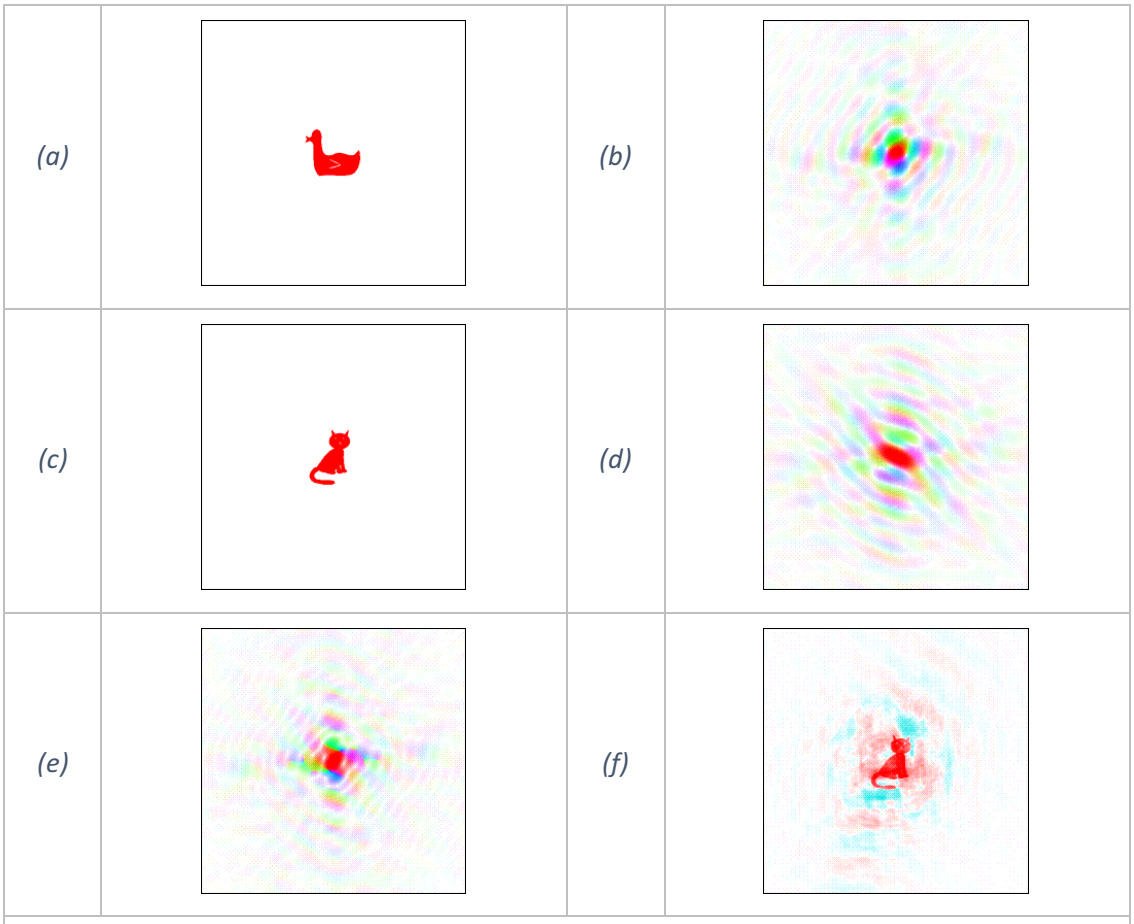


Figure A.3. ***The phases used in image reconstruction dominate the generated image.*** *(a) An image of a duck. (b) The Fourier transform of the duck image in (a). The pixel intensity reflects the amplitude at a point, and the colour reflects the phase. (c,d) The same as (a,b) except for the image of a cat. (e) The combination of the amplitudes (pixel strengths) from (b) and the phases (colours) from (d). (f) The inverse Fourier transform of (e). The phases dominate the reconstruction, and the image looks mostly like the object that was used to generate the phases, the cat. (Images from http://www.ysbl.york.ac.uk/~cowtan/fourier/magic.html).*

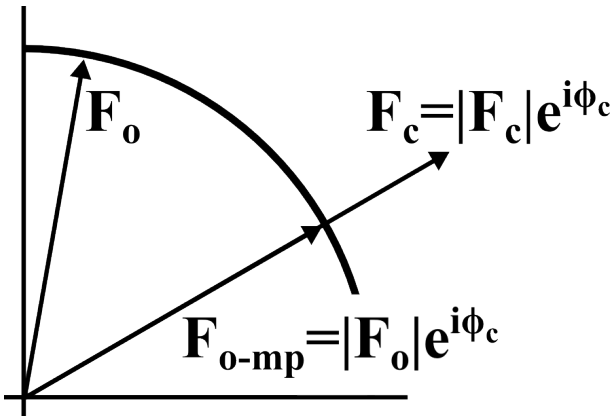


Figure A.4. ***Model-phased maps are biased towards the source of the phases.*** *The true experimental structure factor (F_o), model structure factor (F_c) and the model-phased experimental structure factor ($F_{o\text{-}mp}$) are shown. Once the amplitude of the experimental structure factor, $|F_o|$, is known, the model-phased structure factor is restricted to lie on a circle (as shown). The model-phased structure factor contains information from both the experimental amplitudes and the model phases; the obtained density will therefore contain information from both sources. However, the differences between the observed and calculated amplitudes are likely to be small, relative to the large differences between structure factors that can be caused by a small change in the model phase (ϕ_c); as shown here, the model-phased structure factor is closer to the calculated structure factor than the experimental structure factor. This shows how incorrect phases can dominate the density reconstruction, and the phased data can be biased towards the model, even if it is not truly present in the crystal. (Image based on www-structmed.cimr.cam.ac.uk/Course/Fourier/Fourier.html).*

A.3.1 Parseval's theorem and the figure-of merit weighting

Parseval's theorem states that

$$\int_{unit\ cell} \rho(\boldsymbol{x})^2 d^3\boldsymbol{x} = \frac{1}{V_{unit\ cell}} \sum_{\boldsymbol{h}} |\mathrm{F}(\boldsymbol{h})|^2, \qquad \text{A.16}$$

which arises from taking the integral over the unit cell of the modulus-squared of equation A.15. The rms (root-mean-squared) of the density in the unit cell is therefore directly proportional to the rms of the structure amplitudes, as

$$rms(\rho) = \sqrt{\int_{unit\ cell} \frac{\rho(\boldsymbol{x})^2 d^3\boldsymbol{x}}{V_{unit\ cell}}} = \frac{1}{V_{unit\ cell}} \sqrt{\sum_{\boldsymbol{h}} |\mathrm{F}(\boldsymbol{h})|^2} \propto rms(|\mathrm{F}(\boldsymbol{h})|). \qquad \text{A.17}$$

This also applies to differences between series (Fourier series are linear), such that

$$rms(\rho_1 - \rho_2) \propto rms(|\mathrm{F}_1(\boldsymbol{h}) - \mathrm{F}_2(\boldsymbol{h})|), \qquad \text{A.18}$$

and so the rms error in the density is proportional to the error in the structure factors. To minimise the error in the density, it is necessary to minimise the rms errors in the structure factors. For the experimental amplitudes, we do this by taking a probability-weighted average over all possible phases, ϕ, for a particular amplitude, resulting in a *figure of merit weighting*, $m = \langle \cos(\phi_{model} - \phi) \rangle$, for each experimental amplitude (Figure A.5).

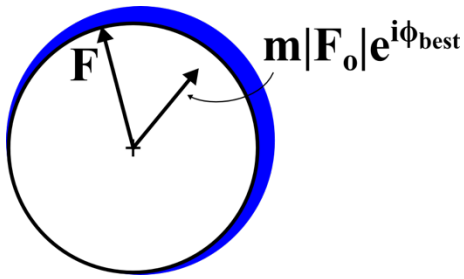


Figure A.5. ***Probability-weighted integration over all possible phases results in a "figure of merit" weighting for the observed diffraction amplitudes.*** *All possible structure factors for a given amplitude ($|F_o|$) are described by a circle (black circle). One of these possible structure factors, F, is shown. Integrating over all possible phases (with weightings corresponding to the thickness of the blue line) results in a best estimate for the experimental structure factor, with weighting m on the amplitude $|F_o|$ and a "best" phase, ϕ_{best}. For a model-phased map, this is the model phase. The figure of merit, m=$|\int P(\phi)\, exp(i\phi)\, d\phi\,|$, is one when the phase is certain ($P(\phi) = \delta(\phi - \bar{\phi})$), and zero when the phase is completely unknown ($P(\phi) = k$). The figure of merit represents our confidence in the quality of the phases used for map generation. (Image adapted from Randy Read at www-structmed.cimr.cam.ac.uk/Course/Fourier/Fourier.html).*

A.3.2 Structure factor probability distributions

The Wilson distribution

The Wilson distribution (Wilson 1949) describes the probability distribution of diffraction amplitudes for a certain wavevector, $\boldsymbol{q} = 2\pi\boldsymbol{h}$, for a crystal where we know the atomic content, but not the location of the atoms. When the atom locations are unknown, the phases can take any value. Summing over all atoms (as point particles) in the unit cell, we obtain the total structure factor

$$\mathrm{F}(\boldsymbol{h}) = \sum_{atoms} f_j e^{2\pi i(\boldsymbol{h}\cdot\boldsymbol{x_j})} = A + iB, \qquad \text{A.19}$$

where the scattering constant for atom j is f_j, located at position x_j, for *N* atoms. Splitting the structure factor into real and imaginary parts (A and B, respectively), the variance of each is calculated by integrating over all possible phases, θ, where each phase is equally likely.

$$\sigma^2(A) = \sum_{atoms} f_j^2 \int_0^{2\pi} \frac{1}{2\pi}\cos^2(\theta)d\theta = \frac{1}{2}\sum_{atoms} f_j^2 = \frac{1}{2}\Sigma_N, \qquad \text{A.20}$$

and the result is the same for B. The expected value of each is zero. The probability distribution for an observed structure factor, under the central limit theorem, is therefore

$$\mathrm{p}(\mathbf{F}) = \mathrm{p}(A,B) = \mathrm{p}(A)\mathrm{p}(B) = \frac{1}{\sqrt{\pi\Sigma_\mathrm{N}}} e^{-\frac{A^2+B^2}{\Sigma_\mathrm{N}}} = \frac{1}{\pi\Sigma_\mathrm{N}} e^{-\frac{|\mathbf{F}|^2}{\Sigma_\mathrm{N}}}. \qquad \text{A.21}$$

This distribution is derived for a specific structure factor, not for a specific amplitude. Furthermore, it covers only acentric reflections (centric reflections are those for which there is a symmetry element of the spacegroup which maps the miller index to the negative of itself). For centric reflections, the structure factor is required to lie along a 1D line (as the structure factor must be invariant under symmetry).

The Sim distribution

The Wilson distribution is derived where the location of none of the atoms is known. The Sim distribution (G. A. Sim 1960; G A Sim 1960) accounts for the known location of a set of atoms, *P*, and the unknown location of a set of atoms, *Q*, contributing structure factors $\mathbf{F_P}$ and $\mathbf{F_Q}$ respectively. The effect of this is to move the centre of the distribution in equation A.21 to $\mathbf{F_P}$, the structure factor of the known set of atoms. This leads to a probability distribution for the combined structure factor

$$\mathrm{p}(\mathbf{F}) = \frac{1}{\pi \Sigma_Q} e^{-\frac{|\mathbf{F}-\mathbf{F_P}|^2}{\Sigma_Q}}, \qquad \text{A.22}$$

where Σ_Q is the sum over the set of unmodelled atoms, *Q*, rather than all atoms as previously.

Structure factors for atoms with coordinate errors

Next we consider the effects of errors in the atomic coordinates in the model we are using to generate the phases. The effect of coordinate uncertainty is to introduce phase uncertainty similar to that in section A.3.1. This is shown schematically in Figure A.6. The value d_j is calculated by integrating over all probability-weighted atomic positions, such that

$$d_j = \int \mathrm{p}(\Delta \boldsymbol{x_j}) e^{2\pi i(\boldsymbol{h}\cdot\Delta \boldsymbol{x_j})} d^3\Delta \boldsymbol{x_j}, \qquad \text{A.23}$$

where $\mathrm{p}(\Delta \boldsymbol{x_j})$ is the probability for the atom to be at position $\boldsymbol{x_j} + \Delta \boldsymbol{x_j}$, the exponential term is the phase difference introduced through the displacement $\Delta \boldsymbol{x_j}$, and the integral is over all possible displacements, $\Delta \boldsymbol{x_j}$.

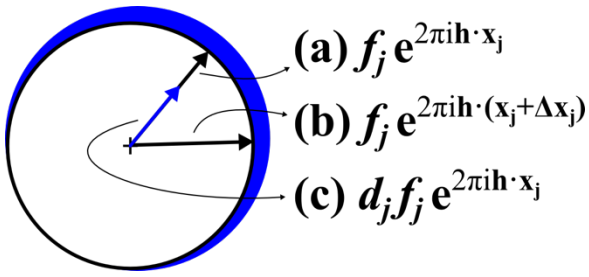


Figure A.6. ***Estimation of structure factor for an atom with coordinate uncertainty.*** *(a) The structure factor for atom j at position x_j, with atomic scattering factor f_j. (b) The structure factor for atom j at the shifted position $x_j + \Delta x_j$. (c) Probability-weighted integration over all possible positions (weightings shown as thickness of external blue band) results in a weighting d_j on the original structure factor from (a) (resulting structure factor shown as the blue arrow). For symmetrical phase probability distributions (such as a Gaussian coordinate error function) the value d_j is a real number (as shown). The method for calculating d_j in the general case is shown in equation A.23. (Image adapted from www-structmed.cimr.cam.ac.uk/Course/Fourier/Fourier.html).*

When the error in each atom's coordinates is accounted for as described above, the overall result is a total weighting to the model structure factors, D (Luzzati 1952). This weights the model structure factors (i.e. $\mathrm{D}\mathbf{F_c}$) equivalently to the weighting of the observed data when calculating the figure of merit (i.e. $m\mathbf{F_o}$).

The resulting probability distribution for the structure factors, taking atomic coordinate error into account, is similar to the Sim distribution (equation A.22), but centred on $\mathrm{D}\mathbf{F_c}$. The application of the weighting factor D to the model structure factor means that, effectively, $(1-\mathrm{D})f_j$ of the scattering factor of each atom is now unaccounted for. This manifests as an adjusted variance term, $\sigma_\Delta^2 = \Sigma_\mathrm{N} - \mathrm{D}^2\Sigma_\mathrm{P}$, which accounts for any unmodelled atoms, or partially accounted-for atoms. Σ_N and Σ_P are the sum-of-squares of all atomic scattering factors and the atomic scattering factors for atoms in the model, respectively. The resulting probability distribution for the structure factor is

$$\mathrm{p}(\mathbf{F};\mathbf{F_c}) = \frac{1}{\pi\varepsilon\sigma_\Delta^2} e^{-\frac{|\mathbf{F}-\mathrm{D}\mathbf{F_c}|^2}{\varepsilon\sigma_\Delta^2}}, \qquad \text{A.24}$$

where ε is an "expected intensity" factor, to account for symmetry-related atoms in the model, that do not contribute independently to the structure factor (as is assumed otherwise).

If the model is complete ($\Sigma_\mathrm{N} = \Sigma_\mathrm{P}, \Sigma_\mathrm{Q} = 0$), the variance term simplifies to

$$\sigma_\Delta^2 = (1-\mathrm{D}^2)\Sigma_\mathrm{N}, \qquad \text{A.25}$$

and the distribution in equation A.24 becomes the Luzzati distribution (Luzzati 1952).

The Srinivasan distribution

The Sim distribution describes an incomplete model with perfect coordinates, and the Luzzati distribution describes a complete model with imperfect coordinates. The

Srinivasan distribution (Srinivasan & Ramachandran 1966) combines the two models through the use of normalised structure factors

$$\mathbf{E} = \frac{\mathbf{F}}{\sqrt{\varepsilon \Sigma_{\mathrm{N}}}} \text{ and } \mathbf{E_c} = \frac{\mathbf{F_c}}{\sqrt{\varepsilon \Sigma_{\mathrm{P}}}}. \quad \text{A.26}$$

Using these normalised structure factors, and

$$\sigma_A = \mathrm{D}\sqrt{\frac{\Sigma_{\mathrm{P}}}{\Sigma_{\mathrm{N}}}}, \quad \text{A.27}$$

(Srinivasan 1966) the probability distribution can be re-written as

$$\mathrm{p}(\mathbf{E}; \mathbf{E_c}) = \frac{1}{\pi(1-\sigma_A^2)} e^{-\frac{|\mathbf{E}-\sigma_A \mathbf{E_c}|^2}{1-\sigma_A^2}}. \quad \text{A.28}$$

This is now the probability distribution for $\mathbf{E}$, given $\mathbf{E_c}$, but before we have measured the amplitude of $\mathbf{E}$.

Probability distribution for known amplitudes

Once the amplitude of the structure factor $\mathbf{F}$ (and therefore $\mathbf{E}$) is known, only the structure factors with that amplitude have non-zero probabilities, and so the distribution of probabilities is reduced to the curve that cuts through the Gaussian probability distribution in the complex plane (Figure A.7). Integration of the re-normalised probability distribution over this line gives an estimate for the D-factor.

A.3.3 Composite maps: Minimisation of phase bias

In previous sections, we have described the derivations of weightings for different structure factors, so as to decrease phase bias. Here, we look at the different combinations of these weighted structure factors, for the purpose of further reducing phase bias. We broadly follow the derivation from Read 1986.

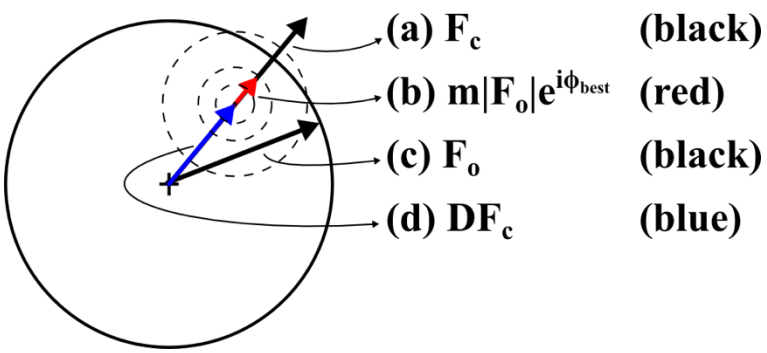


Figure A.7. ***Structure factor probability distribution for known amplitude.*** *(a) The calculated structure factor for the model. (b) The figure-of-merit-weighted estimation of the observed structure factor. (c) A possible structure factor with the observed amplitude, $|F_o|$. (d) D-weighted calculated structure factor for the model. When the amplitude of the observed structure factor is known, the probability distribution for the true F_o, which is centred on DF_c, collapses to where the probability distribution intersects with the circle of the measured amplitude (where the solid circle intersects with the dashed rings representing the probability distribution). (Image adapted from Randy Read at www-structmed.cimr.cam.ac.uk/Course/Fourier/Fourier.html).*

The true structure factor, $\mathbf{F_o}$, can be written in terms of the approximated structure factor derived from the partial model, $\mathrm{D}\mathbf{F_c}$, and a difference term, such that

$$\mathbf{F_o} = \mathrm{D}\mathbf{F_c} + \boldsymbol{\Delta}\mathbf{F}. \tag{A.29}$$

The term $\mathrm{D}\mathbf{F_c}$ has been used since this minimises the rms difference to $\mathbf{F_o}$. Defining the angle between $\mathbf{F_o}$ and $\mathrm{D}\mathbf{F_c}$ as $\Delta\alpha$, we can use the cosine rule to write

$$|\boldsymbol{\Delta}\mathbf{F}|^2 = |\mathbf{F_o}|^2 + \mathrm{D}^2|\mathbf{F_c}|^2 - 2\mathrm{D}|\mathbf{F_o}||\mathbf{F_c}|\cos(\Delta\alpha). \tag{A.30}$$

Taking the expectation of this equation, we obtain

$$\langle|\boldsymbol{\Delta}\mathbf{F}|^2\rangle = |\mathbf{F_o}|^2 + \mathrm{D}^2|\mathbf{F_c}|^2 - 2m\mathrm{D}|\mathbf{F_o}||\mathbf{F_c}|, \tag{A.31}$$

where we have used the $\langle\cos(\Delta\alpha)\rangle = m$, the figure of merit. Inserting the identity

$$|\mathbf{F_o}|^2 = \mathbf{F_o}\mathbf{F_o^*} = \mathbf{F_o}(\mathrm{D}\mathbf{F_c^*} + \boldsymbol{\Delta}\mathbf{F}^*), \tag{A.32}$$

into equation A.31, and multiplying by $e^{\mathrm{i}\phi_c}$, the phase from the model, we obtain

$$\langle|\boldsymbol{\Delta F}|^{\mathbf{2}}\rangle e^{\mathrm{i}\phi_c} = \mathbf{F_o}(\mathrm{D}\mathbf{F_c^*} + \boldsymbol{\Delta F}^*)e^{\mathrm{i}\phi_c} + \mathrm{D}^2|\mathbf{F_c}|^{\mathbf{2}}e^{\mathrm{i}\phi_c} - 2m\mathrm{D}|\mathbf{F_o}||\mathbf{F_c}|e^{\mathrm{i}\phi_c}. \quad \text{A.33}$$

Dividing through by $\mathrm{D}|\mathbf{F_c}|$, and using the identities

$$|\mathbf{F_c}|e^{\mathrm{i}\phi_c} = \mathbf{F_c}, \quad \text{and} \quad \mathbf{F_c^*}e^{\mathrm{i}\phi_c} = |\mathbf{F_c}|, \quad \text{A.34}$$

equation A.33 becomes

$$\frac{\langle|\boldsymbol{\Delta F}|^{\mathbf{2}}\rangle}{\mathrm{D}\mathbf{F_c^*}} = \mathbf{F_o} + \mathbf{F_o}\left(\frac{\boldsymbol{\Delta F}^*}{\mathrm{D}\mathbf{F_c^*}}\right) + \mathrm{D}\mathbf{F_c} - 2m|\mathbf{F_o}|e^{\mathrm{i}\phi_c}. \quad \text{A.35}$$

Expanding the second instance of $\mathbf{F_o}$ once more, we obtain

$$\frac{\langle|\boldsymbol{\Delta F}|^{\mathbf{2}}\rangle}{\mathrm{D}\mathbf{F_c^*}} = \mathbf{F_o} + \boldsymbol{\Delta F}^*e^{\mathrm{i}\phi_c} + \frac{|\boldsymbol{\Delta F}^*|^{\mathbf{2}}}{\mathrm{D}\mathbf{F_c^*}} + \mathrm{D}\mathbf{F_c} - 2m|\mathbf{F_o}|e^{\mathrm{i}\phi_c}. \quad \text{A.36}$$

We ignore any error terms, $\mathcal{O}(\boldsymbol{\Delta F})$, as these are random and therefore simply contribute noise when Fourier-transformed (Main 1979). Rearranging equation A.36, we arrive at

$$\mathbf{F_o} \approx 2m|\mathbf{F_o}|e^{\mathrm{i}\phi_c} - \mathrm{D}\mathbf{F_c} = (2m|\mathbf{F_o}| - \mathrm{D}|\mathbf{F_c}|)e^{\mathrm{i}\phi_c}. \quad \text{A.37}$$

From this we conclude that our best estimate of the electron density for the crystal is to use the structure factors as in equation A.37; these are minimally biased by the set of model phases, and the multiplier of two on the figure-of-merit-weighted observed amplitude serves to up-weight unmodelled features, which are observed at approximately half strength in an $m|\mathbf{F_o}|e^{\mathrm{i}\phi_c}$ map (Luzzati 1953). An illustrative example is shown in Figure A.8.

The complimentary map to the $2m\mathrm{F_o}$-$\mathrm{DF_c}$ map, which represents our best estimate of the complete crystal density, is the $m\mathrm{F_o}$-$\mathrm{DF_c}$ map. This *difference map* represents our best estimate of the difference between the model and the observed data. This is used extensively in model-building to highlight errors and omissions in the model.

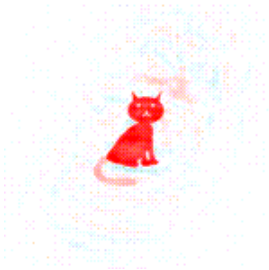
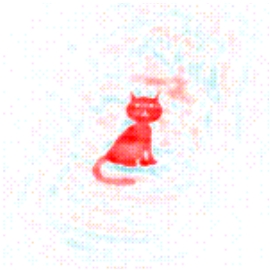
(a)		*(b)*	
(c)		*(d)*	

Figure A.8. ***Weighted combination of available structure factors increases the signal from unmodelled features.*** *(a) An image of a cat. (b) An image of a tailless (Manx) cat. (c) Inverse Fourier transform (FT) of the amplitudes of the FT of (a) combined with the phases of the FT of (b). The tail, which is not present in (b), which was used to generate the phases, is present at about half strength. (d) Inverse FT with the same phases as used for (c), but now with 2x the amplitudes from the FT of (a) minus the amplitudes from the FT of (b). The signal from the tail is increased, although there is an increase in the noise in the image. This is representative of the crystallographic modelling process, where the phases are derived from an incomplete model of the protein, and the weighted combination of the known amplitudes increases the density strength for unmodelled sections. (Images by Kevin Cowtan from http://www.ysbl.york.ac.uk/~cowtan/fourier/coeff.html).*

A.4 Electron density validation metrics derivations

The electron density metrics RSZD and RSZO from Tickle (2012) are re-derived below.

A.4.1 Real-space Z-difference score

The RSZD is a measure of the significance of the difference density in a specified region. The difference density over the whole unit cell is approximately normally distributed, and so we can calculate Z-scores,

$$Z_{\Delta\rho}(x) = \frac{\Delta\rho(x)}{\sigma(\Delta\rho)}, \qquad \text{A.38}$$

at each point, x, where the difference density, $\Delta\rho(x)$, is divided by the standard uncertainty in the difference density, $\sigma(\Delta\rho)$. This provides a statistical measure of how significant the difference density is at a point, relative to the global quality of the model; as the model improves, the metric becomes more sensitive to any difference density present. Since the amount of difference density is directly related to the accuracy of the model, sets of these Z-scores are a natural measure of the local accuracy of the model.

The parameter $\sigma(\Delta\rho)$, the uncertainty in the difference density due to random errors in the observed data and the crystallographic phases, is assumed to be constant over the unit cell. Tickle estimates this value from the bulk-solvent regions of the crystal, by assuming that the bulk solvent model is an adequate description of the completely disordered regions of the crystal, and that any difference density that arises here represents random noise. Plotting the bulk-solvent difference density quantiles against the theoretical normal distribution quantiles, the parameter $\sigma(\Delta\rho)$ is estimated as the slope of the central portion of the quantile-quantile (Q-Q) plot (between the ±1.5 quantiles).

For a group of atoms, difference density is sampled on a grid within a certain radius of the atoms. The finite resolution of the crystallographic map means that grid points must be spaced every $d_{min}/2$ for the sampled values to be independent of each other (Shannon 1949; Tickle 2012), where d_{min} is the resolution limit of the dataset. Samples taken with a smaller sampling distance will contain values that are correlated with each other. However, larger sampling distances may miss peaks in the difference density, which is what we are most interested in.

Tickle compensates for this difficulty by over-sampling the density – at a spacing of $d_{min}/4$, resulting in a 8-fold over-sampling in 3D – and then down-sampling the ordered list of resulting values by the same factor. The methodical resampling of the *ordered* list of difference values ensures a representative sampling of the density distribution, whilst ensuring that the extreme values of difference density are kept. Linear interpolation of the density values is used whenever the down-sampled value falls between measured values. The sampled difference density values are converted to Z-scores as in equation A.38.

So as to consider positive and negative values equally, it is convenient to take the absolute value of the difference density. The absolute values of the Z-scores are assumed to be independent and half-normally distributed. Therefore, the joint probability distribution of the absolute values of the Z-scores, where the ith Z-score is denoted by x_i to prevent confusion with other statistical parameters/variables, is given by

$$\mathrm{P}(x_1, \dots, x_i, \dots x_n) = \prod_{i=1}^{n} 2\psi(x_i) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{x_i^2}{2}\right) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\chi^2}{2}\right), \quad \text{A.39}$$

where the probability distribution for x_i is given by $\psi(x_i)$, the normal distribution (the factor of two compensates for considering the absolute value of x_i). Since Z-scores are normalised values, $\psi \sim \mathcal{N}(0,1)$.

As indicated by the form of the final term in equation A.39, we subsequently apply a chi-squared test to the observed values, to calculate a p-value for observing such a set of Z-scores. The standard chi-squared test cumulative probability, p, is calculated as

$$p\left(\chi^2 \leq \sum_{i=1}^{n} x_i^2\right) = \mathrm{P}\left(\sum_{i=1}^{n} x_i^2 ; \frac{n}{2}\right), \qquad \text{A.40}$$

where $\mathrm{P}(\dots)$ is the lower-regularised gamma function; the corresponding p-value is $1 - p$. This p-value measures the significance of the collection of Z-scores, assuming that *all* values in the sample are significant.

However, by considering the case where half of a residue is correctly modelled and the other half is not, we conclude that in the general case it is not reasonable to assume that all difference density values may be significant. Recognising this, Tickle reformulates the p-value calculated in equation A.40 to consider the significance of the *(n-k)* largest absolute difference density values (from $k = i$ to n), ordered in increasing size. By considering multiple null hypotheses – that the *(n-k)* largest values may be significant for all values of *k* – the relevant p-value corresponds to where the probability, p, is maximised for *k*:

$$p_{max} = \max_{\mathrm{k}} \left(p\left[\chi_k^2 \leq \sum_{i=k}^{n} x_{(i)}^2\right]\right). \qquad \text{A.41}$$

In the general case, the probability, p, is no longer given by the low-regularised gamma function, since a bias is introduced through the methodical sub-selection of the (n-k) largest values. Due to the high dimensionality of this expression, which introduces

difficulties in the empirical calculation of p-values, Tickle suggests the following expression for practical usage:

$$p_{max} \approx \max_{\mathrm{k}}\left(P\left(\sum_{i=k}^{n} x_{(i)}^{2}\,;\frac{n-(k-1)}{2}\right) I\left\{2\Phi\left[x_{(k)}\right]-1; k-1, n-(k-1)\right\}\right). \quad \text{A.42}$$

The term P once more is the lower-regularised gamma function, but is now followed by I, the regularized incomplete beta function. The I term is the multiple testing correction which compensates for the bias introduced through the methodical sub-selection of the difference density values. $\Phi\left[x_{(k)}\right]$ is the cumulative probability distribution of ψ evaluated at $x_{(k)}$.

The corresponding p-value is given by 1-p_{max}. The RSZD is presented as the equivalent standard Z-score for this p-value, as Tickle argues this is more familiar to crystallographers.

A.4.2 Real-space Z-observed score

The real-space z-observed (RSZO) score, measuring precision of the density, is much simpler to derive. It is a signal-to-noise measure calculated by taking the unweighted average density over the atoms of interest, and dividing by the error in the density:

$$\mathrm{RSZO} = \frac{mean(\rho_{obs})}{\sigma(\rho_{obs})} \approx \frac{mean(\rho_{obs})}{\sigma(\Delta\rho)}, \quad \text{A.43}$$

where it is has been assumed that the error in the observed density, $\sigma(\rho_{obs})$, is approximately the same as the error in the difference density, $\sigma(\Delta\rho)$.

Appendix B

Dataset for refitting of crystallographic ligands

The dataset of ligands used for the re-fitting analysis is shown in Table B.1.

Table B.1. ***Dataset used to test current ligand-fitting programs.*** *The protein name is the designation used by the SGC. The model refers to an incremental numbering system used in the SGC. The Residue column contains the chain of the original ligand, the residue number and any alternate conformer ID.*

Protein Name	Model	Residue	Res. (Å)	Atoms
ABCB10A	m001	A_900_	3.25	31
ABCB10A	m002	A_1719_	3.30	31
ABCB10A	m003	A_1719_	2.91	31
ABL2A	m001	A_501_	2.05	37
ABL2A	m001	A_502_	2.05	37
ABL2A	m002	B_1_	1.65	28
ABL2A	m003	A_547_	2.81	33
ABL2A	m003	B_547_	2.81	33
ABL2A	m003	B_548_	2.81	33
ABL2A	m003	B_549_	2.81	33
ABL2A	m003	C_547_	2.81	33
ABL2A	m003	C_548_	2.81	33
ABL2A	m003	C_549_	2.81	33
ACVR1A	m001	D_1_	2.00	25
ACVR1A	m002	H_1_	2.15	26
ACVR1A	m002	I_1_	2.15	26
ACVR1A	m003	E_1_	1.82	31
ACVR1A	m003	E_2_	1.82	31
ACVR1A	m003	E_3_	1.82	31
ACVR1A	m003	E_4_	1.82	31
ACVR1A	m004	D_1_	2.42	23
ACVR1A	m005	A_1000_	2.56	30
ACVR1A	m005	B_1000_	2.56	30
ACVR1A	m005	C_1000_	2.56	30
ACVR1A	m005	D_1000_	2.56	30
ACVR2A	m001	D_1_	1.96	30
ACVR2A	m001	D_2_	1.96	30
ACVR2A	m002	G_1_	1.95	29
ACVR2A	m002	H_1_	1.95	29
ACVR2A	m003	C_1_	2.05	17
ACVR2A	m003	E_1_	2.05	17
AK3A	m002	X_2_	2.05	59
BAZ2BA	m002	B_1_	1.86	13
BAZ2BA	m003	B_1_	2.08	17
BAZ2BA	m004	B_1_	2.06	21
BAZ2BA	m006	B_1_	2.24	18
BAZ2BA	m007	B_1_	2.28	18
BAZ2BA	m009	B_1_	2.54	23
BAZ2BA	m010	B_1_	2.16	24
BAZ2BA	m011	B_1_	2.30	18
BAZ2BA	m012	B_1_	2.29	18

Protein Name	Model	Residue	Res. (Å)	Atoms
DHRS4A	m001	D_602_	1.70	48
DHRS4A	m001	D_603_	1.70	48
DHRS4A	m001	D_604_	1.70	48
DNPEPA	m001	C_4_	2.20	10
DYRK1AA	m001	A_600_	2.40	32
DYRK1AA	m001	B_600_	2.40	32
DYRK1AA	m001	C_600_	2.40	32
DYRK1AA	m001	D_600_	2.40	32
DYRK1AA	m002	A_600_	2.50	32
DYRK1AA	m002	B_600_	2.50	32
DYRK1AA	m003	A_700_	3.15	23
DYRK1AA	m003	B_700_	3.15	23
DYRK1AA	m003	C_700_	3.15	23
DYRK1AA	m004	D_1_	1.40	22
DYRK2A	m002	A_600_	2.55	23
DYRK2A	m003	B_1_	2.28	25
DYRK2A	m004	A_900_	2.80	21
DYRK2A	m005	X_1_	2.30	33
DYRK2A	m006	X_1_	2.70	27
ERAP1A	m001	X_960_	2.70	22
FDPSA	m003	A_901_	2.27	16
FDPSA	m004	A_901_	1.79	17
FESA	m001	D_1_	1.78	35
FESA	m004	B_1_	1.84	42
FLJ13798A	m001	A_501_	2.60	10
FLJ13798A	m001	B_501_	2.60	10
FLJ21802A	m001	A_701_	2.40	12
FLJ21802A	m001	B_701_	2.40	12
GSG2A	m004	B_1_	2.00	24
GSG2A	m006	B_1_	2.00	27
GSG2A	m007	A_800_	1.81	23
GSG2A	m008	C_1_	2.00	20
GSG2A	m011	H_1_	1.90	20
GYG1A	m003	D_1_	2.26	25
GYG1A	m004	D_1_	1.82	25
GYG1A	m006	C_3_	1.80	25
GYG1A	m006	C_4_	1.80	25
GYG1A	m010	D_1_	1.98	25
GYG1A	m010	F_1_	1.98	25
JMJD1BA	m001	D_1_	2.18	10
JMJD2AA	m002	A_500_	1.98	10

BAZ2BA	m013	B_1_	2.66	21
BAZ2BA	m014	B_1_	2.35	20
BAZ2BA	m015	B_1_	2.40	18
BAZ2BA	m016	B_1_	2.34	18
BAZ2BA	m017	B_1_	2.28	28
BAZ2BA	m018	A_4000_	1.88	11
BAZ2BA	m019	A_4000_	2.11	13
BAZ2BA	m021	B_1_	2.00	19
BAZ2BA	m022	C_1_	2.05	24
BAZ2BA	m023	B_1_	1.70	27
BAZ2BA	m024	B_1_	1.80	29
BAZ2BA	m025	C_4000_	1.78	11
BAZ2BA	m026	A_4000_	1.84	11
BAZ2BA	m027	A_4000_	1.80	12
BAZ2BA	m028	H_1_	1.85	17
BAZ2BA	m029	B_1_	1.98	13
BAZ2BA	m030	B_1_	2.30	17
BRD2A	m001	B_1_	1.61	31
BRD2A	m002	D_1_	1.63	24
BRD2A	m003	C_1_	1.59	35
BRD2A	m003	C_2_	1.59	35
BRD2A	m004	B_1_	1.15	32
BRD2A	m005	C_1_	1.10	33
BRD2A	m006	C_1_	1.55	35
BRD2A	m006	C_2_	1.55	35
BRD2A	m008	C_1_	1.10	34
BRD2A	m009	D_1_	1.67	27
BRD3A	m003	B_1_	1.36	31
BRD3A	m004	B_1_	2.06	31
BRD4A	m003	B_1_	1.60	31
BRD4A	m004	C_1_	1.61	19
BRD4A	m006	C_1_	1.68	19
BRD4A	m007	D_1_	1.39	23
BRD4A	m008	B_1_	1.52	24
BRD4A	m009	B_1_	1.60	22
BRD4A	m010	E_1_	1.80	23
BRD4A	m010	E_2_	1.80	23
BRD4A	m010	E_3_	1.80	23
BRD4A	m010	E_4_	1.80	23
BRD4A	m011	B_1_	1.58	37
BRD4A	m016	C_1_	1.63	13
BRD4A	m017	B_1_	1.78	31
BRD4A	m018	B_1_	1.62	20
BRD4A	m019	B_1_	1.59	22
BRD4A	m021	C_1_	1.77	17
BRD4A	m022	C_1_	1.81	20
BRD4A	m024	C_1_	1.66	27
BRD4A	m025	C_1_	1.72	22
BRD4A	m026	C_1_	1.84	22
BRD4A	m027	B_1_	1.60	23
BRD4A	m028	D_1_	1.69	18
BRD4A	m030	C_1_	1.68	24
BRD4A	m031	D_1_	1.73	38
BRD4A	m031	D_2_	1.73	38
BRD4A	m032	D_1_	1.65	37
BRD4A	m032	D_2_	1.65	37
BRD4A	m033	C_1_	1.65	22
BRD4A	m034	B_1_	1.85	26
BRD4A	m035	C_1_	1.64	32
BRD9A	m002	B_1_	1.73	37

JMJD2AA	m002	B_500_	1.98	10
JMJD2AA	m003	A_500_	2.59	10
JMJD2AA	m003	B_500_	2.59	10
JMJD2AA	m007	A_500_	1.93	10
JMJD2AA	m007	B_500_	1.93	10
JMJD2AA	m017	D_4000_	2.16	25
JMJD2AA	m017	D_4001_	2.16	25
JMJD2AA	m018	D_4000_	1.93	25
JMJD2AA	m018	D_4001_	1.93	25
JMJD2AA	m019	D_4000_	1.97	25
JMJD2AA	m019	D_4001_	1.97	25
JMJD2AA	m020	D_4000_	1.83	24
JMJD2AA	m020	D_4001_	1.83	24
JMJD2DA	m001	B_4000_	1.22	12
JMJD2DA	m011	D_4000_	1.24	11
JMJD3A	m001	A_3001_	1.70	14
JMJD3A	m001	B_3001_	1.70	14
JMJD3A	m002	C_1_	2.13	29
LIMK1A	m001	C_1_	1.65	35
LIMK1A	m001	E_1_	1.65	35
LOC148158A	m005	D_1_	2.60	15
LOC148158A	m005	E_1_	2.60	15
LOC148158A	m005	F_1_	2.60	15
LTB4DHA	m002	A_801_	2.20	34
LTB4DHA	m002	A_802_	2.20	34
MAP2K6A	m001	B_1_	2.26	35
MAPK7A	m001	B_1_	2.80	46
MAT2BB	m001	A_501_	2.80	17
MAT2BB	m001	A_502_	2.80	17
MAT2BB	m001	B_501_	2.80	17
MAT2BB	m001	B_502_	2.80	17
MAT2BB	m001	C_501_	2.80	17
MAT2BB	m001	C_502_	2.80	17
MAT2BB	m001	D_501_	2.80	17
MAT2BB	m001	D_502_	2.80	17
MAT2BB	m001	E_501_	2.80	17
MAT2BB	m001	E_502_	2.80	17
MGC45594A	m002	C_1_	1.90	19
MGC45594A	m002	E_1_	1.90	19
MGC45594A	m002	G_1_	1.90	19
MGC45594A	m002	H_1_	1.90	19
MGC45594A	m002	I_1_	1.90	19
MGC45594A	m002	K_1_	1.90	19
MGC45594A	m002	L_1_	1.90	19
MGC45594A	m002	M_1_	1.90	19
MINAB	m001	D_601_	2.57	10
MTRA	m001	B_1_	2.70	32
NUDT1A	m003	A_200_	1.70	16
NUDT1A	m003	B_200_	1.70	16
NUDT1A	m004	A_600_	1.85	18
NUDT1A	m004	B_600_	1.85	18
PHF8A	m002	B_1_	2.55	11
PHIPA	m003	X_1_	1.91	7
PHIPA	m005	B_1_	1.97	13
PHKG2A	m001	E_1_	2.50	29
PHKG2A	m001	F_1_	2.50	29
PHKG2A	m001	G_1_	2.50	29
PHKG2A	m001	H_1_	2.50	29
PIM1A	m002	B_1_	1.80	31
PIM1A	m003	D_1_	2.55	31

BRD9A	m003	B_1_	1.70	19
BRDTA	m002	C_1_	2.20	31
BRDTA	m002	C_2_	2.20	31
CDKL1A	m001	A_500_	2.40	32
CDKL1A	m001	B_500_	2.40	32
CDKL1A	m001	C_500_	2.40	32
CDKL2A	m002	A_500_	1.53	28
CDKL3A	m001	A_350_	2.20	27
CDKL5A	m001	A_1000_	2.00	27
CLK3A	m003	B_1_	1.92	28
CLK3A	m005	B_1_	2.25	22
CLK3A	m006	C_1_	2.09	23
CLK3A	m006	C_2_	2.09	23
CREBBPA	m003	G_1_	1.82	13
CREBBPA	m004	E_1_	1.86	7
CREBBPA	m004	E_2_	1.86	7
CREBBPA	m006	E_1_	1.63	12
CREBBPA	m006	E_2_	1.63	12
CREBBPA	m008	C_1_	1.80	19
CREBBPA	m008	C_2_	1.80	19
CREBBPA	m008	D_1_	1.80	19
CREBBPA	m008	D_2_	1.80	19
CREBBPA	m009	C_1_	1.43	28
CREBBPA	m010	B_1_	1.40	28
CREBBPA	m011	B_1_	1.66	34
CREBBPA	m012	B_1_	1.66	32
CREBBPA	m013	C_1_	1.20	36
CREBBPA	m014	C_1_	1.69	24
CREBBPA	m014	C_2_	1.69	24
CREBBPA	m014	C_3_	1.69	24
CREBBPA	m015	E_1_	1.83	13
CREBBPA	m015	E_2_	1.83	13
CREBBPA	m015	E_3_	1.83	13
CREBBPA	m015	E_4_	1.83	13
CREBBPA	m016	B_1_	1.10	30
DDR1A	m001	A_1000_	1.92	39
DDR1A	m001	B_1000_	1.92	39
DDR1A	m001	C_1_	1.92	39
DDR1A	m002	A_1000_	1.95	43
DDR1A	m002	B_1000_	1.95	43
DDR1A	m003	A_1000_	1.70	37
DDR1A	m003	B_1000_	1.70	37
DDR1A	m004	A_1000_	1.70	40
DHRS4A	m001	D_601_	1.70	48

PIM1A	m005	C_1_	2.45	31
PIM1A	m008	D_1_	1.90	23
PIM1A	m009	C_1_	1.90	35
PIM1A	m012	D_1_	2.10	25
PIM1A	m014	C_1_	2.22	21
PIM1A	m015	C_1_	2.35	18
PIM1A	m016	C_1_	2.20	31
PIM1A	m017	C_1_	2.00	23
PIM1A	m018	C_1_	1.83	24
PRKCL2A	m001	B_2_	2.75	31
RIPK2A	m001	A_1000_	2.75	39
RIPK2A	m001	B_1000_	2.75	39
SRPK2A	m001	A_800_	2.50	30
STK10A	m003	B_1_	2.45	32
STK10A	m004	B_1_	2.41	33
STK10A	m007	A_500_	3.05	35
STK10A	m007	B_500_	3.05	35
STK17BA	m001	B_1_	2.35	26
STK17BA	m002	B_1_	2.29	22
TXNL2A	m003	A_1261_	1.90	20
TXNL2A	m003	B_1261_	1.90	20
TYK2A	m001	C_1_	2.15	18
UTYC	m002	A_1900_	1.81	29
UTYC	m002	B_1900_	1.81	29
VRK1A	m001	E_1_	2.40	25
VRK1A	m001	F_1_	2.40	25
VRK1A	m001	G_1_	2.40	25
VRK1A	m001	H_1_	2.40	25
XX01BMPR1BA	m001	E_1_	2.05	31
XX01BMPR1BA	m001	G_1_	2.05	31
XX01GAKA	m001	K_1_	2.55	27
XX01GAKA	m001	K_2_	2.55	27
XX01MAPK14A	m001	C_1_	1.95	25
XX01MAPK14A	m002	E_1_	2.05	25
XX01MAPK14A	m002	E_2_	2.05	25
XX01MAPK14A	m002	E_3_	2.05	25
XX01MAPK14A	m002	E_4_	2.05	25
XX01MAPK14A	m003	E_1_	2.32	25
XX01MAPK14A	m003	E_2_	2.32	25
XX01MAPK14A	m003	E_3_	2.32	25
XX01MAPK14A	m003	E_4_	2.32	25
XX03GAKA	m001	G_1_	2.16	25
XX03GAKA	m002	D_1_	2.80	27
XX07AURKBA	m001	A_500_	2.75	33

Appendix C

Maximum Likelihood Methods

C.1 Bayes Theorem

Bayes theorem states that the probability for events *A* and *B* to occur, $P(A, B)$, can be written

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A), \quad \text{C.1}$$

where $P(A|B)$ is the probability of *A*, given that *B* has occurred, and vice versa for $P(B|A)$. $P(A)$ and $P(B)$ are the probabilities of *A* and *B*, respectively. Rearranging equation C.1, we can obtain

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad \text{C.2}$$

i.e. we can write the probability of *B* as a function of the probability of *A*, and the probabilities of observing *A*, given *B*, and vice versa. This has significant application in the development of maximum likelihood methods.

C.2 Maximum likelihood methods

Frequently in analyses, the parameters of a statistical model are fitted to a set of observed data. The estimation of parameters in the statistical model is used to draw conclusions about the data; thus it is important to obtain the best estimates of those values. For the probability of a *model* given some observed *data*, from equation C.2 we can write

$$P(model|data) = \frac{P(data|model)P(model)}{P(data)}. \quad \text{C.3}$$

When the data has been collected, $P(data) = 1$, and $P(model)$ is some constant. We therefore have

$$P(model|data) \propto P(data|model). \quad \text{C.4}$$

This enables us to determine the most likely set of model parameters, given a set of observed data. For an example, consider the normal distribution

$$P(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad \text{C.5}$$

This gives the probability to observe a value, x, given parameters μ and σ^2. The probability of observing a set of values $x_1, x_2, x_3, \ldots$, which are independent and identically distributed, is equal to the product of the individual probabilities. This joint probability distribution function (JPDF) is equal to

$$P(x_1, x_2, x_3, \ldots|\mu, \sigma^2) = \prod_i P(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\sum_i \frac{(x-\mu)^2}{2\sigma^2}\right). \quad \text{C.6}$$

From equation C.4, we may re-write this as a probability distribution for one of the model parameters, e.g.

$$\begin{aligned} P(\mu|\sigma^2, x_1, x_2, x_3, \ldots) &\propto P(x_1, x_2, x_3, \ldots|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\sum_i \frac{(x-\mu)^2}{2\sigma^2}\right). \end{aligned} \quad \text{C.7}$$

Therefore, given a set of observed data, $\{x_i\}$, and a value for the other statistical model parameters, i.e. σ, we can estimate a value for μ by maximising the probability in equation C.7 as a function of μ: by evaluating

$$\frac{dP(\mu|\sigma^2, x_1, x_2, x_3, \ldots)}{d\mu} = 0. \quad \text{C.8}$$

Appendix D

Results from re-analysis of the Schiebel datasets

The ligands identified by the original analysis of the Schiebel et al (2016) datasets and the hits subsequently identified by PanDDA are shown in Table D.1.

Table D.1. ***Ligand identification by original analysis and PanDDA analysis.*** *The rows with medium-confidence hits from PanDDA are highlighted in* ***bold****.*

Fragment	PDB	Identified by Schiebel	Identified by PanDDA
4	5OYT	YES	YES
5	5OYU	YES	YES
8	5OYX		YES
14	5OZ3	YES	YES
17	5OZ6	YES	YES
22	**5OZB**		**YES**
29	5OZI		YES
30	5OZJ		YES
31	5OZK	YES	YES
34	**5OZN**	**YES**	**YES**
35	5OZO	YES	YES
39	5OZS	YES	YES
41	5OZU	YES	YES
42	**5OZV**	**YES**	**YES**
48	5P01	YES	
51	5P04	YES	YES
52	5P05	YES	YES
54	5P07	YES	YES
56	5P09	YES	YES
58	5P0B	YES	YES
63	5P0G	YES	YES
66	**5P0J**	**YES**	**YES**
73	5P0Q	YES	YES
75	5P0S	YES	YES
78	5P0V	YES	YES
80	**5P0X**		**YES**
81	5P0Y	YES	YES
106	5P1N		YES
109	5P1Q	YES	YES
110	5P1R		YES
112	5P1T	YES	YES
114	5P1V	YES	YES
122	5P24		YES
124	**5P26**		**YES**
125	5P27	YES	YES

Fragment	PDB	Identified by Schiebel	Identified by PanDDA
206	5P4H	YES	YES
207	**5P4I**	**YES**	**YES**
209	5P4K	YES	YES
211	5P4M	YES	YES
213	5P4O		YES
214	5P4P		YES
216	5P4R	YES	YES
218	**5P4T**	**YES**	**YES**
224	5P4Z	YES	YES
227	5P52	YES	YES
228	5P53		YES
230	5P55		YES
231	5P56	YES	YES
234	**5P59**		**YES**
236	5P5B	YES	YES
240	5P5F	YES	YES
253	5P5S		YES
254	5P5T		YES
255	5P5U	YES	YES
260	5P5Z	YES	YES
261	5P60	YES	YES
266	5P65	YES	YES
267	5P66	YES	YES
268	5P67	YES	YES
272	5P6B	YES	YES
273	5P6C	YES	YES
274	5P6D	YES	YES
277	**5P6G**		**YES**
278	5P6H	YES	YES
279	5P6I		YES
283	5P6M		YES
284	5P6N		YES
285	5P6O	YES	YES
286	5P6P	YES	YES
290	5P6T	YES	YES

127	**5P29**		**YES**
131	5P2D	YES	YES
133	5P2F		YES
138	5P2K		YES
158	5P35	YES	YES
162	5P39	YES	YES
164	5P3B	YES	YES
168	5P3F		YES
171	5P3I	YES	YES
177	5P3O	YES	YES
181	5P3S	YES	YES
188	**5P3Z**		**YES**
189	5P40	YES	YES
191	5P42		YES
196	5P47		YES
198	**5P49**		**YES**
201	5P4C		YES
203	5P4E	YES	YES
205	5P4G	YES	YES

291	5P6U	YES	YES
292	5P6V		YES
299	5P72		YES
301	5P74		YES
305	5P78	YES	YES
306	5P79	YES	YES
311	5P7E	YES	YES
321	5P7O		YES
323	5P7Q	YES	YES
324	5P7R		YES
328	5P7V	YES	YES
329	**5P7W**		**YES**
330	5P7X	YES	YES
333	5P80	YES	YES
337	5P84	YES	YES
338	5P85	YES	YES
348	5P8F		YES
351	5P8I		YES
355	**5P8M**		**YES**

References

Adams, P.D. et al., 2010. PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2), pp.213–221.

Afonine, P. V. et al., 2012. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography*, 68(4), pp.352–367.

Alberts, B. et al., 2014. *Molecular Biology of the Cell* 6th ed., Garland Science.

Anderson, A.C., 2003. The Process of Structure-Based Drug Design. *Chemistry & Biology*, 10(9), pp.787–797. Available at: http://www.cell.com/article/S1074552103001947/fulltext.

Van Den Bedem, H. et al., 2009. Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. *Acta Crystallographica Section D: Biological Crystallography*, 65(10), pp.1107–1117.

Berg, J.M., Tymoczko, J.L. & Stryer, L., 2002. Biochemistry. *W H Freeman*, New York., pp.320–323. Available at: papers2://publication/uuid/7EB6183C-F1A3-4903-A72F-B1A659CECF68.

Berman, H., Henrick, K. & Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. *Nature structural biology*, 10(12), p.980.

Berman, H.M. et al., 2000. The Protein Data Bank. *Nucleic acids research*, 28(1), pp.235–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102472&tool=pmcentrez&rendertype=abstract.

Bhat, T.N., 1989. Correlation between occupancy and temperature factors of solvent molecules in crystal structures of proteins. *Acta Crystallographica Section A*, 45, pp.145–146.

Bhat, T.N. & Cohen, G.H., 1984. OMITMAP: an Electron Density Map Suitable for the Examination of Errors in a Macromolecular Model. *Journal of Applied Crystallography*, 17(pt 4), pp.244–248.

Blundell, T.L., Jhoti, H. & Abell, C., 2002. High-Throughput Crystallography for Lead Discovery in Drug Design. *Nature Reviews Drug Discovery*, 1(1), pp.45–54. Available at: http://www.nature.com/doifinder/10.1038/nrd706.

Bränden, C.-I. & Alwyn Jones, T., 1990. Between objectivity and subjectivity. *Nature*, 343(6260), pp.687–689.

Bricogne, G. et al., 2011. BUSTER version 1.10.0. *Cambridge, United Kingdom: Global Phasing Ltd.*

Brown, E.N. & Ramaswamy, S., 2007. Quality of protein crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 63(9), pp.941–950.

Brunger, A.T., 1992. Free R-Value - a Novel Statistical Quantity for Assessing the Accuracy of Crystal-Structures. *Nature*, 355(6359), pp.472–475. Available at: papers3://publication/uuid/C53A6ADB-4599-44A0-A160-22130E7A2D5F.

Brünger, A.T., Karplus, M. & Petsko, G.A., 1989. Crystallographic refinement by simulated annealing: application to crambin. *Acta Crystallographica Section A*, 45(1), pp.50–61.

Burnley, B.T. et al., 2012. Modelling dynamics in protein crystal structures by ensemble

refinement. *eLife*, 1, p.e00311.

Carolan, C.G. & Lamzin, V.S., 2014. Automated identification of crystallographic ligands using sparse-density representations. *Acta crystallographica. Section D, Biological crystallography*, 70(Pt 7), pp.1844–53. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4089483&tool=pmcentrez&rendertype=abstract.

Carr, R. a E. et al., 2005. Fragment-based lead discovery: Leads by design. *Drug Discovery Today*, 10(14), pp.987–992.

Carr, R. & Jhoti, H., 2002. Structure-based screening of low-affinity compounds. *Drug Discovery Today*, 7(9), pp.522–527.

Chen, V.B. et al., 2010. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1), pp.12–21.

Chilingaryan, Z., Yin, Z. & Oakley, A.J., 2012. Fragment-based screening by protein crystallography: Successes and pitfalls. *International Journal of Molecular Sciences*, 13(10), pp.12857–12879.

Congreve, M., Murray, C.W. & Blundell, T.L., 2005. Keynote review: Structural biology and drug discovery. *Drug Discovery Today*, 10(13), pp.895–907.

Deller, M.C. & Rupp, B., 2015. Models of protein-ligand crystal structures: trust, but verify. *Journal of computer-aided molecular design*, 29(9), pp.817–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25665575.

Emsley, P. et al., 2010. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4), pp.486–501.

Emsley, P. & Cowtan, K., 2004. Coot: Model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, 60(12 I), pp.2126–2132.

Evans, P., 2006. Scaling and assessment of data quality. *Acta Crystallographica Section D: Biological Crystallography*, 62(1), pp.72–82.

Evans, P.R. & Murshudov, G.N., 2013. How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography*, 69(7), pp.1204–1214.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–874.

Fischer, M., Shoichet, B. & Fraser, J., 2015. One Crystal, two Temperatures - Cryocooling Penalties alter Ligand Binding to Transient Protein Sites. *ChemBioChem*, 16, pp.1560–1564. Available at: http://doi.wiley.com/10.1002/cbic.201500196.

Foadi, J. et al., 2013. Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*, 69(8), pp.1617–1632. Available at: http://scripts.iucr.org/cgi-bin/paper?S0907444913012274.

Fraser, J.S. et al., 2011. Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences*, 108(39), pp.16247–16252.

Grosse-Kunstleve, R.W. et al., 2002. The Computational Crystallography Toolbox: Crystallographic algorithms in a reusable software framework. *Journal of Applied Crystallography*, 35(1), pp.126–136.

Halgren, T. a, 1996. Merck Molecular Force Field. *J. Comput. Chem.*, 17(5–6), pp.490–519. Available at: http://doi.wiley.com/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

Hartshorn, M.J. et al., 2005. Fragment-Based Lead Discovery Using X-ray Crystallography. *Journal of Medicinal Chemistry*, 48(2), pp.403–413. Available at: http://pubs.acs.org/doi/abs/10.1021/jm0495778.

Hassell, A.M. et al., 2006. Crystallization of protein-ligand complexes. *Acta Crystallographica Section D: Biological Crystallography*, 63(1), pp.72–79.

Headd, J.J. et al., 2012. Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallographica Section D: Biological Crystallography*, 68(4), pp.381–390.

Holton, J.M. et al., 2014. The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *FEBS Journal*, 281(18), pp.4046–4060. Available at: http://doi.wiley.com/10.1111/febs.12922.

Hughes, J.P. et al., 2011. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6), pp.1239–1249.

Jones, T.A. et al., 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica Section A*, 47(2), pp.110–119.

Kabsch, W., 2010. XDS. *Acta Crystallographica Section D: Biological Crystallography*, 66(2), pp.125–132.

Keedy, D.A. et al., 2014. Crystal cryocooling distorts conformational heterogeneity in a model michaelis complex of DHFR. *Structure*, 22(6), pp.899–910. Available at: http://dx.doi.org/10.1016/j.str.2014.04.016.

Keedy, D.A., Fraser, J.S. & van den Bedem, H., 2015. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. *PLoS Computational Biology*, 11(10), pp.1–22.

Kleywegt, G.J. & Brünger, A.T., 1996. Checking your imagination: Applications of the free R value. *Structure*, 4(8), pp.897–904.

Kozakov, D. et al., 2015. Ligand deconstruction: Why some fragment binding positions are conserved and others are not. *Proceedings of the National Academy of Sciences of the United States of America*, 112(20), pp.E2585-94. Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1501567112\nhttp://www.pnas.org/content/112/20/E2585.

Kuzmanic, A., Pannu, N.S. & Zagrovic, B., 2014. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat Commun*, 5, p.3220. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3926004&tool=pmcentrez&rendertype=abstract.

Lang, P.T. et al., 2010. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Science*, 19(7), pp.1420–1431.

Lang, P.T. et al., 2014. Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc. Natl. Acad. Sci. USA*, 111(1), pp.237–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3890839&tool=pmcentrez&rendertype=abstract.

Langer, G. et al., 2008. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature protocols*, 3(7), pp.1171–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2582149&tool=pmcentrez&rendertype=abstract.

Langer, G.G. et al., 2012. Fragmentation-tree density representation for crystallographic modelling of bound ligands. *Journal of Molecular Biology*, 419(3–4), pp.211–222. Available at: http://dx.doi.org/10.1016/j.jmb.2012.03.012.

Liu, J. & Nussinov, R., 2016. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Computational Biology*, 12(6), pp.3–7.

Ludlow, R.F. et al., 2015. Detection of secondary binding sites in proteins using fragment screening. *Proceedings of the National Academy of Sciences*, 2015, p.201518946. Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1518946112.

Luzzati, V., 1953. Resolution d'un structure cristalline lorsque les positions d'une partie des atoms sont connues: traitement statistique. *Acta Crystallographica*, 6(2), pp.142–152. Available at: http://scripts.iucr.org/cgi-bin/paper?S0365110X53000508\npapers3://publication/doi/10.1107/S0365110X53000508.

Luzzati, V., 1952. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallographica*, 5(6), pp.802–810.

Macarron, R. et al., 2011. Impact of high-throughput screening. *Nature*, 10(March 2011), pp.188–195. Available at: http://dx.doi.org/10.1038/nrd3368.

Main, P., 1979. A theoretical comparison of the Beta, Gamma' and 2Fo-Fc syntheses. *Acta Crystallographica Section A*, 35(5), pp.779–785.

Matthews, B.W., 1968. Solvent content of protein crystals. *Journal of Molecular Biology*, 33(2), pp.491–497. Available at: http://linkinghub.elsevier.com/retrieve/pii/0022283668902052.

McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models, Second Edition*,

McPherson, A. & Gavira, J.A., 2014. Introduction to protein crystallization. *Acta Crystallographica Section F:Structural Biology Communications*, 70(1), pp.2–20.

Mooij, W.T.M. et al., 2006. Automated protein-ligand crystallography for structure-based drug design. *ChemMedChem*, 1(8), pp.827–838.

Moriarty, N.W., Grosse-Kunstleve, R.W. & Adams, P.D., 2009. Electronic ligand builder and optimization workbench (eLBOW): A tool for ligand coordinate and restraint generation. *Acta Crystallographica Section D: Biological Crystallography*, 65(10), pp.1074–1080.

Murray, C.W. & Blundell, T.L., 2010. Structural biology in fragment-based drug design. *Current Opinion in Structural Biology*, 20(4), pp.497–507. Available at: http://dx.doi.org/10.1016/j.sbi.2010.04.003.

Murray, C.W. & Rees, D.C., 2009. The rise of fragment-based drug discovery. *Nature chemistry*, 1(3), pp.187–192.

Murshudov, G.N. et al., 2011. *REFMAC* 5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D Biological Crystallography*, 67(4), pp.355–367. Available at: http://scripts.iucr.org/cgi-bin/paper?S0907444911001314.

Nicholls, R.A., Long, F. & Murshudov, G.N., 2012. Low-resolution refinement tools in REFMAC5. *Acta Crystallographica Section D: Biological Crystallography*, 68(4), pp.404–417.

Nienaber, V.L. et al., 2000. Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nature biotechnology*, 18(10), pp.1105–1108.

Pozharski, E., Weichenberger, C.X. & Rupp, B., 2013. Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 69(2), pp.150–167.

Price II, W.N. et al., 2009. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nature Biotechnology*, 27(1), pp.51–57. Available at: http://www.nature.com/nbt/journal/v27/n1/full/nbt.1514.html\nhttp://www.nature.com/nbt/journal/v27/n1/pdf/nbt.1514.pdf.

Ramachandran, G.N., Ramakrishnan, C. & Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7(1), pp.95–99. Available at: http://dx.doi.org/10.1016/S0022-2836(63)80023-6.

Read, R.J., 2006. 15.2. Model phases: probabilities, bias and maps. In M. G. Rossmann & E. Arnold, eds. *Crystallography of biological macromolecules*. International Tables for Crystallography Volume F, pp. 325–331.

Read, R.J., 1986. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallographica Section A Foundations of Crystallography*, 42(3), pp.140–149.

Renaud, J.-P. et al., 2016. Biophysics in drug discovery: impact, challenges and opportunities. *Nature reviews. Drug discovery*, 15(10), pp.1–20. Available at: http://www.nature.com/doifinder/10.1038/nrd.2016.123\nhttp://www.ncbi.nlm.nih.gov/pubmed/27516170.

Rould, M. a. & Carter, C.W., 2003. Isomorphous Difference Methods. *Methods in Enzymology*, 374, pp.145–163.

Rupp, B., 2010. *Biomolecular crystallography : principles, practice, and application to structural biology*, New York: Garland Science.

Schiebel, J., Krimmer, S.G., et al., 2016. High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits. *Structure*, pp.1–12. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0969212616301356.

Schiebel, J., Radeva, N., et al., 2016. Six Biophysical Screening Methods Miss a Large Proportion of Crystallographically Discovered Fragment Hits: A Case Study. *ACS Chemical Biology*, p.acschembio.5b01034. Available at: http://pubs.acs.org/doi/abs/10.1021/acschembio.5b01034.

Scott, D.E. et al., 2012. Fragment based approaches in drug discovery and chemical biology. *Biochemistry*, 51, pp.4990–5003.

Shannon, C.E., 1949. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1), pp.10–21.

Sheldrick, G.M., 2007. A short history of SHELX. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1), pp.112–122.

Silvestre, H.L. et al., 2013. Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery. *Proceedings of the National Academy of Sciences of the United States of America*, 110(32), pp.12984–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23872845.

Sim, G.A., 1960. A note on the heavy-atom method. *Acta Crystallographica*, 13(6), pp.511–512. Available at: http://scripts.iucr.org/cgi-bin/paper?S0365110X60001266.

Sim, G.A., 1960. The Cumulative Distribution of Structure Amplitudes. *Acta Cryst.*, 13, pp.58–59.

Smart, O.S. et al., 2012. Exploiting structure similarity in refinement: Automated NCS and target-structure restraints in BUSTER. *Acta Crystallographica Section D: Biological Crystallography*, 68(4), pp.368–380.

Smart, O.S. et al., 2011. GRADE, version 1.1.1. Global Phasing Ltd, Cambridge, United Kingdom. http://www.globalphasing.com.

Srinivasan, R., 1966. Weighting functions for use in the early stages of structure analysis when part of the structure is known. *Acta Cryst*, 20, pp.143–144.

Srinivasan, R. & Ramachandran, G.N., 1966. Probability distribution connected with structure amplitudes of two related crystals. VI. On the Significance of the Parameter sigma-A. *Acta Crystallographica*, 20, pp.570–571. Available at: http://scripts.iucr.org/cgi-bin/paper?S0365110X65004796.

Stanfield, R., Pozharski, E. & Rupp, B., 2016. Comment on Three X-ray Crystal Structure Papers. *The Journal of Immunology*, 196(2), pp.521–524. Available at: http://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1501343.

Taylor, G.L., 2010. Introduction to phasing. *Acta Crystallographica Section D: Biological Crystallography*, 66(4), pp.325–338.

Terwilliger, T.C. et al., 2006. Automated ligand fitting by core-fragment fitting and extension into density. *Acta crystallographica. Section D, Biological crystallography*, 62(8), pp.915–22. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2745883&tool=pmcentrez&rendertype=abstract.

Terwilliger, T.C. et al., 2008. Iterative-build OMIT maps: Map improvement by iterative model building and refinement without model bias. *Acta Crystallographica Section D: Biological Crystallography*, 64(5), pp.515–524.

Terwilliger, T.C. et al., 2007. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D: Biological Crystallography*, 64(1), pp.61–69.

Tickle, I. et al., 2004. High-throughput protein crystallography and drug discovery. *Chemical Society reviews*, 33(8), pp.558–565.

Tickle, I.J., 2012. Statistical quality indicators for electron-density maps. *Acta Crystallographica Section D: Biological Crystallography*, 68(4), pp.454–467.

Trueblood, K.N. et al., 1996. Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica Section A*, 52(5), pp.770–781. Available at: http://dx.doi.org/10.1107/S0108767396005697\nhttp://scripts.iucr.org/cgi-bin/paper?S0108767396005697.

Verlinde, C.L. & Hol, W.G., 1994. Structure-based drug design: progress, results and challenges. *Structure*, 2(7), pp.577–587.

Weichenberger, C.X., Pozharski, E. & Rupp, B., 2013. Visualizing ligand molecules in twilight electron density. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 69(2), pp.195–200.

Williams, M.A. & Daviter, T., 2013. *Protein-ligand interactions: Methods and Applications*,

Wilson, A.J.C., 1949. The probability distribution of X-ray intensities. *Acta Crystallographica*, 2(5), pp.318–321. Available at: http://scripts.iucr.org/cgi-bin/paper?S0365110X49000813.

Winn, M.D. et al., 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, 67(4), pp.235–242.

Wlodek, S., Skillman, A.G. & Nicholls, A., 2006. Automated ligand placement and refinement with a combined force field and shape potential. *Acta crystallographica. Section D, Biological crystallography*, 62(7), pp.741–749. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16790930 [Accessed August 12, 2014].

Womack, T.O. et al., 2010. RHOFIT, version 1.2.4. Global Phasing Ltd, Cambridge, United Kingdom. http://www.globalphasing.com.

Zerbe, B.S. et al., 2012. Relationship between Hot Spot Residues and Ligand Binding Hot Spots in Protein − Protein Interfaces.

Zwart, P.H., Langer, G.G. & Lamzin, V.S., 2004. Modelling bound ligands in protein crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 60(12 I), pp.2230–2239.