# Multiple crystallographic dataset analysis to explore crystallographic occupancy and the conformational states of proteins




Elliot Robin Nelson

St Hilda's 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 |
| Dr Carien Dekker | Novartis |

---

A thesis submitted for the degree of

*Doctor of Philosophy*

Trinity 2019

This thesis is dedicated to
Robin Nelson
for
kindling my scientific curiosity

# Acknowledgements

I am indebted to both my supervisors Frank and Charlotte. Charlotte always knew how to motivate me, and keep everything in understandable terms. Frank, whilst being focused on the big picture, helped me with every stage from idea to execution and delivery of this thesis. I would like to thank my industrial supervisors who kept the project relevant in the wider scientific context. Nicholas Pearce, who preceded me in this group was vital to outlining the original hypotheses, and in getting me started in the protein crystallography group at the SGC.

Tobias Krojer provided key insights into using mass spectrometry to calibrate occupancy, and Rod Chalk was invaluable in teaching me how to use the mass spectrometry equipment and how to troubleshoot the mass spectrometry of crystals. Janine Gray provided the challenge of how to resolve electron density of the GEF/GTPase complex, and the corresponding data. I'm glad I could be of assistance to her project. Joy Kristin Petrick and Lena Muenzker made me realise how much I enjoyed acting as a local contact at Diamond Light Source, which has helped me chose my career path. Joy's *T. cruzi* FPPS data provided the challenge of modelling metal ions in superposed models, and I am grateful for her letting me use this as a test case in my thesis. Romain Talon and Laura Diaz Saez provided the NUDT22 crystals. Michael Fairhead helped me learn the basics of protein purification, and supplied the initial batches of protein.

Thanks to the protein crystallography group, the Diamond Light Source I04-1 and XChem teams, which have provided me amazing resources and insights into many areas of structural biology. Thanks to OPIG for always reminding me of the fun side of science. Rachael Skyner, Mihaela Smilova and Conor Wild have helped with improving my code quality. Philippa Hulley, my college advisor provided amazing pastoral support and persuaded me to keep going on many an occasion. The whole protein crystallography group have helped in keeping me sane throughout my four years, especially Romain Talon, Elizabeth MacLean and Yuliya Dubianok. Many people have been vital proofreaders; special thanks go to Yuliya, Janine, Joy, Lizbe Koemkemoer and Andrew Thompson.

I am of course grateful to my mother, brother, friends and family for supporting me. I am awed by my wife Alice, she has made every step of this thesis possible and more enjoyable.

## 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 conatins no other person's work except where stated in the text.

# Abstract

Protein crystallography is a powerful technique, to elucidate binding poses of ligands, which when sufficiently high-resolution data are available, gives atomic detail. Different conformations of the protein can also be revealed, potentially identifying which states are relevant to function and thus informing structure-based design. Recent advances in interpreting electron density from multiple datasets, such as Pan Dataset Density Analysis (PanDDA) have improved ligand detectability. This work focuses on producing higher quality models of protein-ligand complexes and alternate conformations of proteins, by utilising electron density information from multiple datasets.

In chapter 2, a recently developed modelling strategy for protein structures was used to determine conformational differences between datasets. The structural model used is a superposition of the ground state and an altered state. The ground state is modelled from the average electron density of datasets which are sufficiently similar, and do not have significant regions of electron density different to other datasets. The altered state corresponds to local perturbations to the ground state, such as ligand binding. To improve the modelling of regions which undergo large conformational changes, the real space correlation coefficient was used to select the structural model which best explained the electron density.

The superposition of a ground and altered state was for the first time used to model ligands binding near a metal ion site, for models of *T. cruzi* farnesyl pyrophosphate synthase. In order to achieve a well-behaved model external distance restraints which relate the superposed states, had to be augmented with distance restraints between the metal ions, the coordinating protein residues, and the coordinating water molecules. Furthermore, the occupancy of the metal ions needed to vary between unoccupied (0) and fully occupied (1) and not inherit the refined value of occupancy for the nearby ligand.

Understanding what proportion of each dataset is well represented by the ground state, and thus the crystallographic occupancy of the altered state, is important in determining whether these superposed models appropriately explain the electron density without overfitting. To explore crystallographic occupancy of weakly binding ligands, repeated crystallographic datasets of protein ligand complexes were collected. Surprisingly a wide distribution of occupancies for the same ligand was observed (chapter 3). This was confirmed across different refinement programs, with both conventional models of protein-ligand complexes, and by modelling the ligand binding as a perturbation to

the ground state. To determine whether this represented variation in the crystals, a brute force search of ligand occupancy and B factor values, which minimised difference density in the ligand binding region was developed. The exhaustive search method had a smoothly varying function near the minima of difference density. The method led to similar distributions of ligand occupancy to that of the refinement programs. Also note that occupancies below 0.4 were not refined to, or found in the brute force search. This suggests that the wide distribution of occupancies was most likely due to the poor parametrisation of the model rather than a underlying variability between crystals of the same protein ligand complex.

Chapter 4 applied the same modelling strategies, and refinement programs to approximately two thousand protein ligand complexes found by XChem fragment screening. These refinements also showed that very few models had ligand occupancies below 0.4. As in chapter 3, refinement programs resulted in different occupancy and B factor values when comparing the same dataset.

To calibrate the refinement of crystallographic occupancy an orthogonal method to measure the amount of a ligand present in a protein-ligand complex is required. In chapter 5 the development of such a method is discussed by measuring the proportion of a covalently labelled protein (NUDT7) by intact mass spectrometry of crystals. The proportion is determined by mixing two batches of protein, one labelled with a covalent electrophile, one not. The proportion can be accurately measured well below 0.4, and corresponds with the expected amount of labelled compound. Crystals grown with the same proportion of covalent compound were used to calibrate the crystallographic occupancy. It was found that for crystallographic models to sample the occupancy equivalent to say a 0.3 proportion of protein with the covalently bound ligand, the B factors of the ligand needed to be fixed, so that the model did not compensate for low occupancy by increasing the B factors.

Lastly, in chapter 6 multiple datasets were sampled to identify new alternate side chain conformations, with the aim to determine whether the new conformations were correlated with one another. Such a relationship between two distant conformations could indicate allosteric signal propagation across the protein. In a case study of *T. cruzi* farnesyl pyrophosphate synthase, alternate conformations were detectable by sampling electron density around the dihedral angles of side chains, using a tool called Ringer. These appeared at both the allosteric ligand binding site and near the metal ions in the active site. Methods to automate the detection of such related events were attempted but did not isolate such examples.

Overall this thesis presents evidence to show that the use and careful analysis of multiple datasets improves protein-ligand complex models, and can help detect large conformational changes in the protein. By orthogonal measurements of the proportion of a ligand present in a crystal, it was seen that the current parameterisation of occupancy is challenged when modelling weakly binding fragments.

# Contents

**List of Figures** **xv**

**List of Tables** **xix**

**List of Abbreviations** **xxi**

**1 Introduction** **1**

1.1 Why improving crystallographic models of ligands and alternate protein conformations is important? . . . . . 2

1.1.1 Types of interaction between a ligand and a protein . . . . . 3

1.1.2 Approaches to identifying ligands which modulate protein function 3

1.1.3 Producing appropriate crystallographic models of protein-ligand complexes . . . . . 4

1.1.4 The importance of binding affinity, and why it isn't measurable in crystallography . . . . . 5

1.1.5 Can crystallographic occupancy be used to rank compound binding? 6

1.1.6 Can better models of alternate conformations help the understanding of allostery? . . . . . 7

1.2 Target focused screening methods to identify new ligands . . . . . 8

1.2.1 High throughput screening . . . . . 8

1.2.2 Fragment based lead discovery . . . . . 10

1.2.3 Fragment screening techniques . . . . . 11

1.2.4 Lack of agreement between screening methods . . . . . 15

1.2.5 Crystallographic screening . . . . . 17

1.2.6 Introducing compounds to crystals . . . . . 23

1.2.7 Developing screening hits into lead compounds . . . . . 23

1.3 Macromolecular crystallography . . . . . 24

1.3.1 Crystallisation . . . . . 25

1.3.2 Limitations of crystallography for determining protein ligand complexes . . . . . 27

1.3.3 Data collection . . . . . 28

1.3.4 Diffraction data . . . . . 30

1.3.5 B factor . . . 33
1.3.6 Wilson scaling . . . 34
1.3.7 Phase problem in crystallography . . . 36
1.3.8 Data quality metrics . . . 37
1.3.9 Electron density maps in crystallography . . . 38
1.3.10 Detecting ligands in crystals . . . 40
1.3.11 Electron density & model metrics . . . 42
1.3.12 Crystallographic occupancy . . . 46
1.3.13 Model building and refinement . . . 48
1.4 Detecting weak binders using the PanDDA algorithm . . . 59
1.4.1 Group deposition of PanDDA analysis using XChemExplorer . . 62
1.5 Modelling multiple states in crystals . . . 63
1.5.1 Independent refinement of multiple structures . . . 64
1.5.2 Multi copy ensemble . . . 64
1.5.3 Utilising molecular dynamics to produce an ensemble of structures 65
1.5.4 Multi-conformer modelling . . . 65
1.5.5 Limitations of ensemble modelling and fitting multiple states in crystallography . . . 66
1.5.6 Sampling side chain density using Ringer . . . 67
1.6 Modelling protein ligand complexes as superposed states . . . 67
1.6.1 Using the mean of ground state electron density maps to improve apo-model . . . 70
1.6.2 Refinement of superposed models . . . 70
1.7 Intact mass spectrometry of proteins . . . 71
1.7.1 Ultra high performance liquid chromatography . . . 73
1.7.2 Intact mass spectrometry of crystals . . . 73
1.8 Thesis statement . . . 74

**2 Utilising superposed modelling 79**
2.1 Introduction . . . 79
2.2 Adaptations to the superposed modelling paradigm . . . 80
2.3 Metal ions near the ligand binding site of *T. cruzi* required additional distance and occupancy restraints . . . 81
2.3.1 Introduction . . . 81
2.3.2 Methods . . . 84
2.3.3 Results . . . 85
2.3.4 Discussion . . . 90
2.4 Local real space correlation coefficient determined appropriate switch I loop models in a GEF/GTPase complex . . . 91

2.4.1 Introduction . . . 91
2.4.2 Methods . . . 93
2.4.3 Results and discussion . . . 97
2.5 Discussion and future work . . . 100

**3 Crystallographic refinement programs did not converge on occupancy for repeat datasets of protein-ligand complexes 101**
3.1 Introduction . . . 101
3.2 Methods . . . 102
3.2.1 Repeat datasets of protein-ligand complexes of fragment sized ligands . . . 102
3.2.2 Exhaustive search algorithm . . . 106
3.2.3 Sampling coordinate space . . . 108
3.3 Results . . . 111
3.3.1 Refinement programs do not converge on occupancy for weakly binding ligands . . . 111
3.4 Discussion and future work . . . 115

**4 Crystallographic refinement did not model weakly binding fragment ligands with a low occupancy 119**
4.1 Introduction . . . 120
4.2 Methods . . . 121
4.2.1 XChem database . . . 121
4.2.2 Refinement methods . . . 122
4.2.3 Refinements were run using distributed task scheduling . . . 122
4.2.4 Preparation of datasets and refinement by multiple methods . . . 123
4.3 Results . . . 127
4.3.1 Existing refinements showed that low occupancy states are uncommon . . . 127
4.3.2 Longer refinement increased the convergence of occupancy . . . 127
4.3.3 Occupancy distribution varied drastically by refinement method, unlike the B factor distribution . . . 129
4.3.4 Superposed methods consistently had lower ligand occupancy . . . 132
4.3.5 Exhaustive search showed similar behaviour to refinement of superposed models . . . 133
4.3.6 Ligands were sufficiently supported by density to be a representative subset . . . 134
4.3.7 PanDDA background density correction calibrated by refined occupancy of XChem hits . . . 138
4.4 Discussion and future work . . . 139

**5 Best practice for determining crystallographic occupancy using mass spectrometry to calibrate refinement methods 143**
5.1 Introduction . . . 144
5.1.1 Covalent binders as chemical probes . . . 145
5.1.2 Intact mass spectrometry for detection of covalent binding . . . 145
5.2 Methods . . . 145
5.2.1 Compound selection . . . 145
5.2.2 Protein preparation . . . 146
5.2.3 Mass spectrometry methods for protein in solution and protein crystals . . . 149
5.2.4 Pre-crystallisation mass spectrometry detected ratio of covalent compound . . . 150
5.2.5 Post crystallisation mass spectrometry detected ratio of covalent compound . . . 152
5.2.6 Crystallographic data collection and refinement methods . . . 165
5.3 Results . . . 166
5.3.1 Crystallographic occupancy was correlated with ratio of covalent compound detected with intact mass spectrometry . . . 166
5.3.2 Restraining B factor allowed post crystallisation mass spectrometry to calibrate refinement methods . . . 170
5.4 Discussion and future work . . . 170

**6 Detecting correlated motions by sampling rotamer distributions across many electron density maps 175**
6.1 Introduction . . . 176
6.2 Methods . . . 177
6.2.1 Datasets . . . 178
6.2.2 Scaling structure factor amplitudes allowed comparison of electron density across multiple datasets . . . 178
6.2.3 Ringer routine on fragment screening datasets . . . 179
6.2.4 Pairwise analysis of datasets . . . 180
6.2.5 Clustering Ringer plots over datasets for each residue . . . 181
6.2.6 Scoring to detect sub-populations of similar datasets . . . 183
6.3 Results . . . 185
6.3.1 Idealised rotamers were detected in multiple datasets . . . 185
6.3.2 Conformational heterogeneity between datasets varied by residue 185
6.3.3 Gaussian fits approximate the Ringer plot . . . 187
6.3.4 Ringer across multiple datasets detected correlated motion in *T. cruzi* FPPS case study . . . 187

6.3.5 Sub populations of datasets such as bound ligands were not detected by automated method . . . 188
6.4 Discussion and future work . . . 190
6.4.1 Weaknesses in the multiple dataset ringer routine and opportunities for improvement . . . 191
6.4.2 Future work . . . 193

**7 Conclusions and future work** **195**
7.1 Superposed modelling improves the description of the crystal state . . 196
7.1.1 Superposed modelling can better describe crystal state . . . 196
7.1.2 Metal ions require independent restraints in superposed models 197
7.1.3 Future work: model comparison metrics are needed to improve models of weakly binding ligands . . . 197
7.2 Refinement across repeated datasets of protein-ligand complexes and many weakly binding fragment ligands show occupancy refinement is inconsistent . . . 198
7.2.1 Future work: developing a benchmarking set for refinement of weakly binding ligands . . . 198
7.3 Intact mass spectrometry of crystals calibrated occupancy refinement methods . . . 199
7.3.1 Future work: detailed measurement of crystal parameters may be required to ensure that maximal ligand occupancy is achieved prior to cryo-cooling . . . 200
7.4 Correlated motions were detected by sampling electron density across multiple datasets . . . 200
7.4.1 Future work: relating sub-populations detected globally over multiple datasets to associated local residue movements . . . 200
7.5 Summary . . . 201

**Appendices**

**A Code and data accessibility** **205**

**B Fragment screening data** **207**
B.1 DCP2 (NUDT20) . . . 207
B.1.1 Biological relevance of DCP2 . . . 207
B.1.2 Crystal preparation . . . 210
B.1.3 Fragment screening libraries . . . 211
B.1.4 DCP2 structure determination . . . 211
B.1.5 Fragment hits from screening . . . 211

B.2 *Trypanosoma. Cruzi* Farneysl pyrophosphate synthase . . . . . 221
B.3 NUDT22 . . . . . 226
B.3.1 Protein preparation . . . . . 226
B.4 NUDT7 purification . . . . . 227
B.5 Purification details . . . . . 228
B.5.1 *E.coli* metals . . . . . 228

**C Superposed modelling** **229**

**D Repeat datasets of protein-ligand complexes for assessing refinement programs** **231**

**E Intact mass spectrometry of NUDT7** **237**
E.1 Alternate sample preparation: Titanium dioxide spin columns . . . . . 237
E.2 Signal to noise ratio . . . . . 238

**References** **241**

# List of Figures


1.1 Differences between high throughput screening and fragment based screening . . . 9
1.2 XChem fragment screening pipeline . . . 19
1.3 Posied fragment library (DSPL) . . . 22
1.4 Phase diagram of crystallisation . . . 25
1.5 Electron density map types for detecting ligands . . . 41
1.6 Refinement programs used for refinement of crystal structures in the protein databank . . . 57
1.7 `phenix.refine` macro-cycle flowchart . . . 58
1.8 Detecting weakly bound ligands with PanDDA . . . 60
1.9 Superposed modelling schematic diagram . . . 68
1.9 Superposed modelling schematic diagram (continued) . . . 69
1.10 Electrospray ionisation for mass spectrometry . . . 72

2.1 FPPS seen in fragment screening . . . 86
2.1 Two zinc ions at partial occupancy are the most appropriate model of the active site of *T. Cruzi* . . . 87
2.2 Switch I loop and switch II loops adopt different conformations during nucleotide exchange in GEF/GTPase complexes . . . 92
2.3 A covalent binder resulted in loop rearrangement of switch I region of Kalirin/ Rac1 complex . . . 94
2.4 Real space correlation coefficient per residue identifies which switch loop conformation dominates . . . 95
2.5 Real space correlation coefficient for assessing the most appropriate model 96
2.6 Mean real space correlation of loop residues shows when each loop model dominated, or the superposed model is appropriate . . . 98
2.7 High $R_{Work}$ of some structures which could not be classified . . . 99

3.1 Flowchart describing exhaustive search algorithm . . . 107
3.2 Flatness of difference density maps at different occupancy and B factors 109
3.3 Exhaustive search produced a smoothly varying function around the minima 109
3.4 Validation of exhaustive search algorithm . . . 110
3.5 Electron density evidence for repeated datasets of protein ligand complexes 112

3.6 Refinement programs did not agree on occupancy for repeat datasets of protein-ligand complexes . . . 113
3.7 Occupancy distribution varies more than B factor distribution for repeat datasets of protein-ligand complexes . . . 114

4.1 Anatomy of a Luigi task . . . 123
4.2 Flowchart showing re-refinement of existing fragment screening hits . . 124
4.3 Flowchart for methods to prepare ligand CIF . . . 126
4.4 Existing refinement of datasets using REFMAC5 showed that low occupancies were not sampled by refinement . . . 128
4.5 Longer refinements were needed for the convergence of occupancy . . 128
4.6 Refinement programs did not converge on occupancy, but the B factor distributions of the ligand were similar . . . 130
4.7 Occupancy and B factor distributions for XChem hits using different models and refinement programs . . . 131
4.8 Superposed methods had a lower ligand occupancy . . . 132
4.9 Occupancies from exhaustive search were more uniformly distributed than other methods . . . 133
4.10 RSCC distribution shows XChem hits had a good fit to density . . . 134
4.11 RSZD distribution shows XChem hits did not have an excess of difference density over the model. . . . 135
4.12 RSZO across XChem hits . . . 136
4.13 Distribution of the ratio of ligand B factor to the protein atoms surrounding the ligand . . . 137
4.14 Occupancy of ligands was consistently higher than estimated by background density correction . . . 138

5.1 Covalent binder for NUDT7 was a result of fragment merging . . . 146
5.2 NUDT7 wass 100% efficiently labelled by covalent binder . . . 148
5.3 Intact MS measured the proportion of covalently labelled protein without systematic bias prior to crystallisation . . . 151
5.4 Degradation of NUDT7 to lower mass species seen in mass spectra of crystals did not effect proportion of covalent labelling . . . 154
5.5 Degradation peak analysis . . . 156
5.6 Ion suppression reduced MS signal. . . . 157
5.7 Mass spectra of unlabelled NUDT7 showed high signal to noise . . . 158
5.8 Adding 0.05% PEG 10 kDa to NUDT7 reduced the signal to noise ratio of theMS . . . 159
5.9 Alternate crystallisation conditions include conditions with lower molecular weight PEG, and no PEG . . . 160

5.10 Titanium dioxide spin columns did not displace PEG 10 kDa . . . . . 162
5.11 $C_4$ ZipTip™ reduced the amount of PEG present with NUDT7. . . . . 163
5.12 $C_{18}$ ZipTip™ did not reduce presence of PEG . . . . . 164
5.13 Increasing length of methanol gradient did not improve PEG separation 164
5.14 A covalent binder was identified in the electron density of NUDT7 crystals at 40% and 100% ratio of compound present, but not at 5% . . . . . 167
5.15 Crystallographic occupancy increased with the proportion of compound measured by MS . . . . . 168
5.15 Crystallographic occupancy increased with the proportion of compound measured by MS (continued) . . . . . 169
5.16 Ligand occupancies refined with superposed methods were lower . . . 169
5.17 Fixing B factor of ligand improved fit to ratio of covalently labelled protein 171

6.1 Multiple dataset Ringer method summary . . . . . 177
6.2 Dihedral angles in a side chain . . . . . 179
6.3 Clustering ringer plots into groups of similar datasets for each residue . 182
6.4 Distribution of clusters from ringer plot clustering . . . . . 184
6.5 Dominant peaks found by Ringer matched the theoretical distribution of side chain rotamers . . . . . 185
6.6 Ringer plots show residues had different amounts of conformational heterogeneity . . . . . 186
6.7 Gaussian fits captured most but not all features of Ringer plots . . . . . 188
6.8 Correlated motion in *T.Cruzi* FPPS was seen in Ringer plots . . . . . 189
6.9 Automated clustering could not discriminate bound ligands . . . . . 190

B.1 DCP2 structure in complex of Dcp1-Dcp2-Edc3-$m^7$GDP . . . . . 208
B.2 DCP2 domain structure used for fragment screening . . . . . 209

C.1 *T. cruzi* FPPS metal binding site, shown without extra water molecules which were added to structure shown in main text. . . . . 230

D.1 Electron density evidence for ligands in repeat datasets of protein-ligand complexes . . . . . 232
D.2 Electron density evidence for ligands in repeat datasets of protein-ligand complexes . . . . . 233
D.3 Distribution of occupancy and B factor for repeat datasets of protein-ligand complexes . . . . . 234
D.4 Distribution of occupancy and B factor for repeat datasets of protein-ligand complexes . . . . . 235

E.1 Adding 0.5% PEG 10 kDa to NUDT7 reduced signal to noise ratio of MS . 238
E.2 Adding 5% PEG 10 kDa to NUDT7 reduced signal to noise ratio of MS . . 239

# List of Tables

1.1 Crystal systems . . . . . 29

2.1 Geometry and coordination of ions in alternate refinements of *T. cruzi FPPS* 89
2.2 R values for different refinements of switch I loop region . . . . . 97

3.1 Concentration of compounds used in repeat datasets experiments . . . 103

5.1 Location of NUDT7 degradation peaks . . . . . 155

B.1 DCP2 dataset summary . . . . . 212
B.2 DCP2 initial hits . . . . . 212
B.3 DCP2 follow up hits . . . . . 215
B.5 *T. cruzi* FPPS active site hits . . . . . 221
B.6 *T. cruzi* FPPS allosteric site hits . . . . . 223
B.4 *T. cruzi* FPPS dataset summary . . . . . 225
B.7 *T. cruzi* FPPS novel site hits . . . . . 225

# List of Abbreviations

**DMSO** . . . . . Dimethyl sulfoxide

**DSPL** . . . . . Diamond and SGC poised library

**DSiP** . . . . . Diamond poised library

**PDB** . . . . . Protein data bank/ Protein coordinate data bank file format

**MS** . . . . . Mass spectrometry

**ESI** . . . . . Electrospray ionisation

**FBLD** . . . . . Fragment based lead discovery

**MALDI** . . . . . Matrix assisted laser desorption/ionization

**NMR** . . . . . Nuclear magnetic resonance

**PanDDA** . . . . . Pan dataset density analysis [1]

**RSCC** . . . . . Real space correlation coefficient

**RSR** . . . . . Real space residual

**RSZO** . . . . . Real space Z-observed score

**RSZD** . . . . . Real space Z difference

**SGC** . . . . . Structural genomics consortium

**STD-NMR** . . . . Saturation transfer difference nuclear magnetic resonance

**TCRU** . . . . . *Trypanosoma cruzi*

**IPP** . . . . . Isopentenyl pyrophosphate

**DMAPP** . . . . . Dimethylallyl pyrophosphate

**FPP** . . . . . Farnesyl pyrophosphate

**GPP** . . . . . Geranyl pyrophosphate

**FPPS** . . . . . Farnesyl pyrophosphate synthase

**HTS** . . . . . High throughput screening

**PEG** . . . . . Polyethylene glycol

**TFA** . . . . . Trifluoroacetic acid

**HPLC** . . . . . High pressure (performance) liquid chromatography

**UHPLC** . . . . . Ultra high pressure (performance) liquid chromatography

**DHB** . . . . . 2,5-Dihydroxybenzoic acid

**GEF** . . . . . Guanine nucleotide exchange factor

**GTPase** . . . . . Hydrolase enzymes binding the guanine triphosphate nucleotide

**GTP** . . . . . Guanine triphosphate

**ATP** . . . . . Adenosine triphosphate

**GDP** . . . . . Guanine disphosphate

**DCP2** . . . . . Human m7GpppN-mRNA hydrolase

**NUDT20** . . . . . Human m7GpppN-mRNA hydrolase

**DCP2B** . . . . . Human m7GpppN-mRNA hydrolase construct used at Structural Genomics Consortium

**NUDT22** . . . . . Uridine diphosphate glucose pyrophosphatase

**NUDT7** . . . . . Peroxisomal coenzyme A diphosphatase

**NUDT7A** . . . . Peroxisomal coenzyme A diphosphatase construct used at Structural Genomics Consortium

**XChem** . . . . . XChem crystallographic fragment screening facility at Diamond Light Source

**CIF** . . . . . Crystallographic information file

**AceDRG** . . . . . A stereochemical dictionary generator for ligands [2]

**SMILES** . . . . . Simplified molecular input line entry system

**HPC** . . . . . High performance computing

**Coot** . . . . . Crystallographic Object-Oriented Toolkit [3]

**GRADE** . . . . . Global Phasing geometrical restraint generator [4]

**eLBOW** . . . . . Electronic ligand builder and optimisation workbench [5]

**DIMPLE** . . . . . A pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands [6]

**NUDIX** . . . . . Nucleoside diphosphate linked to some other moiety X

**TLS** . . . . . Translation, liberation, screw model for B factors

**DDXXD** . . . . . Asparatic acid rich motif in FPPS

**BDC** . . . . . Background density correction

**TATA box** . . . . Non-coding DNA sequence containing a consensus sequence characterized by repeating T and A base pairs

**TOF** . . . . . Time of flight

**QTOF** . . . . . Quantitative time of flight

**BART** . . . . . Robot used on Diamond Light Source macromolecular beamlines for sample loading

**BUSTER** . . . . . Maximum likelihood macromolecular crystallographic refinement program [7]

**TNT** . . . . . Stereo-chemistry and restraint handling alongside BUSTER [7]

**REFMAC5** . . . . Maximum likelihood macromolecular crystallographic refinement program [8]

**ZipTips™** . . . . Pipette tips packed with functionalised silica beads for chromatographic separation

**$C_3$, $C_4$, $C_{18}$** . . . . Silica beads for chromatographic separation, functionalised with alkyl carbon chains of specified length

**MST** . . . . . Microscale thermophoresis

**mmCIF** . . . . . Macromolecular Crystallographic Information File

**ITC** . . . . . Isothermal Titration Calorimetry

**SPR** . . . . . Surface Plasmon Resonance

**TSA** . . . . . Thermal Shift Assay

**MD** . . . . . Molecular Dynamics

**NCS** . . . . . Non-crystallographic symmetry

**DLS** . . . . . Diamond Light Source

# 1 Introduction

## Contents


**1.1 Why improving crystallographic models of ligands and alternate protein conformations is important? . . . . . 2**
1.1.1 Types of interaction between a ligand and a protein . . . . . 3
1.1.2 Approaches to identifying ligands which modulate protein function . . . . . 3
1.1.3 Producing appropriate crystallographic models of protein-ligand complexes . . . . . 4
1.1.4 The importance of binding affinity, and why it isn't measurable in crystallography . . . . . 5
1.1.5 Can crystallographic occupancy be used to rank compound binding? . . . . . 6
1.1.6 Can better models of alternate conformations help the understanding of allostery? . . . . . 7
**1.2 Target focused screening methods to identify new ligands . . . . . 8**
1.2.1 High throughput screening . . . . . 8
1.2.2 Fragment based lead discovery . . . . . 10
1.2.3 Fragment screening techniques . . . . . 11
1.2.4 Lack of agreement between screening methods . . . . . 15
1.2.5 Crystallographic screening . . . . . 17
1.2.6 Introducing compounds to crystals . . . . . 23
1.2.7 Developing screening hits into lead compounds . . . . . 23
**1.3 Macromolecular crystallography . . . . . 24**
1.3.1 Crystallisation . . . . . 25
1.3.2 Limitations of crystallography for determining protein ligand complexes . . . . . 27
1.3.3 Data collection . . . . . 28
1.3.4 Diffraction data . . . . . 30
1.3.5 B factor . . . . . 33

1.3.6 Wilson scaling . . . . . 34
1.3.7 Phase problem in crystallography . . . . . 36
1.3.8 Data quality metrics . . . . . 37
1.3.9 Electron density maps in crystallography . . . . . 38
1.3.10 Detecting ligands in crystals . . . . . 40
1.3.11 Electron density & model metrics . . . . . 42
1.3.12 Crystallographic occupancy . . . . . 46
1.3.13 Model building and refinement . . . . . 48
**1.4 Detecting weak binders using the PanDDA algorithm . . . . . 59**
1.4.1 Group deposition of PanDDA analysis using XChemExplorer . 62
**1.5 Modelling multiple states in crystals . . . . . 63**
1.5.1 Independent refinement of multiple structures . . . . . 64
1.5.2 Multi copy ensemble . . . . . 64
1.5.3 Utilising molecular dynamics to produce an ensemble of structures . . . . . 65
1.5.4 Multi-conformer modelling . . . . . 65
1.5.5 Limitations of ensemble modelling and fitting multiple states in crystallography . . . . . 66
1.5.6 Sampling side chain density using Ringer . . . . . 67
**1.6 Modelling protein ligand complexes as superposed states . . . . . 67**
1.6.1 Using the mean of ground state electron density maps to improve apo-model . . . . . 70
1.6.2 Refinement of superposed models . . . . . 70
**1.7 Intact mass spectrometry of proteins . . . . . 71**
1.7.1 Ultra high performance liquid chromatography . . . . . 73
1.7.2 Intact mass spectrometry of crystals . . . . . 73
**1.8 Thesis statement . . . . . 74**

## 1.1 Why improving crystallographic models of ligands and alternate protein conformations is important?

Producing appropriate models of protein-ligand complexes is key to understanding the structure to function relationship of proteins [9]. Ligands are sought to modify the functional behaviour of proteins. Understanding the location, and strength of binding (binding affinity) is key to choosing which ligands may be appropriate lead candidates for drug molecules.

### 1.1.1 Types of interaction between a ligand and a protein

When small molecule ligands bind to proteins, they may have a functional effect. They can act in many different ways such as: substrates, inhibitors or allosteric modulators [18]. Substrates are the molecules on which the protein acts [19]. Substrates typically bind to a protein in the region defined as an active site. Non-covalent inhibitors usually block an active site, by having a higher affinity than the protein's substrate, and reduce the functional activity of the protein [20].

Covalent inhibition can be considered a two step process, first an inhibitor reversibly associates with the protein, bringing the covalent warhead close to a targeted residue, then the reactive entity in the covalent warhead and the nucleophile in the target form an irreversible covalent bond [21]. Covalent inhibition can be both in the active site [22–24], or in distal allosteric sites [25].

Allosteric modulators bind to the protein distally from the active site, and regulate the function of the protein [18]. Activators are modulators which increase the activity of the protein, and usually act allosterically [26]. Small molecules can also act as cargo, transported alongside proteins across membranes.

### 1.1.2 Approaches to identifying ligands which modulate protein function

Drug development can be target focused, where a specific protein target is identified and then chemical matter is sought to inhibit or alter the function of that target. Alternatively a systems or phenotypic approach may be taken, whereby the effects on a physiologically relevant biological system (such as a cell, tissue or organ) are interrogated to identify biologically active compounds [27–29]. Analysis of 113 first in class drugs, released between 1999 and 2013 show that 71% were developed using taget-based approaches (41% small molecules, 30% biologics) and 29% were developed using systems approaches. Of the systems based approaches 7% of drugs were found in phenotypic screening, and 22% were found through a chemocentric approach which begins the drug development process by using compounds with known pharmacology [30].

To identify modulators of protein function a screening strategy to identify chemical matter against that target is carried out. High throughput screening (HTS) approaches, discussed in section 1.2.1, look for relatively strong binders, but have to search a large chemical library [31]. Modifications to an initial hit act to alter the chemical moieties of a compound to make it more potent [32], or adjust its properties to be more drug-like [33]. Fragment based approaches, discussed in section 1.2.2 usually identify weaker binders which can be grown [34, 35], linked [36, 37] or merged [38] to form larger and more potent molecules [39, 40].

This thesis focuses on target-based approaches, for which a target has been identified and can be isolated to be screened against potential functional modulators. Furthermore, this thesis focuses on producing appropriate crystallographic models of weakly binding fragment sized precursors to small molecule drugs.

### 1.1.3 Producing appropriate crystallographic models of protein-ligand complexes

Crystallographic protein-ligand complexes, at sufficiently high-resolution can give atomic detail on the ligand binding pose, and the interactions made with the protein residues. Binding pose information from crystallography can hugely aid hit elaboration [10], as it can be used to determine which vectors from the compound are less likely to sterically clash with the protein, altering the binding pose [11, 12]. Compounds which follow these vectors can be found by efficient database searches [11] using techniques such as reaction vectors.

Producing the most appropriate ligand model, requires an understanding of macromolecular crystallography, as discussed throughout section 1.3. Most importantly an understanding of model building and refinement, as outlined in section 1.3.13. To choose the most appropriate model, it is important to understand how to verify the quality of a model, the quality of the underlying dataset, and how well the model fits the dataset, as discussed in section 1.3.11 on metrics.

Using multiple models to describe the conformations that the protein can adopt, can be beneficial, a residue may be best described when it is able to occupy two distinct

conformations. If a single structure was assumed for such a residue, information would be lost, with the B factor describing the conformational heterogeneity for each atom. Existing methods to model multiple states of a protein into the electron density are discussed in section 1.5. When modelling multiple states, it is important firstly to assess whether the model is overfitting the data, as increasing the number of states increases the number of free parameters, making it easier to fit to the density.

This thesis focuses on using multiple datasets to inform what model is the most appropriate. Section 1.4 discusses previous work, PanDDA, which uses multiple datasets to detect perturbations from a ground state which can correspond to both ligand binding and larger protein conformational changes. Section 1.6 discusses the modelling framework introduced alongside PanDDA, whereby perturbations to a ground state of the protein are modelled in an altered (bound) state which is superposed with the ground state.

### 1.1.4 The importance of binding affinity, and why it isn't measurable in crystallography

The binding affinity measures the strength of binding to a protein. When developing a ligand which may inhibit or otherwise modify the function of a target protein, it is preferable to know the change in binding affinity when the molecule is elaborated, so that the relative improvement of different compounds can be compared.

The binding affinity, which is typically measured by equilibrium dissociation constant:

$$K_d = \frac{[P][L]}{[LP]} = \frac{k_{\text{on}}}{k_{\text{off}}} \tag{1.1}$$

where

- $[P]$ is the concentration of the protein.
- $[L]$ is the concentration of the ligand.
- $[LP]$ is the concentration of the protein ligand complex.
- $k_{\text{on}}$ is the on-rate constant.
- $k_{\text{off}}$ is the off-rate constant.

There are therefore two general strategies for determining binding affinity: an equilibrium experiment where one determines the extent of the reaction as a function of the concentration of one of the reactants, or a kinetic experiment which determines the rates of forward and reverse reactions as the concentration of one of the reactants [13].

Crystallography cannot directly determine ligand binding affinities, as it cannot determine either on or off rate, or the concentration of the reactants.

Biophysical methods such as surface plasmon resonance (SPR) [14], isothermal titration calorimetry (ITC) [15], micro-scale thermophoresis (MST) [16] and nuclear magnetic resonance (NMR) [17] can be used to quantify the affinity of a ligand. These are discussed as techniques used in fragment screening in section 1.2.2.

### 1.1.5 Can crystallographic occupancy be used to rank compound binding?

Crystallographic occupancy measures the proportion of sites in the crystal lattice for which an atom is present. It is logical to consider comparing the crystallographic occupancy of two ligands bound in the same site to compare the relative strength of binding. It is possible to estimate the required concentration of compound in a crystallisation drop, to lead to highly occupied sites. To do this, the derivation given in [41] is followed:

Consider that the fraction $X_{PL}$ of occupied receptor sites in equilibrium is given by the amount of receptor in the complex divided by the total amount of receptor present:

$$X_{PL} = \frac{[PL]}{[P] + [PL]} \tag{1.2}$$

Which can be expressed in the terms of the ligand concentration, by using equation 1.1:

$$X_{PL} = \frac{[L]}{K_d + [L]} \tag{1.3}$$

Thus:

$$[L] = \frac{X_{PL} \cdot K_d}{1 - X_{PL}} \tag{1.4}$$

Therefore, if one is seeking to obtain a site occupancy of 80-90%, these estimates would require a concentration of the ligand in the crystallisation drop to be 4-10 times the $K_d$ to get a highly occupied site. This is problematic with low affinity binders, as the required

concentration is often far too high to be tolerated by the crystal, or in some cases too high for such a stock solution of the compound to soluble.

This estimate of site occupancy suggests that it may be feasible to rank ligands using their occupancy, along with the soaking concentration required to reach that occupancy. However, for that to be the case, it would require the occupancy of the ligand which is measured by crystallography to be consistent. Chapter 3 shows that the crystallographic occupancy for a single protein-ligand complex has a wide distribution of values. Chapter 4 looks at the distrubution of ligand occupancies across refinement of approximately two thousand protein ligand complexes from XChem crystallographic fragment screening. It showed that low occupancy values are not refined to without further restraints, and that the same dataset would refine to have a different occupancy in different refinement programs. Chapter 5 addresses the issue of occupancy values in crystallographic models, by investigating calibration of crystallographic occupancy using intact MS of crystals. There it was shown that low occupancies can be modelled if the ligand B factors are restricted from varying.

### 1.1.6 Can better models of alternate conformations help the understanding of allostery?

Allostery refers to the ability for a local change, often associated with the binding of a ligand, to be transmitted to another location on a macromolecule, resulting in regulation of activity [42]. This thesis aimed to compare multiple datasets of the same protein, some with local perturbations such a ligand binding, and determine whether these local changes can be correlated to other changes in the protein structure and thus be signatures of allostery. Chapter 6 uses the Ringer method on multiple datasets, this samples electron density by rotating each side chain around each dihedral axis, to detect alternate conformations of those side chains. In a case study, it was possible upon ligand binding to identify new conformations for both active site and distal residues in the same datasets. Chapter 6 attempts to automate the detection of such related conformational changes.

#### 1.1.6.1 Allostery

Allosteric communication is conjectured to proceed either via a sequential series of conformational changes between the allosteric binding site and distal functional site, or through changes to available vibrational modes in a protein [43]. The vibrational model for protein allostery assumes the protein has multiple vibrational modes. The binding of an allosteric ligand creates a change between vibrational modes, without a series of steric changes between allosteric and functional sites. An allosteric network is the series of local interactions which are responsible for the conformational changes from the allosteric ligand binding site to a distant functional site [44]. These local interactions can be both movements of the protein side chains, as well as shifts of the backbone atoms [45]. Another way to characterise the protein is the set of conformations it can take, known as the conformational ensemble [46].

When investigating allostery, the aim was to understand the affect of ligand binding and associated protein conformational changes, at a distant site on the activity of the protein. Most allosteric signal pathway detection methods, utilise techniques such as MD that predict correlated motion from structure [47, 48]. These usually do not use electron density information, only structural models, potentially reducing the sensitivity of the methods. This is because a structural model usually only represents the protein conformation with highest evidence in the electron density.

## 1.2 Target focused screening methods to identify new ligands

### 1.2.1 High throughput screening

Identifying a hit compound in drug discovery often begins with HTS. An assay screens a target against libraries of up to multiple millions of drug-like compounds, usually at rates of 10,000 to 100,000 compounds per day [49, 50].

Assays can measure catalytic activity from purified enzymes, reconstituted complexes [51], cellular extracts [52, 53], or measure a phenotypic response in intact cells [54].

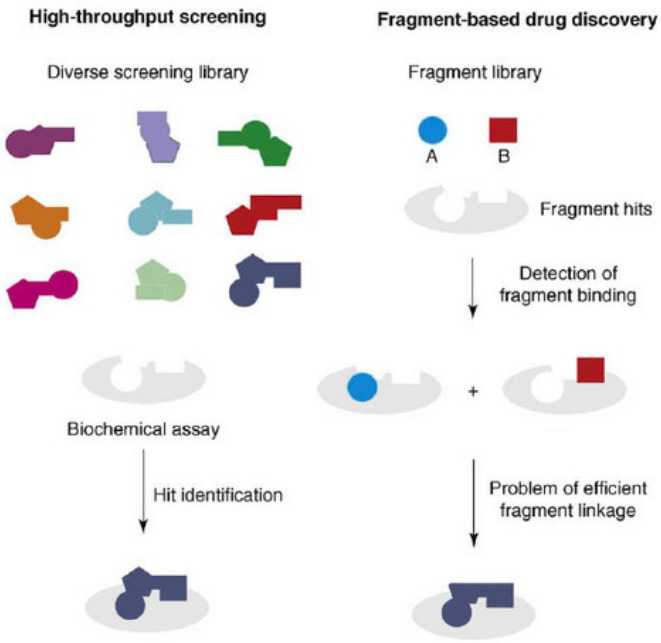



**Figure 1.1:** HTS libraries are larger than fragment screening libraries as they use larger more complex molecules. Hit rates are typically higher in fragment screening, but hits tend to be weaker and require linking, growing or merging which may be non-tractable. Adapted from [55].

HTS assays must be scalable, reproducible, cost effective and sufficiently sensitive [49] to detect compounds of low potency or efficacy.

The biological response in HTS assays is most frequently measured by detecting photons which were produced by florescent or luminescent mechanisms. Typically, an assay strategy will be designed for a recombinant protein target identified to be involved in the disease pathway of interest. Inhibition of enzyme activity is the most common property to measure with a HTS assay. The assay strategy should involve orthogonal secondary screening techniques, to identify false positives, and allow prioritisation of identified hits [56].

These hits can then be compared in biological response, chemical properties and diversity to inform the synthesis or purchase of analogue compounds. Further development of the hit compound towards a chemical probe or lead drug-like molecule involves iterating between assay result and compound elaboration.

HTS has led to the development of many drugs and drug candidates [27]. The relatively small size of such screening libraries (approximately $10^6$ compounds) poorly sample the chemical space of drug-like molecules. This is estimated to comprise of

approximately $10^{63}$ compounds [57]. Other potential issues with HTS include that some novel target classes are not tractable with HTS, and do not yield progressible lead candidates in up to 50% of projects [58]. HTS can suffer from false positives [59, 60], such as aggregates of many individual molecules forming particles of 30-400nm, which leads to non-specific inhibition [61].

### 1.2.2 Fragment based lead discovery

Fragment based lead discovery (FBLD) aims to efficiently sample chemical space with small molecules of low complexity [62, 63], and then develop those fragments into larger more potent molecules by growing, linking and merging fragment moieties [40]. Fragments are typically considered to be between 150-300 Da and contain up to 20 non-hydrogen atoms [64].

The premise of initially screening with small fragment molecules is based on molecular complexity theory [62, 63]. Briefly, as larger molecules have more features, there is the potential to form more interactions with a protein target. These interactions include any form of contact, i.e. hydrogen bond formation or hydrophobic interactions. Each contact contributes to the free energy of binding. Moreover, binding requires presence of a kinetically accessible pathway by which the bound state can be formed. A molecule and its potential binding site can be respectively divided into features which are either favourable or unfavourable. A molecule which was perfectly complementary to a binding site could have its interaction removed by the placement of a single unfavourable feature, for example, a sterically clashing methyl group. This complementarity becomes less likely as a molecule becomes larger. Therefore, fragment sized molecules are promiscuous: should be able to bind to a greater number of sites in proteins. Fragment screens usually lead to higher hit rates than HTS. Due to the lower number of potential interactions available in a smaller molecule, the affinity of fragment hits is usually low.

### 1.2.3 Fragment screening techniques

#### 1.2.3.1 Protein observed nuclear magnetic resonance

In protein observed NMR, fragment binding is seen by chemical shift changes in an isotope-labelled target protein ($^{15}$N and/or $^{13}$C). The binding affinity can be measured by measuring the dissociation constant, $K_d$, which can be obtained by ligand titration. The typical range of detection is high and lies between 1 nM and 1 mM. Additionally resonance assignment can provide information on the location of the ligand binding site [65]. Major disadvantages of the method include requiring large amounts of isotope labelled protein, and a protein size limitation (<50 kDa) [66]. Overall, this technique is responsible for less than 5% of primary screening [67].

#### 1.2.3.2 Ligand observed nuclear magnetic resonance

Ligand based NMR methods utilise a variety of techniques including:

- Saturation transfer difference (STD NMR)[68]
- Relaxation experiments [69, 70]
- WaterLOGSY [71]
- $^{19}$F Reporter screen [72]

Ligand observed NMR techniques are prevalent and account for approximately 18% of primary screening [67]. High throughput is often achieved by screening ligand cocktails. Ligand cocktails are required to be seperable by NMR and thus have disperse non-overlapping peaks, such cocktails can contain up to twenty different fragments [73]. These methods typically require less protein than protein observed NMR. However, no binding site information is obtained, except in quantitative STD NMR [74], or by $^{19}$F reporter screen NMR [72]. In $^{19}$F reporter screening the fragment's ability to displace a reporter ligand containing a $^{19}$F nucleus is measured. For this, an existing binder containing a $^{19}$F nucleus is required. For further details of the multitude of NMR techniques for ligand screening see [70, 75].

#### 1.2.3.3 Biochemical binding assays

Biochemical assays used in HTS of larger molecules can be customised to work with fragment libraries [76]. Optimisations for weak fragments allow detection of inhibition at low mM. Detection range is dependent on the assay for each protein and lies between low nM and µM.

As HTS of larger compounds is an established technique, companies and laboratories are primed to attempt such assays with fragment sized compounds, therefore these techniques are high prevalent accounting for approximately 30% of primary screening [67].

#### 1.2.3.4 Thermal shift assay

TSA uses differential scanning fluorimetry to study the thermal unfolding of a protein [77]. This can be measured by a fluorescent dye that binds to hydrophobic patches of the protein as they are exposed [78]. As ligand binding can thermally stabilise the protein, a shift in the melting temperature is expected upon ligand binding. Other TSA techniques include measuring intrinsic tryptophan fluorescence [79], and GFP tagging of the protein [80]. In some ligand series the extent of stabilisation relates to affinity. As TSA is a plate based technique high throughput can be achieved [81]. As protein denaturation is determined by the thermodynamics of protein folding, which may not be affected by the presence of an affine ligand, TSA is not amenable to all proteins. False positives and false negatives are common [82]. One of the disadvantages of the technique is the lack of information on binding location. TSA accounts for approximately 15 % of primary screening [67].

#### 1.2.3.5 Surface plasmon resonance

Using SPR for the detection of ligand binding to protein requires immobilisation of the proteins to a sensor chip. Any fragment which binds to the protein increases the mass near the sensor surface. The change in the refractive index of the interface is detected by the instrument. If the target protein can be immobilised, and remains active on immobilisation, then only small amounts of protein are needed, to test a large number of fragments, as they are flowed over the sensor [83]. SPR can yield dose-response

curves and thereby affinity measurements, although this requires ligand solubility to be much higher than the $K_d$ [14]. The location of the active site is only determinable by comparing wild type protein to active site mutants, or through competition assays with compounds with known binding sites.

SPR can measure a wide range of affinities 1pM to approximately 2 mM, but can struggle to detect weakly binding ligands. Especially when the ligands are solubilised in DMSO, a solvent with high refractive index, adjusting an appropriate DMSO background can be challenging [84]. As a ligand screening technique SPR is quite prevalent, accounting for $\sim$7% of primary screening [67].

#### 1.2.3.6 Weak affinity chromatography

In weak affinity chromatography compounds are passed over a column with protein which is immobilised on small particles. Differences in compound retention time caused by interactions with the immobilised protein allow affinity measurement of fragment mixtures [85]. This technique is still in development and not in prevalent use. Although it can reveal site specificity by competition assays, it currently has low throughput. By design this technique is focused on weaker binders, with a detection range of 10 mM to 1 µM [86].

#### 1.2.3.7 Isothermal titration calorimetry

ITC directly measures the heat that is released or absorbed during a ligand binding event [87]. A titration of ligand is performed against the protein, and can be used to determine the affinity $K_d$, stoichiometry, enthalpy $\Delta H$ and entropy $\Delta S$ and in some cases $k_{\mathrm{on}}$ and $k_{\mathrm{off}}$ [88]. ITC requires a relatively large amount of protein, whilst being low throughput, thereby it is typically a secondary screening method. ITC is also subject to systematic errors, such as incorrect concentration measurement, or incorrect estimation of experimental parameters (i.e. dilution heats), which can lead to disparity between measurements [89]. The throughput lies between ten and fifty compounds a day on a single calorimeter.

ITC is one of most well-regarded methods for reliable measurement of affinity, and has a wide range of low nM to 5 mM. Fragments which bind with low affinity may not

be fully classified through an ITC experiment, but measures such as enthalpic efficiency [87] can be obtained from ITC on low affinity fragments.

#### 1.2.3.8 Microscale thermophoresis

Thermophoresis involves movement of a protein along a temperature gradient. Some proteins move toward higher temperatures, others lower, the movement can be monitored by labelling the protein with fluorescent dyes. A MST instrument loads protein in solution, typically in low volume (i.e. 4 µl on Nanotemper Monolith NT1.15) capillaries. The fluorescent dye is excited by a appropriate light source. An infrared laser provides local sample heating. A typical experiment begins by measuring fluorescence without a heat source, at this stage the sample is homogenously distributed. After the infrared laser heats the sample volume, within 1 second a large change in the fluorophore properties is detected due to the fast temperature change. The thermophoretic movement of the fluorescently labelled molecules is measured. After 30 seconds the infrared laser is turned off. Mass diffusion leads to a jump in fluorescent signal. In a ligand binding experiment a titration of a nonfluorescent ligand results in a gradual change in thermophoresis which can be plotted to yield a binding curve, which can be fitted to derive binding constants (i.e. $K_D$). Ligand binding effects the migration behaviour, by altering the overall charge or hydration shell of the protein [16]. MST has a wide reported detection range of 1 pM to 10 mM [90]. This method is medium throughput measuring up to hundreds of compounds per day.

#### 1.2.3.9 Covalent libraries by intact mass spectrometry

Covalent inhibition is in general a two step process. First, an inhibitor associates reversibly with the protein of interest. This brings the covalent moiety (warhead) in proximity with target residue. After this a reaction occurs between the reactive residue and the reactive inhibitor which forms a covalent bond [21]. Covalent inhibitors have a prolonged duration of action, and high potency. Covalent inhibitors contain electrophiles that react with nucleophilic residues in enzymes, such as cysteine, serine or methionine. Electrophilic warheads include epoxides, aziridines, esters, ketones and nitriles [91].

To achieve selectivity the nucleophilic residue should not be conserved across the protein family. Historically, covalent inhibitors have been avoided due to concerns about compound promiscuity [92]. Covalent screening can be carried out by incubating a library of disulphide containing fragments with the target, this is known as disulphide tethering [93]. Direct screening of electrophiles by intact MS [94], requires checking the library for high reactivity species, as well as assessment of compound promiscuity. Intact MS allows detection of the total mass of the protein, the fraction of labelling at different concentrations can be used to determine relative binding of compound series. Stoichiometry can be determined, and likely binding sites can be identified. However, in proteins with multiple nucleophilic residues, an alternative assay is needed for validation [94, 95].

#### 1.2.3.10 Native mass spectrometry

Native MS allows detection of weakly binding non-covalent species. As well as the determination of stoichiometry both relative and absolute binding affinities can be obtained. The protein is transferred into the gas phase as ions with no molecular fragmentation [96]. Successful ionisation and retention of the native state requires optimisation of the pH, and buffer conditions. The binding of ligands is assessed by comparing spectra acquired with and without the ligand species present [97]. If binding occurs, additional peaks will be visible corresponding to the mass over charge ratio of the complex. Ligand titration can be used to determine affinities. Native MS is generally limited by mass resolution to proteins under 80 kDa [98]. Screening can be high throughput if coktails of ligands are used. Site specificity can be determined in competitive assays [99].

### 1.2.4 Lack of agreement between screening methods

NMR and crystallography were originally used as primary screening techniques in FBLD [100–102], as these techniques elucidate the structural activity relationship. However, the relatively low throughput of X-ray crystallography led to cascade approaches [40, 103], where primary screening is carried out by faster techniques, able to detect low

affinity binders ($K_d$: 0.1 to 10 mM), such as NMR, native MS [98], high concentration biochemical screens (HCS), TSA or SPR [14]. Secondary screening on hits from these techniques, is usually performed by protein observed NMR or X-ray crystallography to obtain structural information on the hit.

In a recent study, a single target validation of the cascade approach was carried out. 361 compounds were screened using six methods, notably no hits were detected in all methods. 44% of hits (71 hits) were only detectable in crystallography [104]. If a screening cascade of any two techniques would have been used, at least 73% of the crystallographic hits would have failed to progress through the cascade. Independently, screening of the core domain of HIV-1 integrase, was carried out by two groups each using different screening cascades [105]. Their library screens shared 455 common compounds from the Maybridge fragment library [106]. One group identified 84 hits by cocktail screening in saturation transfer difference NMR. Of these 62 were confirmed as single compounds by $^{15}$N heteronuclear single-quantum coherence (HSQC) and only 16 of these were detectable in X-ray crystallography. The second group screened the compounds singly by SPR techniques, finding 16 hits, 6 of which could be confirmed by X-ray. Only 3 hits overlapped between the groups, one of which was detectable in crystallography.

The lack of agreement in screening methods highlights opportunities to improve fragment screening. It is not yet clear whether the low overlap is primarily due to the different sensitivities of the experiments, or the difference in repeatability between the techniques. It could be argued that hits that are detected in multiple screening techniques are more likely to have a higher affinity, assuming that these hits are sufficiently specific. It is also important to question whether techniques that have a low number of repeats of either mixed compounds in cocktails, or singletons are the most appropriate experiment design in fragment screening.

For a technique to become prevalent for fragment lead discovery it needs to be high throughput, automatable and verifiable. The agreement of other techniques with crystallography will likely continue to be explored as fragment screening by crystallography increases in throughput. As other techniques can rank by binding affinity, they

are required for quantitative ranking of ligands. Techniques such as SPR are easier to parallelise thus are likely to remain prevalent.

### 1.2.5 Crystallographic screening

Crystallographic fragment screening is the process of investigating small ligands which bind to protein crystals. The compound is introduced either by soaking with the apo protein crystal, or by co-crystallising with the ligand in the crystallisation condition. Crystallography can give atomic detail on the pose of the ligand in its binding site and interactions formed between the fragment ligand and the protein.

Historically crystallography has been too low throughput to act as a primary screening technique. However, recent increases in beamline flux, developments in sample loading robotics [107], crystal soaking with acoustic droplet ejection [108] and automated data collection queues [109] have enabled the medium throughput of up to 500 crystals per day on a single beamline.

Developments in downstream analysis of crystals has also been important to increase the throughput of crystallographic fragment screening. Firstly, the automation of data reduction pipelines which take diffraction images through from spot finding, indexing, scaling, merging, integration and molecular replacement to produce an initial structural model allows for rapid data analysis [6, 110]. Hit detection can be laborious as it requires manual inspection of many electron density maps. The Pan Dataset Density Analysis (PanDDA) [1] algorithm discussed in section 1.4 has improved this, by identifying outlying density across the many nearly-isomorphous crystals and allowing a user only to inspect those with significant changes for the original ground state of the protein.

Crystallographic fragment screening requires a stable crystal system, which can produce many similarly diffracting crystals [111]. Those crystals need to be amenable to soaking with compounds, usually solubilised in solvents such as dimethyl sulfoxide (DMSO) which can damage the crystal integrity and reduce the maximal resolution of diffraction [111].

Compounds can be introduced one per crystal well, or as a cocktail of multiple ligands. A cocktail reduces the number of crystals required for X-ray screening, but requires a single hit to dominate and be clearly identifiable in the electron density [112, 113].

#### 1.2.5.1 XChem pipeline for high throughput crystallographic fragment screening

XChem at Diamond Light Source (DLS) is a fragment screening pipeline which facilitates high throughput crystallographic screening overviewed in figure 1.2 [111]. As of late 2018, the facility typically allows a naïve user with a reliably crystallising protein, and diffracting system to be characterised and screened against approximately 500 compounds in a week.

Diffraction datasets in this thesis have been collected using the XChem pipeline.

XChem utilises a series of high throughput methods. A user will first image their plates (ROCK IMAGER®, Formulatrix), which have been integrated with TexRank [114] to allow quick viewing of plates, ranked according to the prevalence of crystals. Targeting coordinates can be saved through a visual interface, to indicate where in the crystal drop compound should be introduced.

Acoustic droplet ejection (Echo 550®, LABCYTE) allows precise compound delivery to the crystal well in small increments (2.5 nl). Diffusion across the drop has been shown to be the most effective targeting strategy [108].

Crystal mounting is aided by the Shifter (Oxford Lab Technologies), a semi-automated x-y stage. It increases the typical mounting rate of approximately 20 mounts per hour to 100-200 crystal mounts per hour, without a reduction in diffraction quality (manuscript submitted, Nathan Wright, SGC). This is achieved by removing the tasks of cutting open the well, and tracking the experiment by utilising a laboratory management system, SoakDB [111]. This leaves the user free to focus on the dextrous challenge of crystal mounting.

In cases where cryoprotectant is needed the Echo can be configured to dispense the cryoprotectant onto the crystallisation drop. This is usually undertaken just prior to mounting, and the cryoprotectant is added at the same targeted location as the compound addition. It is possible to use the shifter to perform a dip-soak experiment

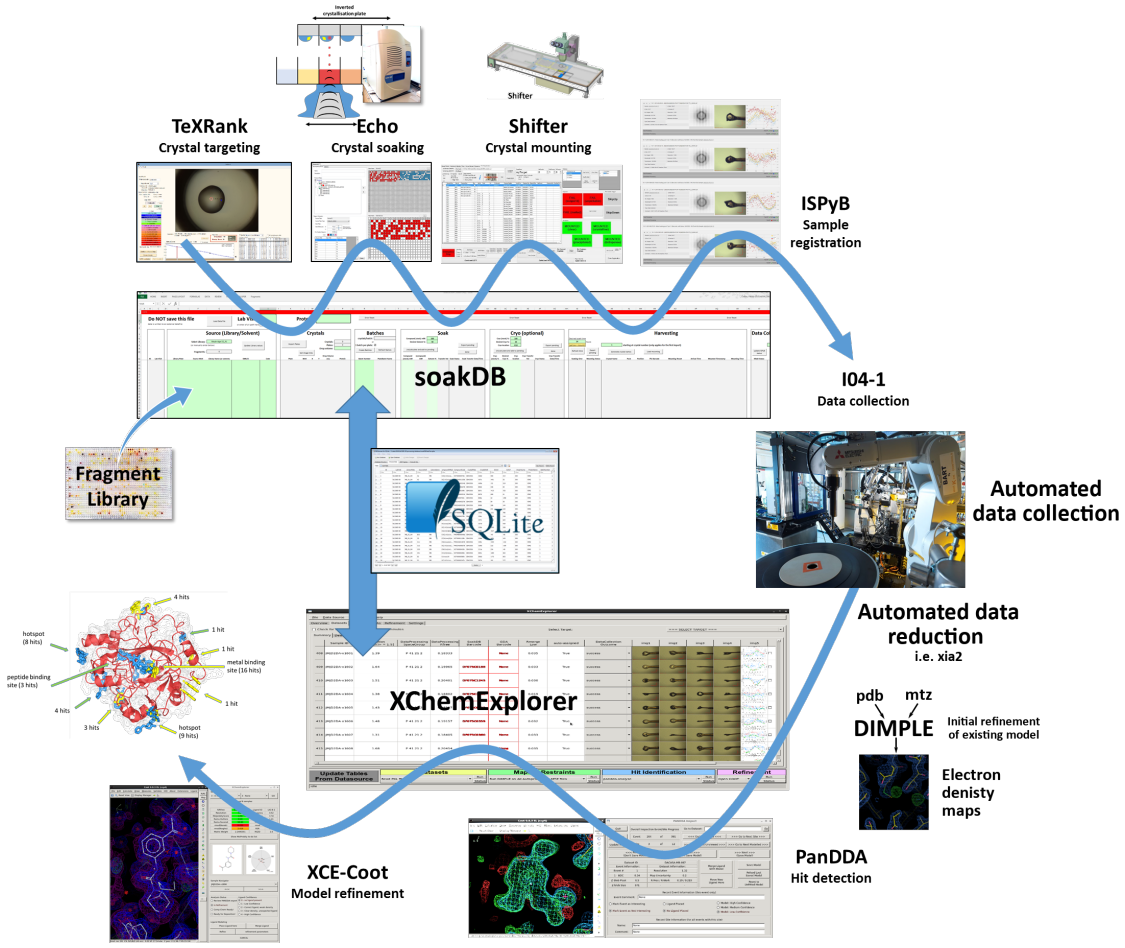



**Figure 1.2:** Overview of XChem platform at Diamond Light Source. Crystal plates are imaged and subsequently ranked by TexRank [114], to determine whether a crystal is present. A location on the plate is specified by the user, for soaking with fragments by acoustic droplet ejection [108] using an Echo®liquid handler (Labcyte). Crystal mounting is aided by the Shifter (manuscript submitted, Nathan Wright, SGC), a semi-automated x-y crystal harvesting stage. soakDB records laboratory actions to a database. Diffraction data are primarily collected in an automated mode [109] at the I04-1 beamline. Data reduction (such as integration, indexing and scaling) steps are carried out by automated pipelines [110]. Electron density maps are calculated with the existing model for the protein, and the newly collected diffraction data using DIMPLE [6]. Data analysis is performed with XChemExplorer [115], and utilises the PanDDA algorithm [1] for detection of weakly binding fragments. Within the XChemExplorer interface it is possible to refine the model using a coot interface (XCE-COOT) which calls the refinement routines for superposed models discussed in section 1.6. The output of XChem platform is currently a series of structure of protein-ligand complexes with fragment sized molecules.

where the crystal is briefly transferred into cryoprotectant before flash cooling [116]. The high throughput nature of the shifter can be used to test multiple cryogens against diffraction quality, which may be necessary in cases with inconsistent diffraction quality.

On the beamline (usually I04-1 at DLS), the BART robot [107] enables faster sample loading and exchange. Recent improvements to X-ray centring speed and reliability have increased throughput allowing more crystal systems that can be screened automatically without human intervention.

If, as is normally the case for XChem projects, the structure of the apo-protein is known, DIMPLE [6] can be used to obtain an initial model and electron density map. An existing structural model is supplied, either from the PDB [117–119], or as a PDB file alongside a merged MTZ file. The DIMPLE pipeline runs macromolecular refinement using REFMAC5 [8] after preparatory steps such as converting intensities to amplitudes, choosing a $R_{free}$ set and reindexing if necessary. Molecular replacement using phaser [120] is run within DIMPLE if there are significant changes in the unit cell.

The user performs data analysis in XChemExplorer [115], which interfaces with the laboratory management database used in XChem: SoakDB [111]. XChemExplorer can be used to aid deposition to the PDB [117–119], discussed further in section 1.4.1.

As XChem is now a user facing facility, there have been a wide variety of projects. If a project involves a consistently diffracting crystal system (a pre-requisite) and can tolerate solvent, there is normally a hit rate of 3-20%. Often some of the hits are not in functionally interesting locations, and thus the useful hit rate is much lower. Although there are usually many hits, deciphering which to follow up still requires specialist medicinal chemistry advice. Current work is looking to extend the computational chemistry toolset to allow user to determine appropriate compounds which are feasibly synthesisable. These would undergo docking or short MD simulations to determine binding pose, and there are also developments (private communication, Mihaela Smilova, SGC) to assess the ensemble of fragment binders and the local protein environment and determine the most important bonds. Importantly these features are being developed with visualisation tools and backend frameworks to support data management, and hardware to allow users to trigger these jobs and understand the results in a high throughput manner.

XChem fragment screening still has many opportunities to be improved. Tests on using multiple crystals per spine pin are ongoing (private communication, Alice Douangamath, DLS) to increase throughput at the beamline. Fragment screening with room temperature data collection remains massively underexplored. I was involved with testing room temperature fragment screening, initially in spine pins on the I03 beamline, followed by later tests on I24 with crystals *in-situ* in their crystallisation trays (private correspondence, Daniel Keedy, City University of New York). These developments are likely to be best explored by using *in-situ* X-ray diffraction experiments, on beamlines such as the VMXi beamline at DLS.

Further experiments to test the validity of the high concentrations of compound usually used at XChem would be useful. Although hits from high concentration soaks can be found and progressed hits found at lower concentrations may be less frequent, but would likely be stronger binders, which may be easier to measure binding affinity of in biophysical techniques.

#### 1.2.5.2 Screening libraries

Although fragment sized compounds sample chemical space more efficiently than the higher mass compounds used in HTS, libraries of compounds still must be chosen appropriately. Fragment libraries should have a clear design rationale, typically to sample chemical space efficiently with a high compound diversity [121, 122]. However, libraries should remain synthetically accessible [123] and have a reasonably high hit rate of binders [124]. Fragment screening libraries are typically designed to adhere to the ”rule of three” [64]:

- Molecular weight less than 300Da
- At most three hydrogen bond donors and acceptors
- cLogP $\geq 3$

Where cLogP is the logarithm of partition coefficient between n-octanal and water, it measures lipophilicity.

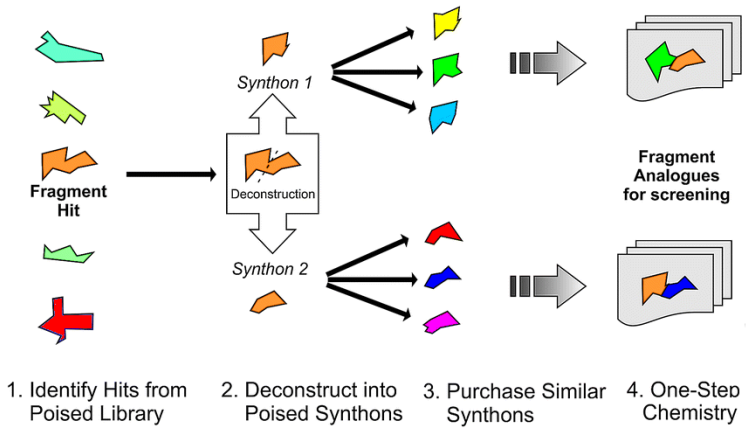



**Figure 1.3:** Poised fragment library is constructed from molecules which can be decomposed into two synthons. These are common building blocks which can be joined by one of the top ten reactions used in medicinal chemistry [129], twelve heterocycle forming reactions [130], or oxazole formation [131, 132]. This relative simplicity of medicinal chemistry should allow for quick follow up with alternate molecules that contain one of the original synthons. Adapted from [123].

Library size is most often limited by the technique used, with libraries usually between 500-20,000 compounds [124, 125]. Compound purity, stability and solubility at the relatively high concentrations used in fragment screening (> 500 μM) are also important factors to design and quality control for in fragment libraries. Libraries should also avoid compounds which are known, or appear after screening multiple targets to be; non-specific binders, reactive covalent modifiers, chelators, or compounds that aggregate [126, 127], as these can show false positives [82]. Filters are applied to remove potentially problematic compounds such as pan-assay interference compounds [128].

XChem has maintained and used a variety of fragment screening libraries. Most targets have been screened against the Diamond Sciences Poised Library (DSPL) [123] or its later iteration (DSiP). This library is based on the ability to decompose the molecules into two synthetically accessible parts (synthons), as shown in figure 1.3. Follow up molecules can be designed by altering either synthon, and choosing compatible reactions form a short-list of highly used medicinal chemistry reactions [129–132]. This potential to follow up was further improved with the introduction of graph network [11] searching of the Enamine REAL database, starting from the synthons. This work is deployed in the Fragaylsis tool used for hypothesis generation for follow ups at XChem (private communication, Rachael Skyner, DLS).

XChem also maintains access to commercially available libraries (Edelris Keymical Fragment ™and the Maybridge rule of three fragment library [106]), as well as libraries designed for synthetic diversity [122].

The average size of compounds in the DSPL library is quite high, and testing libraries with smaller compound sizes for initial exploration of the binding pocket is ongoing (private communication, Alice Douangamath, Romain Talon, Frank von Delft, DLS). These include MiniFrags [133] and Fraglites [134].

### 1.2.6 Introducing compounds to crystals

Soaking is the process by which a compound is introduced to the crystal once it has formed. Introduction of compounds via soaking, is possible as protein crystals typically have large solvent channels, and a solvent content between 30-70% [135–137]. Therefore, the compound must be able to penetrate the crystal lattice. This requires potential binding sites to accessible to the solvent channels in the protein crystal [138]. The compound must also bind in manner where any perturbations to the crystal are tolerated in that crystal form, thereby precluding large conformational changes [139, 140]. Furthermore, disruption of the crystal contacts by the compound could lead to poorer diffraction.

Co-crystallisation is the process where the compound is incubated with the protein prior to crystal growth. Co-crystallisation can yield structures when soaking is not amenable [141]. The incubation of protein and compound is slower than soaking for fragment screening, and if possible, should be reserved for follow up work, such as screening a protein with a fragment binder to identify fragments which could be merged with the original hit. Co-crystallisation can also be useful to find binders which induce large conformational changes on binding which would not be tolerated in an existing crystal system [139]. However, in this case a new crystallisation condition may be required.

### 1.2.7 Developing screening hits into lead compounds

Fragments definitionally are small molecules usually with relatively weak binding affinities. To progress fragment hits into lead compounds, it is desirable to retain the binding pose

and interactions of the fragment molecule [12], and either grow, substitute or merge fragments. Fragment screening methods which give positional information, such as the atomic detail of binding pose given by sufficiently high resolution X-ray crystallography can be used to develop rational design strategies from a set of initial compounds.

One major challenge in developing fragment screening hits is to identify features of the fragment that are key for its binding [142]. Once identified, then one can hypothesise larger chemicals which share a common substructure with the fragment [143], and are likely to maintain the binding pose, whilst adding further chemical moieties to improve potency and selectivity.

The ability to elaborate a fragment relies partly on correctly determining a confident model, which can be difficult if the electron density is weak due to a partly occupied site.

Further discussion of fragment elaboration is beyond scope for this thesis, for a review of recent fragment to lead campaigns see [67].

## 1.3 Macromolecular crystallography

Experimental determination of protein structures has historically been dominated by X-ray crystallography. Currently (2019-09-24) 89% of the 156,101 structures deposited in the PDB [117–119] were determined using X-ray crystallography, with NMR at 8% and cryo-electron microscopy at 2%.

X-ray crystallography at sufficiently high resolution provides atomic detail on the protein structure. A X-ray beam irradiates molecules regularly arranged in a crystalline lattice, leading to a diffraction pattern. The diffraction pattern, alongside phase information can be used to derive electron density maps of the repeated unit cells of the crystal. Phase information can be determined using molecular replacement utilising homologous structures. It can also be found by experimental measurement of anomalous diffraction from heavier atoms found or substituted by soaking or selenomethionine substitution into the crystalline lattice. A model of the protein structure is then built into this electron density map. By iterating between model building and refinement of the model and map, a high quality model of the protein's structure can be produced, if the data are of sufficently high quality.

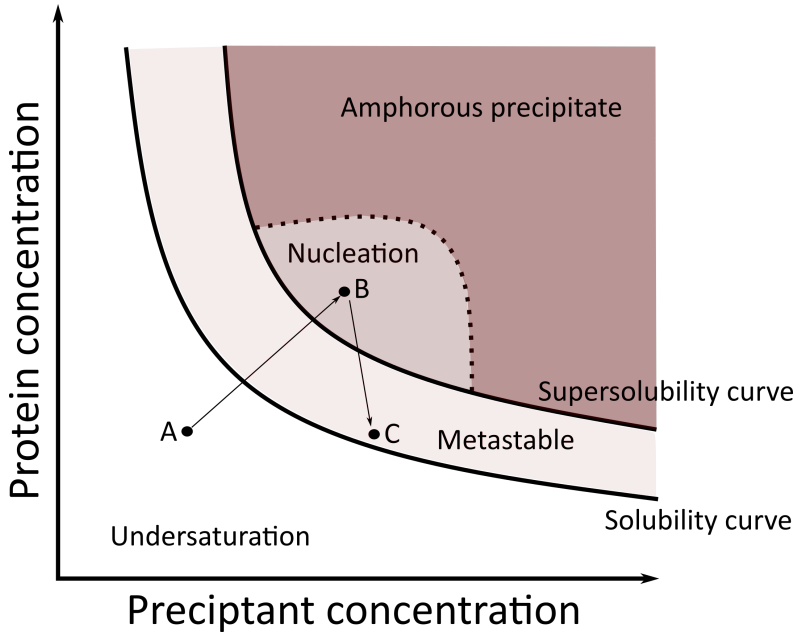



**Figure 1.4:** Phase diagram of crystallisation for proteins. Protein crystallisation is affected by protein and precipitant concentrations. When either or both are low, the system is under-saturated, the protein solution is clear, stable and in a single phase. Above a solubility curve, the system is metastable, the conditions under which a protein may crystallise. Above this the system separates into two phases: precipitant and protein solution. A crystallisation event will often occur via a path similar to A-B-C: the solution begins as under-saturated, vapour diffusion leads to protein and precipitant concentration increasing until a nucleation event occurs, a crystal begins to form, returning to the metastable state, where the crystal can continue to grow.

### 1.3.1 Crystallisation

Macromolecular crystals are formed by repeating lattices of protein molecules formed by weak non-covalent interactions. Obtaining protein crystals of a target protein is typically, laborious iterative process, as crystallisation conditions need to be found that encourage stable self-assembly of the protein molecules. As protein crystals are typically produced from recombinant protein, there are also many challenges obtaining a sufficient amount (typically 0.1-10mg) of high purity protein, which are largely beyond the scope of this thesis.

Screening for crystallisation conditions is usually done in crystallisation microplates in different crystal formation conditions (i.e. vapour diffusion in sitting or hanging drop, microbatch under oil) [144]. Crystallisation is controlled by the precipitant and protein concentration, as shown in figure 1.4. Finding an initial condition is usually done through commercially available coarse screens, which take an informed scatter-gun approach.

This exposes the protein to a wide variety of conditions with precipitants known to induce crystallisation.

At the SGC the established practice is to set-up 4-5 commercially available screens, in 96 well sitting drop vapour diffusion plates, with three sub-wells, at three reservoir to protein ratios (1:2, 1:1, 2:1), and at two temperatures (4°C and 20°C) [145]. This sampling is often sufficient to find a condition that produces some crystalline matter, whether that be crystals or micro-crystalline precipitate. If sufficient, then a finer screen is typically designed to optimise crystal growth based on the conditions at which the initial crystallisation material was found. Variation in pH and precipitant concentration are typically applied in a finer screen.

#### 1.3.1.1 Crystallisation for fragment screening

Crystallisation for fragment screening requires further optimisation, mainly to increase the stability, diffraction consistency and robustness to high concentration compounds which are often solubilised in solvents such as dimethyl sulfoxide (DMSO). Firstly a crystallisation fine screen may lead to a variety of pH and precipitant conditions that produce crystals. If sufficient crystals (>50) with similar sizes, and morphologies are obtainable in an optimisation fine screen, then the diffraction consistency of the system should be tested. If the maximal resolution lies within a tight window (i.e. 1.8-2.1Å) with few outliers, then the target can be progressed. In the case that diffraction is inconsistent, optimisation of other conditions that produced crystalline material should be attempted. At this stage it may also be necessary to explore whether alternate sequence constructs of the protein produce crystals which may diffract with higher consistency [146].

At XChem, fragment screening is usually done by soaking high concentration (100 to 500 mM) compounds solubilised in a solvent. A diffraction test with increasing nominal solvent concentration, typically 5% up to 30%, is done to ascertain whether crystals remain mountable and diffract consistently [108].

### 1.3.2 Limitations of crystallography for determining protein ligand complexes

Although crystallography can give lots of information on the atomic pose of ligands, there are limitations such as a crystalline environment not well representing how protein and compound in solution would interact. Crystal lattices of proteins are held together by a few weak, but specific non-covalent intermolecular interactions [147] known as crystal contacts. Protein crystals are fragile and sensitive to environmental change [148]. If residues involved in the function, or putative ligand binding sites form part of the crystal contact, these residues may be in conformations which are different from the solution state of the protein [147], and thus affect what ligands can bind. However, the binding of ligands can also increase crystal order, improve resolution and stabilise loop conformations [149].

The conditions required to crystallise a protein are likely to be non-representative of the cellular environment that the protein would normally be found. The pH is often strongly biased to trigger crystallisation, as well as the addition of high concentration precipitants that would not be present in a cellular context. These additions allow crystallisation, but can significantly affect the likelihood of compound binding.

Most crystallographic datasets are collected at cryogenic temperatures to reduce the effect of radiation damage [150]. This alters which conformations that the protein can take [151, 152] and thereby potential alters the ligands that can bind [153], and potentially alters the functional response of those ligands. Thermodynamics should favour ligand binding at lower temperatures [153], so at high but non-saturating conditions a greater proportion of the ligand sites would be occupied. However, the heterogeneity observed between different cryogenic datasets appears often to dominate. Transient cryptic sites, which are only revealed on ligand binding) are sometimes only detectable at room temperature [153].

Many fragment sized ligands are only sufficiently soluble in organic solvents to achieve a high enough concentration to have a chance of observing the compounds when soaked onto a crystal. Crystals often do not tolerate high solvent concentrations, however this can be improved by targeting small droplets of compound far from the crystal in the

crystallisation drop and allowing diffusion to introduce the compound to the crystal, reducing shock from an initially high solvent concentration [108]. Solvent tolerance tests used before XChem fragment screening show that some protein crystals are highly susceptible, whilst other can tolerate 30% v/v. The high resolution limit may be affected by solvent addition.

### 1.3.3 Data collection

A macromolecular X-ray diffraction experiment most commonly involves the rotation of a single crystal in an intense X-ray beam. Typically, a series of diffraction images are collected on a detector as the crystal is rotated (typically 0.1-1.5° per image) in the X-ray beam using a sample goniometer.

Most samples are flash cooled in liquid nitrogen when mounted, ideally in an appropriately sized mounting loop. The crystal sample is maintained at cryogenic temperatures to reduce the effect of X-ray dose on the crystal, as accumulated dose causes both specific and global damage to the crystal [150, 154]. The specific damage such as breakage of disulphide bonds, and decarboxylation of acidic groups can be very important to understand the local environment of protein-ligand complexes [155–157].

Cryo-cooling can affect the distributions of conformations observable in the crystal [151]. It can also impart an artificially high degree of order for the water molecules than in structures obtained at room temperature [158]. The rearrangement of networks of ordered waters allows the protein to adopt different conformations [159]. The cryo-cooling of a crystal can affect the ordering of water molecules, thus the degree of conformational flexibility and potentially the ligands which can bind.

The unit cell of a crystal is defined by the arrangement of atoms in the macromolecular assembly, and the shape and symmetry of that cell define the direction and intensity of diffracted beams. Therefore, a protein structure determined by X-ray crystallography will represent the whole molecule(s) present in the lattice, however some parts may remain too disordered to be well modelled.

**Table 1.1:** Crystal systems and their respective symmetry relationships and constraints on the unit cell.

| Crystal System | Minimum symmetry | Constraints on unit cell |
|---|---|---|
| Triclinic | None | None |
| Monoclinic | One 2-fold (along $b$) | $\alpha = \gamma = 90$ |
| Orthorhombic | Three 2-folds (along $a, b, c$) | $\alpha = \beta = \gamma = 90$ |
| Trigonal | 3-fold (along $c$) | $a = b$; $\alpha = \beta = 90$; $\gamma = 120$ |
| Tetragonal | 4-fold (along $c$) | $a = b$; $\alpha = \beta = \gamma = 90$ |
| Hexagonal | 6-fold (along $c$) | $a = b$; $\alpha = \beta = 90$; $\gamma = 120$ |
| Cubic | Four 3-fold (along body diagonal) | $a = b = c$; $\alpha = \beta = \gamma = 90$ |

#### 1.3.3.1 Crystal lattices and space groups

A lattice is an infinite array of points arranged, such that the local environment of any point is identical to the environment of any other point. A unit cell is repeating unit placed on the lattice points which fills space. In three dimensions there are three distances $a, b, c$ and three inter-axial angles $\alpha, \beta, \gamma$ which define a unit cell that forms the repeating unit on the crystal lattice. By convention the unit cell is chosen to have the smallest possible volume, and $\alpha, \beta, \gamma$ are chosen to be obtuse, but as close to 90°as possible.

The allowed rotational symmetries within a normal crystal lattice are 2,3,4 and 6 fold. Other symmetry operations are reflection and inversion. The combination of rotation, inversion and reflection are known as a point group symmetry, as they leave the position of one point unchanged on the transformed object. However, as macromolecules are chiral molecules, reflections and inversions are not valid symmetry options. Rotational symmetry gives seven crystal systems, defined by the symmetry and corresponding limits on the unit cell, as shown in table 1.1.

Lattices in which points lie only at vertices are primitive lattices (P). Unit cells can also be defined such there are lattice points at the centre of one face (C), all faces (F) or the body of the unit cell (I). This allows unit cell axes to remain parallel to symmetry axes. Trigonal crystals can have a rhombohedral cell (R), these can equivalently be treated as hexagonal cells. This leads to 14 Bravais lattices.

In addition to the operations of the point group, the space group defines which translational symmetry operations are applied. There are 65 space groups possible for chiral molecules.

#### 1.3.3.2 Screening and data collection

Data collection strategies for individual crystals, which will maximize the obtainable information whilst reducing the X-ray dose, are usually obtained by assessing a small number of diffraction images [160, Chapter 7] and [161]. These screening images can be used for indexing; determining a set of reciprocal basis vectors that describe the crystal symmetry, space group and unit cell parameters.

When the structure of the protein is known, and only small perturbations are to be observed, such as in crystallographic fragment screening, an existing data collection strategy can be applied. Applying a single strategy to many crystals increases the throughput of the method [111]. In cases where ligand binding can have a large effect on unit cell parameters, it may be appropriate to undertake diffraction screening to determine a data collection strategy. These are uncommon in XChem fragment screening, and such systems would be identified within the pre-screen. The pre-screen is an assessment of the first approximately one hundred compounds to determine any experimental issues or crystallographic pathologies, such as changes in diffraction quality, high rates of precipitation or changes in crystal form. Currently in XChem crystal form changes are dealt with by splitting the analysis by parameters such as the space group or unit cell parameters after data collection (private correspondence, Daniel Pinkas, SGC). If such cases involve structural rearrangement, it is appropriate to proceed with two different ground state models. Further work is needed to support polymorphism induced by ligand binding [162] in XChem screening.

The set of diffraction images collected by rotating a crystal in the X-ray beam, forms the raw data of an X-ray crystallography experiment. Data reduction (indexing, integration, scaling and merging) is then applied. As we use automated pipelines to undertake data reduction we do not discuss details herein.

### 1.3.4 Diffraction data

Diffraction data are obtained fromthe intensities of diffraction spots as recorded on a detector. In most single crystal experiments a crystal is rotated in the X-ray beam. As the crystal rotates, so does the reciprocal lattice. The reciprocal lattice points lie in planes, as

each plane intersects the Ewald Sphere, this results in spots projected onto the detector in ellipses. The diffraction data are collected in a series of images, each representing a small rotation range of the crystal. As the crystal is rotated through a small angle each circle is broadened into a lune, all spots within a lune belong to a plane of the reciprocal lattice. Dependent on the angle range probed a spot may be fully contained within an image, be partially split across images or split across multiple images (fine slicing). With the low background signal on modern detectors, and sufficient indexing algorithms, fine slicing the angle range to sample diffraction spots across multiple images is usually desirable, as the overall signal to noise is improved [163].

Other experiments such as serial synchrotron crystallography, and X-ray free electron lasers, collect many fewer, or single images per crystal, these need to merged together across many crystals to produce sufficiently complete data to solve the underlying macromolecular structure.

Indexing is the process of identifying the diffraction spots and using them to estimate unit cell dimensions, crystal orientation and a first estimate of which Bravais lattice the diffraction pattern corresponds to.

Integration is the process by which the counts of each pixel forming a spot are added together. However, as each image has both signal and background, care must be taken to identify the level of background, some of which may be systematic, i.e. masking dead, hot or cold pixels on the detector, and some are simply random due to detector noise. Weak spots are those which occupy fewer pixels, as the shoulders of the spot peak intensity are below background.

The data measured by integration programs are not a common scale, and this needs to be taken in account. This can be due to changes in incident radiation intensity, illuminated volume of crystal, radiation damage to the crystal and non-uniform detector response. Scaling attempts to make symmetry related and duplicate measurements of a reflection equal by modelling the diffraction experiment as a function of the incident and diffraction beam directions in the crystal. Scaling aims to make data internally consistent, by minimising differences between individual observations $I$ and the weighted mean of all symmetry related equivalents of the reflection $I$ [164].

Once the observations on a common scale the individual parts of recorded reflections are merged into complete reflections. Furthermore, symmetry related reflections are merged into unique observations.

#### 1.3.4.1 Scattering from protein structures

The scattering of X-rays from a single molecule would be too weak to record any signal. The periodicity of a crystal lattice amplifies the molecular scattering function. The total scattering from a crystal in any given reciprocal lattice direction is proportional to the sum of all scattering contributions in the unit cell.

The scattering function in a given direction $\mathbf{h}$, from a crystal irradiated with X-rays, is given by the complex structure factor $\mathbf{F_h}$. The structure factor is a summation of the scattering from every atom in the unit cell:

$$\mathbf{F_h} = \sum_{j=1}^{\text{atoms}} f^0_{\mathbf{S},j} \exp(2\pi \mathbf{h} \mathbf{x}_j) \tag{1.5}$$

where

- $j$ represents the atoms over which the summation is taken.
- $\mathbf{S}$ is the scattering vector; the difference between the incoming and outgoing scattered wave for the atom $j$.
- $f^0_{\mathbf{S},j}$ is the wavelength independent atomic scattering factor representing the scattering from the atom $j$.
- The exponent term represents the relative phase angle for the partial wave emanating from each atom, which depends solely on the direction of the scattering $h$ and the position of the atoms $j$ in the unit cell given by the fractional coordinate vector $\mathbf{x}_j$.

The atomic scattering factor can be approximated as a Gaussian function (centred on $sin\theta = 0$), and decaying rapidly with diffraction angle. The atomic scattering factor represents the Fourier transform of the atoms' real space electron density $\rho(\mathbf{r}_j)$. The summation implies every atom in the unit cell contributes to each reflection [41].

### 1.3.5 B factor

Atoms in a crystal lattice will vibrate around a mean position, dependent on their temperature. A phase difference results from this motion. This phase difference can be modelled as an attenuation to the atoms structure factor. Motion due to thermal fluctuation of the atom is not the only reason for displacement. Crystal heterogeneity, which can arise from crystal mosaicity, domain motion of relative to the coordinate frame, local disorder and the presence of multiple conformations of the protein, also results in displacement of the atoms. Thermal fluctuation and crystal heterogeneity can together be represented by an additional Gaussian, wavelength dependent term in the structure factor equation, called the Debye-Waller factor $T$:

$$T_s = \exp\left(\frac{-B_{iso}|\mathbf{S}|^2}{4}\right) = \exp\left(-B_{iso}\left(\frac{\sin\theta}{\lambda}\right)^2\right) \tag{1.6}$$

where $B_{iso}$ is the isotropic displacement parameter, or B-factor which can be expressed in terms of the mean square isotropic displacement $\langle u_{iso}^2 \rangle$ of the atom from its mean position.

$$B_{iso} = 8\pi^2 \langle u_{iso}^2 \rangle \tag{1.7}$$

The atomic scattering function, $f_{\mathbf{S},j}^0$ as in the structure factor equation 1.5, is attenuated by:

$$f_S^B = f_S^0 \cdot T_S = f_S^0 \exp\left(-B_{\text{iso}}\left(\frac{\sin\theta}{\lambda}\right)^2\right) \tag{1.8}$$

Note that as B-factor describes the attenuation of scattering, the scattering contribution beyond 2.5 Å (i.e. at low resolution) for an atom with B factor 100 Å$^2$ is essentially zero. This implies non-rigid regions at lower resolutions cannot diffract well, and will result in missing electron density.

The B factor should ideally represent the atomic oscillation, which increases with temperature, around an equilibrium position, as well as displacement due to crystal heterogeneity. However, it typically represents the overall uncertainty in atomic position [165]. It can mask certain experimental errors, or changes in structure over the time

in which images are collected (i.e. the increased B factor at sulphur atoms involved in breaking of disulphide bonds due to radiation damage [166]).

The choice of most appropriate B-factor model describing positional uncertainty of atoms, depends on the resolution of the data. The isotropic B factor model uses a single parameter per atom, representing a spherical uncertainty in position about the atoms coordinates. This gives a total of 4 parameters per atom. The anisotropic model considers each atomic position expressed as a thermal ellipsoid, which requires 6 parameters per atom, and therefore is usually only appropriate at resolutions better than 1.4Å [167].

At intermediate resolutions the anisotropic translation, liberation, screw (TLS) model allows assigning anisotropic TLS parameters to groups of atoms, that are thus treated as rigid bodies [168]. As these parameters are applied to groups of atoms, they introduce less parameters than individual atomic isotropic B factors. Assigning TLS groups can be algorithmically determined [169], or applied with knowledge of the biomolecule to regions such as protein domains.

In this thesis we commonly refer to B factors, which unless otherwise stated we are referring to the individual atom isotropic B factor, as this model is usually the most appropriate at the resolution of the data studied.

### 1.3.6 Wilson scaling

To be able to understand whether a structural model is appropriate for the experimental data, first the calculated structure factors from the model and those from the data, are brought onto a common scale. This requires an overall linear scale factor and a correction for the overall B factor attenuation for each specific dataset and crystal. This first approximate scaling is performed through isotropic B-factor scaling, known as Wilson scaling [170]. Here we follow the description from [41].

Although each atom is subject to individual displacements, it is possible to factor out an overall isotropic B-factor:

$$f_j^B = f_j^0 \exp\left(-B_{\text{iso}}(\sin\theta/\lambda)^2\right) \tag{1.9}$$

Given that we know for a structure factor on an absolute scale:

$$\sum_{j=1}^{\text{atoms}} (f_j^0)^2 = \langle I_{\text{abs}} \rangle \tag{1.10}$$

Thus, by squaring 1.11:

$$(f_j^B)^2 = (f_j^0)^2 \exp\left(-2B_{\text{iso}}(\sin\theta/\lambda)^2\right) \tag{1.11}$$

we can write for the sum of all scattering atoms:

$$\langle I_{\text{obs}} \rangle = \langle I_{\text{abs}} \rangle \cdot k \cdot \exp\left(-2B_{\text{iso}}(\sin\theta/\lambda)^2\right) \tag{1.12}$$

and by taking logarithms:

$$\ln \frac{\langle I_{\text{obs}} \rangle}{\langle I_{\text{abs}} \rangle} = \ln k - 2B_{\text{iso}}(\sin\theta/\lambda)^2 \tag{1.13}$$

The Wilson plot, $\ln \frac{\langle I_{\text{obs}} \rangle}{\langle I_{\text{abs}} \rangle}$ versus $(\sin\theta/\lambda)^2$ has a slope which corresponds to the overall B-factor, and an intercept which corresponds to the scale factor $k$. These parameters are estimated by an optimisation procedure. The mean values of observed intensities are normally computed in bins of:

$$(\sin\theta/\lambda)^2 = \frac{1}{(2d)^2} = \frac{S^2}{4} \tag{1.14}$$

These bins equate to shells of reciprocal space, and contain an increasing number of reflections in each bin. For the observed intensities this can be done by sorting reflections by $(\sin\theta/\lambda)^2$ an compute the average intensity over that bin.

To estimate the absolute intensities for the model in each bin $\langle I_{\text{calc}} \rangle_{\text{bin}}$, as the B factors and atom positions are unknown at this point, the unit cell contents are estimated. From the residues in the asymmetric unit one can determine the number of atoms, and by using the Cromer-Mann approximation, which approximates the atomic scattering factor for each atom (C,H,N,O *etc*), one can determine the expected average intensity. This estimate does not consider the B factors of the atoms. Both the scale factor $k$ and the average overall B factor are later refined as model parameters.

### 1.3.7 Phase problem in crystallography

To produce an electron density map (by Fourier transform), the required primary output of a macromolecular crystallography experiment, it is necessary to obtain or estimate the phase information. Initial phase information can be obtained; by direct methods [171, 172], by molecular replacement [120, 173] or by experimental phasing techniques [171, 174, 175].

Direct methods use probabilistic relationships between structure factors of certain groups of reflections to estimate phases, which can then be iteratively improved. This requires diffraction data with a resolution of at least 1.2Å [176, 177], and is used heavily in small molecule crystallography. However, as these probabilistic relationships are inversely proportional to the square root of the number of atoms it is not possible for the vast majority of macromolecular structures [167].

Molecular replacement [173], utilises the similarity of proteins with high sequence identity and presumably structural homology. A model of a previously determined homologous protein is used to generate inter-atomic vectors. These are matched against the peaks of the Fourier transform of the reflection intensities, known as the Patterson function [178]. This elucidates the orientation of the model in the unit cell. The optimal translation function can be estimated, by a grid search, at each position the structure factors for the search model are calculated [120]. Maximum likelihood methods have been applied successfully to molecular replacement [179, 180]. Molecular replacement can be framed as a hypothesis testing problem, where different hypotheses about the orientation, position and quality of the search model are tested against the data. Such a problem can thus use maximum likelihood methods.

Molecular replacement continues to grow in popularity and efficacy as the number of structure in the PDB [117–119] grows, the likelihood of a suitable homologous structure for molecular replacement being available also grows [181]. In difficult cases, where a protein fold may not be observed in the PDB [117–119] it is possible to use protein structure prediction either with *de novo* [181] or homology models [182] as the basis for molecular replacement solution. Furthermore, a brute force search of all possible

models for molecular replacement [181], or searches using small portions of known structures [183] may be successful in determining initial phases.

Experimental phasing techniques are often needed to solve novel macromolecular crystal structures. This requires accurate collection of diffraction data from anomalous scattering. The atomic scattering factor:

$$f(\theta, \lambda) = f_0(\theta) + f'(\lambda) + if''(\lambda) \tag{1.15}$$

where the normal scattering term $f_0$ is dependent on the Bragg angle, and there are two terms dependent on the wavelength. These terms represent the anomalous scattering that occurs at an element's absorption edge when the X-ray photon energy is sufficient to promote an electron from its inner shell [184]. Anomalous scattering is dependent on the scatterer with heavier atoms responsible for a greater anomalous scatter due to the larger number of electrons. The difference between reflections for anomalous scatterers can then be used, by direct or Patterson methods to locate the heavy atom substructure, which provides an initial phase estimate. This phase estimate can then be improved by techniques such as density modification [185, 186]. Phasing methods include single wavelength anomalous diffraction [174] and multiple wavelength anomalous diffraction [175]. These methods utilise heavier elements as anomalous scatterers such as sulphur containing residues methionine and cysteine [187, 188] or introduce other elements by soaking or protein engineering (i.e. selenomethionine substitution [189, 190]).

### 1.3.8 Data quality metrics

After scaling and merging, the measures of internal consistency reported for a dataset are calculated. As the intensity of reflections becomes weaker with increasing resolution, we represent data binned by resolution shell, each with an equal number of reflections. The signal to noise ratio can be represented as an average summed over all reflections $N$ in a resolution shell:

$$\left\langle \frac{I}{\sigma(I)} \right\rangle = \frac{1}{N} \sum_{\mathbf{h}}^{N} \frac{|I_{(\mathbf{h})}|}{\sigma(I_{(\mathbf{h})})} \tag{1.16}$$

During data collection it is desirable to collect a dataset with a high completeness, that is a high percentage of the reflections that are possible to collect for the crystal's

space group and unit cell are collected. It is also desirable to collect multiple redundant observations of the same reflection (multiplicity). The spread of individual intensities of all symmetry equivalent reflections, contributing to the same unique reflection is gauged by the residual $R_{\text{merge}}$:

$$R_{\text{merge}} = \frac{\sum_{\mathbf{h}} \sum_{i=1}^{N} |I_{(\mathbf{h})i} - \langle I_{(\mathbf{h})} \rangle|}{\sum_{\mathbf{h}} \sum_{i=1}^{N} I_{(\mathbf{h})i}} \tag{1.17}$$

where $\mathbf{h}$ represents the unique reflections, and the summation over $N$ redundant observation for a given reflection $h$. Note that $R_{\text{merge}}$ can be calculated for identical but unmerged reflections known as $R_{\text{int}}$, or for symmetrically equivalent reflections known as $R_{\text{sym}}$ and can be calculated either globally or over resolution bins.

However, higher multiplicity leads to an increased $R_{\text{merge}}$ [191], hence the use of:

$$R_{\text{meas}} = \frac{\sum_{\mathbf{h}} \left(\frac{N}{N-1}\right)^{\frac{1}{2}} \sum_{i=1}^{N} |I_{(\mathbf{h})i} - \langle I_{(\mathbf{h})} \rangle|}{\sum_{\mathbf{h}} \sum_{i=1}^{N} I_{(\mathbf{h})i}} \tag{1.18}$$

A high quality dataset should be characterised by a $R_{\text{meas}}$ of 4-5%. Historically the somewhat arbitrary $\langle I/\sigma(I) \rangle = 2.0$ in the highest resolution shell, was used as a resolution cut off, but as weaker high resolution data could improve the refined model, this is subject to discussion. However, a value of $\langle I\sigma(I) \rangle > 2.0$ in the outermost shell, indicates that higher resolution data could have been collected, which is the case in 48% of structures deposited in the PDB [167].

### 1.3.9 Electron density maps in crystallography

An initial set of phases and amplitudes corresponding to observed reflection intensities are used to generate an electron density map, the primary result of a diffraction experiment. A model of the macromolecule acts to explain, and to iteratively improve the electron density but should always be taken in the context of the electron density map.

Choosing the most appropriate representation of electron density is a non-trivial task. When interpreting a partially complete model, a crystallographer would like to be able to identify contiguous electron density from known atoms to the positions of unmodelled atoms evidenced by the diffraction experiment. Furthermore, the crystallographer

should be able to identify any wrongly positioned atoms, and have the lowest number of background peaks possible.

Initial Fourier syntheses for electron density maps included the $\beta$, [192], the $\gamma'$ and the $2F_{obs} - F_{calc}$ synthesis. It was shown that the $\beta$ synthesis contained more background peaks than the $\gamma'$ synthesis and the $2F_{obs} - F_{calc}$ synthesis [193]. The $2F_{obs} - F_{calc}$ synthesis better suppresses peaks of wrongly positioned atoms when compared to the other syntheses [193]. . It was concluded that the $2F_{obs} - F_{calc}$ synthesis should be used for non-centrosymmteric structures, such as the majority of protein structures (centrosymmetric racemic structures [194] excluded).

Note that partly refined models of macromolecules are generally incomplete and have large coordinate errors. Electron density maps calculated using partial structure phases are biased towards the partial structure.

To determine appropriate Fourier coefficients when there are coordinate errors (i.e. in all non-simulated structure solutions) two correlation coefficients are compared; the correlation to the correct map (the Fourier transform of $F_{\rm obs}$) and the correlation to the model map (the Fourier transform of $F_{\rm calc}$). The correlation to the correct map is lowered by any coefficients designed to compensate for model bias. However, model bias in any electron density map makes it difficult to correct errors in the model. Therefore, coefficients that reduce model bias lead to maps that are subjectively improved. Optimal map coefficients will reduce model bias at only a small cost in correlation with the correct map.

A model biased map will have a higher correlation with the model map than that of the correct map. Map coefficients which reduce model bias should lower the correlation to the model map to some extent. A correlation to the model map lower than that between correct and model maps would indicate that correct features of the model were being eliminated. It was shown [195] the most appropriate coefficients are:

$$2mF_{\rm obs} - DF_{\rm calc} \tag{1.19}$$

where $m$ is a figure of merit representing experimental uncertainty, and $D$ [196] represents model error, see [195] for a full derivation.

$$D = \langle \cos(2\pi \mathbf{s} \cdot \delta \mathbf{r}) \rangle \tag{1.20}$$

where $\mathbf{s}$ is the reciprocal lattice vector and $\delta\mathbf{r}$ is the expected error in the atomic coordinates of the model.

$$m = \langle \cos(\alpha_N - \alpha_P^c) \rangle \tag{1.21}$$

where $\alpha_N$ is the phase of the complete structure, and $\alpha_P^c$ is the phase of the known atoms

Sigma scaling is usually used to display normalised electron density maps. This scales electron density in real space, by setting grid point values as a number of standard deviations away from the mean (across all values in the unit cell). Sigma scaling is dependent on solvent content and map quality.

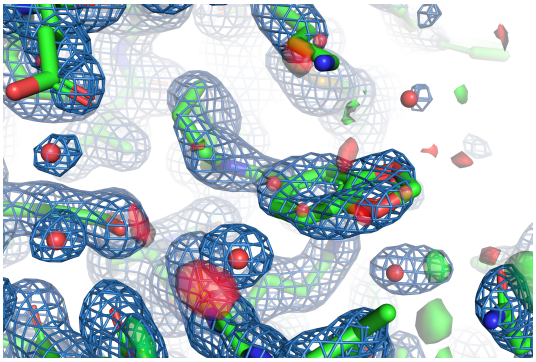
$$\rho_\sigma(x) = \frac{\rho(x)}{rms(\rho(x))} \tag{1.22}$$

The $2mF_{\text{obs}} - DF_{\text{calc}}$ map is most commonly displayed at $1\sigma$ contour level, usually as a blue mesh. The difference map $mF_{\text{obs}} - DF_{\text{calc}}$ is used , and usually displayed at positive and negative contour levels (green and red meshes respectively), typically at $\pm 3\sigma$. Figure 1.5 shows typically contouring. The contour levels are chosen to permit visualisation, whilst reducing noisy density that appears at lower contour levels. Areas outside $\pm 3\sigma$ in the difference map represent significant differences between the model and observed data, which ideally should be corrected by the crystallographer.

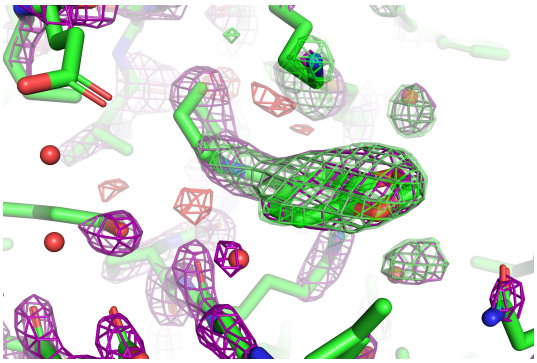
Ideally electron density maps would be scaled such that they could be displayed in absolute units $e/\text{Å}^3$. However, it is impossible to experimentally measure the $F_{000}$ reflection, as it is coincident with the unscattered X-ray beam, which would allow this absolute scaling. It is possible to estimate the $F_{000}$ reflection and use this for scaling electron density [197]. This used to compare electron density between datasets, as in chapter 6. Lower contoured density can be interpreted by the modelling of multiple models, as discussed in section 1.5.

### 1.3.10 Detecting ligands in crystals

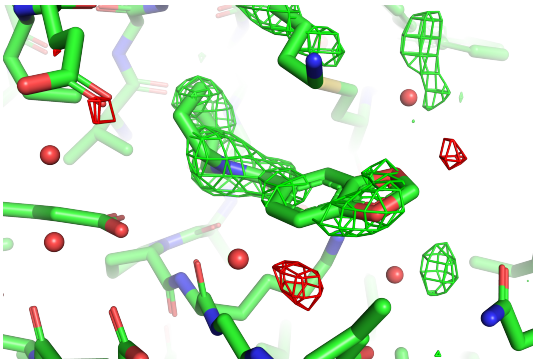
Crystallographic maps are systematically model biased, because model information is used to calculate them. Therefore, if atoms are added to a map, such as when fitting a ligand to an existing map these atoms will have density associated with them, even if there



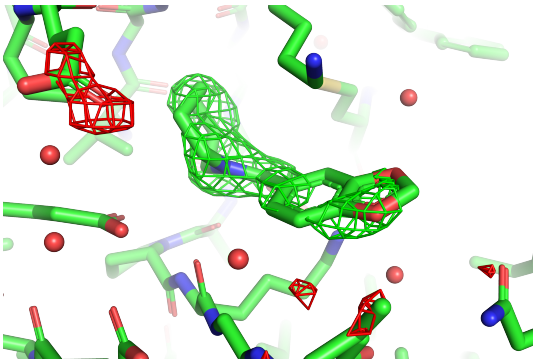
**(a)** $2mF_{observed} - DF_{calc}$ contoured at $1\sigma$ in blue. $F_{\text{obs}} - F_{\text{calc}}$ difference map contoured at $\pm 3\sigma$ in green (positive) and red (negative)



**(b)** Event map contoured at twice the BDC level in purple. Z map countered at $\pm 4\sigma$ in green (positive) and red (negative)



**(c)** Omit map contoured at $\pm 3\sigma$ in green (positive) and red (negative)



**(d)** Polder map contoured at $\pm 3\sigma$ in green (positive) and red (negative)

**Figure 1.5:** Different map types show different ligand detectability.

is no evidence in the observed data. An omit map is an electron density map in which the electron density values in a particular volume are calculated from observed diffraction amplitudes and phases derived from an existing electron density distribution modified so that the electron density values in that particular volume and its neighbourhood are constant [198]. Typically, the area omitted will be an area where model bias may be particularly problematic, such as around the atoms of a newly modelled ligand. In an omit map where the area around a putative ligand has been omitted, the map would show a positive difference density for the ligand, if there was sufficient evidence in the non-model biased electron density.

In an omit map the omitted atoms will cause empty space in the structure, which will be modelled by the bulk solvent model. As this implies constant electron density at any voxel, it may hide density features. Polder omit maps prevent the bulk solvent

model being applied to the omitted region and can better for detecting weak binders [199]. Figure 1.5 shows omit and polder maps for a ligand found during XChem screening. An event map is also shown, which shows stronger evidence for the ligand than the polder map, showing the benefit of contrast maximisation against the ground state of the protein, as discussed in section 1.4.

### 1.3.11 Electron density & model metrics

Quality indicators that gauge the agreement between the model and the observed data, both locally and globally determine whether the model is sufficiently complete and correct, or if further iterative steps of model building and refinement are necessary.

#### 1.3.11.1 Local metrics

The real space correlation coefficient (RSCC) is the correlation between the electron density calculated from the diffraction data (with phase information, from experimental phasing or molecular replacement), $\rho_{\mathrm{obs}}$, to that of the electron density calculated from the atomic model $\rho_{\mathrm{calc}}$:

$$\mathrm{RSCC} = \frac{\mathrm{Cov}(\rho_{\mathrm{obs}}, \rho_{\mathrm{calc}})}{[\mathrm{Var}(\rho_{\mathrm{obs}})\,\mathrm{Var}(\rho_{\mathrm{calc}})]^{\frac{1}{2}}} = \frac{\sum|\rho_{\mathrm{obs}} - \langle\rho_{\mathrm{obs}}\rangle| \times \sum|\rho_{\mathrm{calc}} - \langle\rho_{\mathrm{calc}}\rangle|}{(\sum|\rho_{\mathrm{obs}} - \langle\rho_{\mathrm{obs}}\rangle|^2 \times \sum|\rho_{\mathrm{calc}} - \langle\rho_{\mathrm{calc}}\rangle|^2)^{\frac{1}{2}}} \quad (1.23)$$

where the summation is over all grid points of a map encompassed by the atoms of interest, usually an individual residue or ligand. As a correlation measure, RSCC varies between 0 and 1, values between 0.9 and 1.0 are considered as representing ligands which fit the density [147], corresponding to 67% of protein-ligand complexes in the PDB [117–119]. Values between 0.8 and 0.9 described ligands that partially fit the density, with ligands between 0.7 and 0.8 having significant proportions of the ligand not fitting the density. Values below 0.7 are unlikely to fit the ligand, care should be taken to assess the ligand in these cases, with other evidence required to justify the ligand placement [147].

The real space residual (RSR) corresponds to a normalised difference between the observed and calculated electron density, and is calculated locally over a grid, for a group of atoms.

$$\text{RSR} = \frac{\sum|\rho_{\text{obs}} - \rho_{\text{calc}}|}{\sum|\rho_{\text{obs}} + \rho_{\text{calc}}|} \tag{1.24}$$

The range of RSR is 0 (good) to 1 (bad). Note that RSR can be calculated either with a fixed peak profile for all atoms [200] or with peak profiles obtained from a Fourier transform of the calculated structure factors. Also note that as the limiting radii of atoms used in calculation of RSR varies between programs used, it can vary when generated using alternative software [201].

The real space z-observed score (RSZO) [201] measures precision of the density. It is a signal to noise measure calculated by taking the un-weighted average density over the atoms of interest, and dividing by the error in the density.

$$\text{RSZO} = \frac{mean(\rho_{\text{obs}})}{\sigma(\rho_{\text{obs}})} \approxeq \frac{mean(\rho_{\text{obs}})}{\sigma(\Delta\rho)} \tag{1.25}$$

where the error in observed density, $\sigma(\rho_{\text{obs}})$, is approximately the same as the error in the difference density.

#### 1.3.11.1.1 Real space Z difference score (RSZD)

The real space Z difference score (RSZD) [201] is a measure of the significance of difference density in a region. As the difference density is a measure of how accurate the model is, RSZD is a measure of local accuracy of the model. Note that both RSCC and RSR are both strongly correlated with B factor, but RSZD is not [201]. Scores of less than 3 are considered acceptable [202]. This derivation follows that given in [201, 203].

As the difference density across the unit cell is approximately normally distributed, Z scores can be calculated:

$$Z_{\Delta\rho}(x) = \frac{\Delta\rho(x)}{\sigma(\Delta\rho)} \tag{1.26}$$

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 is 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 sensitive to smaller changes, hence it is a local metric of accuracy.

The uncertainty of difference density due to random errors in observed data, and phases $\sigma(\Delta\rho)$, is presumed constant over the unit cell. This can be estimated from the bulk solvent regions of the crystal, as error here represents random noise, presuming the bulk solvent model is an adequate description of the region. The the slope of the central portion of a quantile-quantile plot of bulk solvent difference density versus the theoretical normal distribution quantile, is used to estimate $\sigma(\Delta\rho)$.

For any group of atoms, the difference density is sampled on a grid within a radius of the atoms. Grid points need to be spaced at $d_{min}/2$, where $d_{min}$ is the resolution limit of the dataset. Any shorter sampling will result in correlated values, and larger sampling may miss difference density features.

This can be tackled by oversampling the density, then down-sampling the ordered list by the same amount. This gets a representative sampling of the density distribution, whilst keeping the most extreme values. These sampled densities are then expressed as Z scores using equation 1.26.

Both positive and negative difference density are considered equally by taking the absolute value of difference density. Absolute Z scores are assumed independent and half-normally distributed. The joint probability distribution of the absolute values of the Z scores is given by

$$P(x_1...,x_i,...x_n) = \prod_{i=1}^{n} 2\psi(x_i) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left(-\sum_{i=1}^{n}\frac{x_i^2}{2}\right) = \left(\frac{2}{\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\chi^2}{2}\right) \tag{1.27}$$

where the i$^{th}$ Z score is denoted by $x_i$ and the probability distribution of $x_i$ is given by $\psi(x_i)$, the normal distribution, which in the case of Z scores is normalised such that $\psi \sim \mathcal{N}(0,1)$.

A chi-squared test can be applied to observed values to obtain a p-value for observing such a set of Z scores. The cumulative probability, $p$ is calculated as:

$$p\left(\chi^2 \leq \sum_{i=1}^{n} x_i^2\right) = P\left(\sum_{i=1}^{n} x_i^2; \frac{n}{2}\right) \tag{1.28}$$

where $P(...)$ is lower regularised gamma function. The corresponding p-value is 1-p. This p-value measures the significance of Z scores, assuming all values are significant.

However, not all difference density values are significant: i.e. If half the residue is correctly modelled, but the other half is not. To account for values of difference density which are not significant, the equation 1.28 is adapted to instead look at the $(n-k)$ largest absolute difference density points from $k=i$ to $n$, ordered in increasing size. By considering multiple alternate null hypotheses, that the $n-k$ largest values may be significant for all possible values of $k$, the relevant p value corresponds to where the probability is maximised for $k$:

$$P_{max} = max_k \left( p \left[ \chi_k^2 \leq \sum_{i=k}^{n} x(i)^2 \right] \right) \tag{1.29}$$

The regularised gamma function is no longer appropriate. As this function is difficult to calculate due as it is high dimensionality (it should be calculated for all $n-k$ possibilities), an alternative expression is used:

$$p_{max} = max_k \left( P \left( \sum_{i=k}^{n} x_{(i)}^2; \frac{n-k-1}{2} \right) \right) I \left( 2\Phi \left[ x_{(k)} \right] - 1; k-1; n-(k-1) \right) \tag{1.30}$$

where $I$ is the regularised incomplete $\beta$ function. The $I$ term is the multiple testing correction which compensates for bias introduced by the sub-selection of the difference density values. $\phi \left[ x_{(k)} \right]$ is the cumulative probability distribution of $\psi$ 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.

#### 1.3.11.2 Global metrics

The overall fit between the diffraction data and the model is quantified by a global linear residual:

$$R = \frac{\sum_{\mathbf{h}} |F_{\text{obs}} - F_{\text{calc}}|}{\sum_{\mathbf{h}} F_{\text{obs}}} \tag{1.31}$$

where the $\sum_{\mathbf{h}}$ a sumamtion over all reflections. As refinement can be subject to over-fitting, due to the low data to parameter ratio a cross validation subset of reflections (often $\sim$ 5%) is reserved from the refinement calculation, these are referred to as free reflections [204]. Residuals that deal with the working and free reflections respectively are:

$$R_{\text{work}} = \frac{\sum_{\mathbf{h} \notin \text{free}} |F_{\text{obs}} - F_{\text{calc}}|}{\sum_{\mathbf{h} \notin \text{free}} F_{\text{obs}}} \tag{1.32}$$

$$R_{\text{free}} = \frac{\sum_{\mathbf{h} \in \text{free}} |F_{\text{obs}} - F_{\text{calc}}|}{\sum_{\mathbf{h} \in \text{free}} F_{\text{obs}}} \tag{1.33}$$

A gap of < 0.05 is considered acceptable in a refined model, above this the current model is an over fit to experimental data. For macromolecular structures determined by X-ray crystallography the minimum values of $R_{\text{work}}$ are dependent on the highest resolution of data available. With sufficiently diligent model building and refinement, with no other major experimental data pathologies, $R_{\text{work}}$ values obtained are typically around 0.18 for 2.0 Å, and 0.14 for 1.2 Å resolution experimental data. The estimate of experimental error, $R_{\text{merge}}$ is typically much lower ($\sim$ 5%) than the error associated with refined structures. Using simulated data, it has been hypothesised this disparity is due to our models not representing the crystal structure, rather than failures in refinement routines, or higher than expected experimental error [205].

### 1.3.12 Crystallographic occupancy

Crystallographic occupancy measures the proportion of sites in the crystal lattice for which an atom is present. For protein atoms, they must be fully occupied over any set of alternate states that they can be present in. For example, a side chain (trans +) rotamer might be present at 30% and the alternate position, (gauche) rotamer would need to present 70% of the time. This constraint of protein atoms to sum to unitary occupancy comes from the requirement that the protein form the crystal lattice.

However, any binding ligands or ions, may be present at lower occupancies, as they need not be bound in every copy of the protein in the lattice. In XChem fragment screening we have knowledge of the apo state of the protein, and therefore typically create a superposed model [202], in which the occupancy of the changed state (ligand bound) sums to unitary occupancy with the unchanged (ground state) model, discussed in section 1.6.

The structure factor equation 1.5 which has been adjusted for local atomic motion by changes to the atomic scattering factors as in 1.8, must also be adjusted for partial occupancy of atoms, represented by $n_j$:

$$\mathbf{F_h} = \sum_{j=1}^{atoms} n_j f^0_{\mathbf{S},j} \exp\left(-B_j \left(\frac{\sin\theta}{\lambda}\right)^2\right) \exp(2\pi i \mathbf{h}\mathbf{x}_j) \tag{1.34}$$

This implies that atomic displacement parameter (B factor) and occupancy are correlated. Thus, a higher B factor leading to attenuated scattering can also be achieved by a lower occupancy. This makes it difficult to refine both the occupancy and B factor with accuracy. Despite this correlation, it is possible to separate high B factor with low occupancy atoms to normal B factor with unitary occupancy by the electron density peak shape. However, due to the noise in electron density maps of proteins, it is not normally possible without prior knowledge about the local chemical environment [41].

As non-covalent ligands are subject to a binding equilibrium, the superposed model where the protein is modelled to have both a bound and ground state it cannot be presumed that either co-crystal or soaked compounds have the ligand bound at unitary occupancy. The majority (92.1%, 2016) of the ligands in the PDB are modelled at unitary occupancy [202]. Whether it is appropriate to model atoms with variable occupancy depends on the resolution of the data, as well as any prior knowledge suggesting a non-unitary occupancy. In this thesis models are of crystals which have a corresponding structure of the protein without the ligand, to which the ligand bound structure can be compared. Without this it may not be justified to model non-unitary occupancy for the ligand. Furthermore, allowing the occupancy to vary improves the fit to density by adding additional parameters [165], this should be justified by a sufficient improvement in model quality.

Furthermore, only a small percentage (3.2%, 2016) of ligands in the PDB are modelled with alternate ground and bound states that sum to unitary occupancy. Therefore some ligands are modelled at lower occupancy, with no contrasting model, the equivalent of assuming that the complementary fraction is modelled by vacuum.

The modelling of protein-ligand complexes using ground and bound states is most beneficial when there are many datasets collected at the same time, and the ligand

binding is considered a perturbation away from the ground state. If there is a variation in crystal form between the ligand free, and ligand bound structure then the superposition would not be appropriate. Modelling the superposition is likely complicating the work of the crystallographer when only ligand bound datasets are collected.

In some instances of model building, some crystallographers use zero occupancy atoms to model regions in which they are uncertain, but have a strong prior reason to include such atoms, i.e. a loop which is disordered, but has two potential endpoints. Atoms with zero occupancy are normally not refined, as the occupancy implies that the atom is not there. This practice can be helpful to indicate a hypothesis, but is likely to be problematic when non-crystallographers interpret the data, as often the model will be interpreted without the corresponding electron density. As this is a somewhat established practice in some communities, care should be taken not to refine these unevidenced portions of structure in modelling or refinement programs. Ideally a model which best explains the electron density should be that which is deposited in the PDB [117–119], and any other hypothetical models should be provided within the associated publication.

### 1.3.13 Model building and refinement

A molecular replacement model, or heavy atom substructure is the starting point for a model of a macromolecule. These need to be improved by iterative model building and refinement.

Model building is the process of fitting an atomic model to the electron density map. This can be done directly by a crystallographer, using tools such as COOT [206] to place residues according to electron density, protein sequence, and stereochemical restraints. Alternatively, automated model building can be attempted using software such as ARP/WARP [207], RESOLVE [208] or BUCCANEER [209]. Typically, these software can produce an acceptable model at resolutions as low as 3.0Å, and are routinely used as the first stages in solving a macromolecular structure.

Refinement is the iterative process of optimising model parameters, using maximum likelihood or least-squares algorithms, to reduce the difference between the structure factors resultant from the model to those resultant from the observed reflections. The

model parameters include the coordinate of the model atoms (x,y,z) and an appropriate B factor model, together with overall parameters such as scale factor, bulk solvent corrections and anisotropy corrections.

#### 1.3.13.1 Local real space refinement

Local real space refinement is the process of fitting the residue into the electron density in real space. Geometry based optimisation of the side chain to be placed, and to the changes made to the protein backbone on fitting the side chain are made. As the model continues to be built in real space it can accumulate errors in the bond length, bond angles, backbone torsion angles and other geometric properties [210]. These can not be trivially corrected against the initial electron density maps, and therefore local real space refinement is applied iteratively by the crystallographer in between cycles of global refinement.

#### 1.3.13.2 Global refinement using maximum likelihood

To correct the small local stereochemical and conformational errors present after the model is built and refined locally in real space, the model is subject to global reciprocal space restrained refinement. The target function to be minimised is the residual between the experimental data $F_{\mathrm{obs}}$ and the model $F_{\mathrm{calc}}$. Effectively, this is reducing the difference between the model and data. Most modern refinement methods to minimise the difference between model and data using the maximum likelihood method.

This discussion of maximum likelihood refinement is drawn from [41, 211].

If we consider the experimental data to be a set of structure factors amplitudes $|F|$, and the model parameters to be estimated as $x$, Bayes theorem can be written as:

$$P(x\,|\,|F|) = \frac{p(x)P(|F|\,|\,x)}{P(|F|)} = p(x)L(x\,|\,|F|) \tag{1.35}$$

where

- $P(x\,|\,|F|)$ is the posterior probability distribution of the model parameters $x$ given the experimental data $|F|$.

- $p(x)$ is the prior probability distribution of model parameter known before the experiment, i.e. expected bond lengths, bond angles, planarity of rings *etc*. This information can be applied as stereochemical restraints.
- $P(|F|\,|\,x)$ is the probability distribution of the experimental data given the model.
- $P(|F|)$ is the probability distribution of the data.
- $L(x|\,|F|)$ is the likelihood function which is proportional to the conditional distribution of experimental data when parameters are known

To estimate the best structural model, the posterior should be maximised. To find the likelihood, $L(x||F|)$, the joint probability distribution of all structure factors would be required, however this is not tractable. Assuming that errors in different reflections are independent the joint probability distribution of all amplitudes of structure factors is used. As the model parameters can be written in terms of the model structure factors $|F_{\text{calc}}|$, the probability distribution of the experimental data given the model can be written as:

$$P\left(|F|\,\middle|\,x\right) = P\left(\left(|F^h_{\text{obs}}|\right)^{\text{all reflections}}\,\middle|\,\left(|F^h_{\text{calc}}|\right)\right) = \prod_{\text{all reflections}} P(|F^h_{\text{obs}}|\,\big|\,|F^h_{\text{calc}}|) \quad (1.36)$$

Given that maximising a function is the same as minimising the negative logarithm of that function, 1.35 can be formulated as:

$$-\ln\left[P\left(\left(|F^h_{\text{calc}}|\right)\,\middle|\,\left(|F^h_{\text{obs}}|\right)^{\text{all reflections}}\right)\right] = -\ln|F^h_{\text{calc}}| - \sum_{\text{all reflections}} \ln L\left(|F^h_{\text{calc}}|\,\middle|\,|F^h_{\text{obs}}|\right) \quad (1.37)$$

In refinement, the aim is to find the form of $P(|F^h_{\text{obs}}|\,\big|\,|F^h_{\text{calc}}|)$ for all reflections. Note that reflections can be either acentric or centric, and are typically handled with different expressions for $P(|F^h_{\text{obs}}|\,\big|\,|F^h_{\text{calc}}|)$. Acentric reflections lack centrosymmtery and have no phase restricted reflections.

Centric reflections are phase restricted reflections resulting from scattering by centrosymmterical related atoms, that is those related by the point group structure of the crystal.

A relatively general form of $P(|F^h_{\text{obs}}|\,\big|\,|F^h_{\text{calc}}|)$ for a number of partial structures $N_{\text{part}}$, is given in 1.38, derived in [211]. Partial structures may be different subsets of atoms,

such as protein atoms, metals or be derived from unparameterised electron density, such as the bulk solvent or a segment of poorly phased electron density which currently cannot be interpreted using an atomic model.

$$P\left(|F_{\rm obs}|\,\middle|\,(F_{\rm calc})^{j=1,N_{\rm part}}\right) = \begin{cases} \dfrac{|F_{\rm obs}|}{\pi\left(2\sigma^2_{F_{\rm obs}}+\Sigma_{wc}\right)}\exp\left[-\left(\dfrac{|F_{\rm obs}|^2+|F_{wc}|^2}{2\sigma^2_{F_{\rm obs}}+\Sigma_{wc}}\right)\right] & \\ \times\displaystyle\int_0^{2\pi} P(\varphi)\exp\left[\left(\dfrac{2|F_{\rm obs}||F_{wc}|}{2\sigma^2_{F_{\rm obs}}+\Sigma_{wc}}\right)\cos(\varphi-\varphi_{wc})\right]d\varphi & \text{acentric} \\ \left[\dfrac{1}{2\pi\left(\sigma^2_{F_{\rm obs}}+\Sigma_{wc}\right)}\right]^{\frac{1}{2}}\exp\left[-\left(\dfrac{|F_{\rm obs}|^2+|F_{wc}|^2}{2\left(\sigma^2_{F_{\rm obs}}+\Sigma_{wc}\right)}\right)\right] & \\ \times\sum_{l=0}^{l} P(\varphi_l)\,\exp\left[\left(\dfrac{|F_{\rm obs}|\,|F_{wc}|}{\sigma^2_{F_{\rm obs}}+\Sigma_{wc}}\right)\cos(\varphi_l-\varphi_{wc})\right] & \text{centric} \end{cases} \tag{1.38}$$

where

- $P(\varphi)$ is the distribution of phases.
- $\Sigma_{wc} = \epsilon\Sigma_{j=1}^{N_{\rm part}}\Sigma_j(1-D_j^2)$ (1.39)
  - $\epsilon$ is the multiplicity of the diffracting plane
  - $D_j = \left\langle \exp\left[-\left(\dfrac{\Delta B_j|s|^2}{4}\right)\right]\right\rangle$ (1.40)
    - $\Delta B_j$ is the error in the B values of atoms in the j[th] partial structure.
    - $s$ is the vector of position in reciprocal space
    - $|s| = 2\sin\theta/\lambda$ (1.41)
- $F_{wc}$ is the weighted sum of partial calculated structure factors
- $\sigma_{F_{\rm obs}}$ is the experimental uncertainty in the amplitude of the structure factor

In the case where there is no prior phase information equation 1.38 becomes:

$$P\left(|F_{\text{obs}}|\left|(F_{\text{calc}})^{j=1,N_{\text{part}}}\right.\right)=\begin{cases}\dfrac{2|F_{\text{obs}}|}{2\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}}\exp\left[-\left(\dfrac{|F_{\text{obs}}|^2+|F_{wc}|^2}{2\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}}\right)\right] & \\ \times I_0\left(\dfrac{2|F_{\text{obs}}||F_{wc}|}{2\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}}\right) & \text{acentric} \\ & \\ \dfrac{2}{\pi(\sigma^2_{F_{\text{obs}}}+\Sigma_{wc})}\exp\left[-\left(\dfrac{|F_{\text{obs}}|^2+|F_{wc}|^2}{2\left(\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}\right)}\right)\right] & \\ \times\cosh\left(\dfrac{|F_{\text{obs}}|\,|F_{wc}|}{\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}}\right) & \text{centric}\end{cases} \tag{1.42}$$

where $I_0(X)$ is the zero order Bessel function of the first kind.

The log likelihood from 1.42 is given as:

$$LLK=\sum_h LLK_h \tag{1.43}$$

where

$$LLK_h=\begin{cases}c_a-\ln(|F_{\text{obs}}|)+\ln(2\sigma^2_{F_{\text{obs}}}+\Sigma_{wc})+\dfrac{|F_{\text{obs}}|^2+|F_{wc}|^2}{2\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}} & \\ -\ln\left[I_0\left(\dfrac{2|F_{\text{obs}}||F_{wc}|}{2\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}}\right)\right] & \text{acentric} \\ & \\ c_c+\dfrac{1}{2}\ln\left(\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}\right)+\dfrac{|F_{\text{obs}}|^2+|F_{wc}|^2}{2\left(\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}\right)}-\ln\cosh\left(\dfrac{|F_{\text{obs}}|\,|F_{wc}|}{\sigma^2_{F_{\text{obs}}}+\Sigma_{wc}}\right) & \text{centric}\end{cases} \tag{1.44}$$

when there is prior phase information the log likelihood is derived from 1.38 instead of 1.42.

Refinement is the optimisation of a function of a set of observations by changing the parameters of a model [212]. The log likelihood is an example of a function to be optimised. The process of optimisation, requires the calculation of the function given the parameters. Searching a set of possible model parameters is the simplest optimisation, however, it is not generally applicable in macromolecular crystallography as there are far too many possible parameters. Most methods utilise the gradient (or derivative) of the function (normally referred to as gradient descent). Methods which require the second derivative (curvature) are known as full matrix methods.

For a more complete description of how maximum likelihood methods are currently implemented see [8, 212].

#### 1.3.13.3 Bulk solvent correction

A significant proportion of a protein crystal is occupied by disordered solvent. The mean solvent content of crystals in PDB is 51% with a standard deviation of 10% [213]) Accurate modelling of solvent is still a challenge to be solved in macromolecular crystallography. The contribution of bulk solvent to structure factors is strongest at lower resolutions.

Different bulk solvent models exist. Babinet's bulk solvent model is based on the assumption that the only difference between the solvent and the protein at low resolution is their scale factor [214].. `REFMAC5` utilises a slight variation of this model [8], assuming that if the protein electron density is convoluted by a Gaussian kernel and multiplied by a scale factor, then the protein and solvent electron densities are equal.

$$\rho_{\mathrm{solvent}} + \mathcal{F}(k_{\mathrm{babinet}}) * \rho_{\mathrm{protein}} = \text{constant} \tag{1.45}$$

where

- $\mathcal{F}(X)$ corresponds to the Fourier transform of $X$.
- $*$ corresponds to convolution

Thus:

$$F_{\mathrm{solvent}} + k_{\mathrm{babinet}} F_{\mathrm{protein}} = 0 \tag{1.46}$$

Thus:

$$F_{\mathrm{solvent}} = -k_{\mathrm{babinet}} F_{\mathrm{protein}} \tag{1.47}$$

therefore:

$$F_{\mathrm{total}} = F_{\mathrm{solvent}} + F_{\mathrm{protein}} = (1 - k_{\mathrm{babinet}}) F_{\mathrm{protein}} \tag{1.48}$$

The second bulk solvent model used in `REFMAC5` is that originally derived as in [215]. This assumes that disordered solvent molecules are distributed evenly across the region of the asymmetric unit which is not occupied by the atoms of the modelled partial structures. The region of the asymmetric unit occupied by the atomic model is masked

out, using cavity detection to find any pockets within the atomic model which would have disordered solvent. A constant electron density is assumed outside the masked region, and the structure factors $F_{\mathrm{mask}}$ are calculated by Fourier transform.

When the atomic model is incomplete, such masking bulk solvent corrections can result in smeared out electron density, which can reduce its height in the unmodelled parts of the crystal. The full form of this solvent model (for `REFMAC5`) is given in [8]. Note that both types of solvent model can be combined.

#### 1.3.13.4 Refinement of initial model parameters

The overall parameters of the solvent models, and overall anisotropy and scale factors can be estimated using a least-squares fit of the amplitude to the observed amplitudes

$$\sum_{\text{working reflections}} (|F_{\mathrm{observed}} - |F_{\mathrm{total}}|)^2 \rightarrow \text{min} \tag{1.49}$$

In `REFMAC5` this is minimised using the Gauss-Newton method. Refinement of crystal twinning, not discussed herein is carried out alongside, see [8] for details.

Once overall parameters such as scale factors and twin fractions have been estimated, the maximum likelihood refinement of the model begins.

#### 1.3.13.5 Restraints in refinement

The prior $p(x)$ in 1.35 is the term which describes the model parameters, which include the restraints placed on the model. There are three main types of information:

1. Chemical information about the protein residues, ligands or nucleic acids, and the covalent links between them
2. Relationship between molecules related by NCS
3. Structural knowledge (i.e. known structures, restraints on interatomic distances, secondary structure elements *etc*).

Chemical information is contained in a CIF dictionary [216], used by COOT [3], `REFMAC5` [8] and `phenix.refine` [217].

Restraints can take the general form:

$$\sum_{\substack{\text{instances}\\ \text{of}\\ \text{parameter}}} \frac{1}{\sigma_r^2} f(r_m, r_i) \tag{1.50}$$

where

- $r_m$ is the current value of the geometric parameter, such as bonds, angles and chiralities.
- $r_i$ is the ideal value of that geometric parameter.
- $\sigma_r$ is the standard deviation of the distribution of the geometric parameter.
- $f(r_m, r_i)$ is a function of the current and ideal values of the geometric parameter.

To express maximum likelihood restraints in a form which is equivalent to least squares, the form of $f()$ is given such that:

$$\sum_{\substack{\text{instances}\\ \text{of}\\ \text{parameter}}} \frac{1}{\sigma_r^2} (r_m - r_i)^2 \tag{1.51}$$

This form of restraints can be applied to covalent bonds.

NCS is the presence of more than one copy of a motif (i.e. protein chain) in the asymmetric unit. This is frequently the case, with over half the PDB [117–119] consisting of oligomers. The molecules are related via translation and rotation. NCS is local and not limited to crystallographic symmetry operations. The differences between NCS related copies of a molecule are minimised in refinement, a detailed description of the parameterisation can be found in [8]. Restraints may also be defined to existing homologous structures, which can be useful especially at lower resolution, or in early stages of refinement.

Applying the isotropic B factor model (4 parameters per atom, plus bulk solvent corrections, TLS groups *etc*) to a protein with 25kDa and which would have approximately 1800 atoms without restraints would require refinement of approxiamtely 7,200 parameters. The number of unique reflections describes the quantity of data that can be refined against. At most resolutions it is insufficient to refine without prior information.

Stereochemical constraints reduce the number of parameters by fixed relation between stereochemical parameters. For example, the occupancy of a side chain with two alternative conformations must sum to unity. The majority of stereochemistry information is represented by restraints.

A restraint is the restriction of a geometric property to within a narrow distribution about the mean of that property, derived from existing high resolution small molecule [218–220] or protein structures [221]. Restraints on bond angle, bond length, dihedral angle, torsion angle can be represented as either distance restraints, or explicitly as angle restraints dependent on the refinement program used [41]. Planarity of amino acids, such as in phenyl or indolyl rings and the backbone peptide bonds, act as relatively strong restraints for groups of atoms. Chirality (i.e. $C_\alpha$ centre of amino acid) is a discrete restraint, as it is either correct or incorrect. It is represented as continuous parameter, as needed for refinement, by calculating the chiral volume [41].

B factor restraints limit the change in B factor between physically close atoms. If an atom has a low B factor, it has a low potential atomic displacement, and atoms covalently bound to it are unlikely to be highly mobile, therefore the change in B factor between two atoms is restrained. This can be done as a restraint of riding motion:

$$Q_B = \sum_{n=1}^{\text{distance}} w_{B(n)} \left( B_n^{\text{origin}} - B_n^{\text{target}} \right)^2 \tag{1.52}$$

Note that the distribution of B factors may also be restrained, i.e. between two NCS related molecules. This can be done by minimising the Kullback-Liebler divergence between the distributions, as discussed in [8].

External structure restraints to known structures can be useful, especially if a high resolution homologue is known. Structural information from the homologue can be applied by presuming the structure is the same, and then apply a restraint with a Geman-McClure function [8]. The use of Geman-McClure functions allow similar interatomic distances in the model and homologue to aid refinement by acting as regularisers, but in areas where there are large differences between the homologue and new model, the restraints are down weighted to allow for differences in the models.

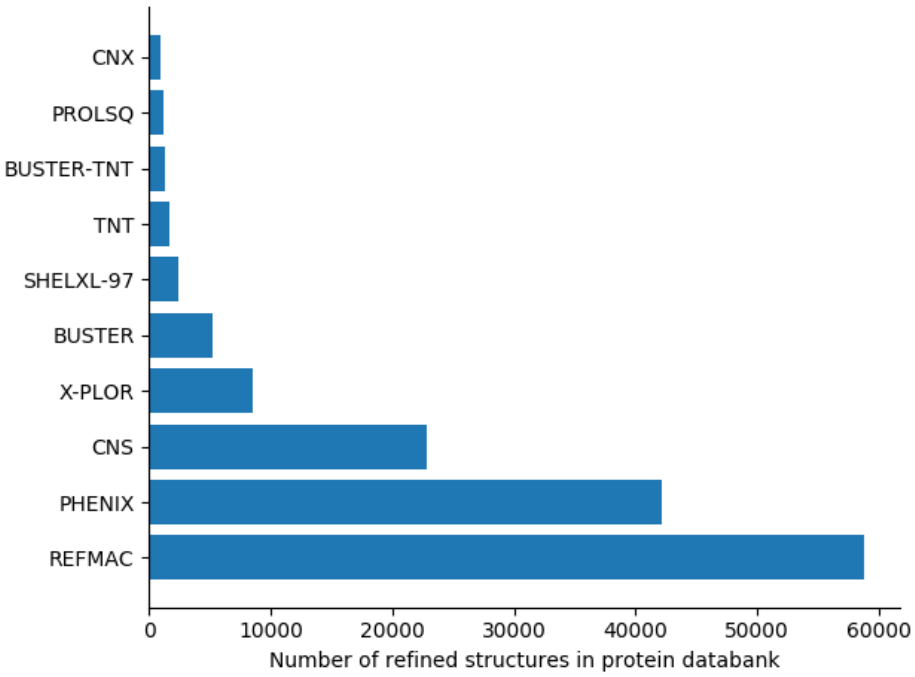



**Figure 1.6:** Refinement programs used for refinement of crystal structures in the PDB (Top ten). Note that TNT and BUSTER-TNT are subsets of the Buster program distribution.

#### 1.3.13.6 Refinement programs

By analysing the utilisation of refinement programs for structures deposited in the PDB [117–119], we can see that `REFMAC5` [8], Buster `refine` [222] and `phenix.refine` [223] represent a large proportion of the refinement programs used. We have chosen to focus on utilising these within this work. `SHELXL` [224] is most appropriate for the resolution of high resolution data, and thus as we have a reasonably wide range of resolutions within the XChem fragment screening data, and data collected for this project, I do not utilise this program. CNS [225, 226] and its predecessor X-PLOR although prevalent in the PDB [117–119] were not used in this work due to time constraints, and as they are no longer as commonly used as other software packages. To highlight the steps taken within macromolecular refinement program, see the flowchart overviewing `phenix.refine` in figure 1.7.

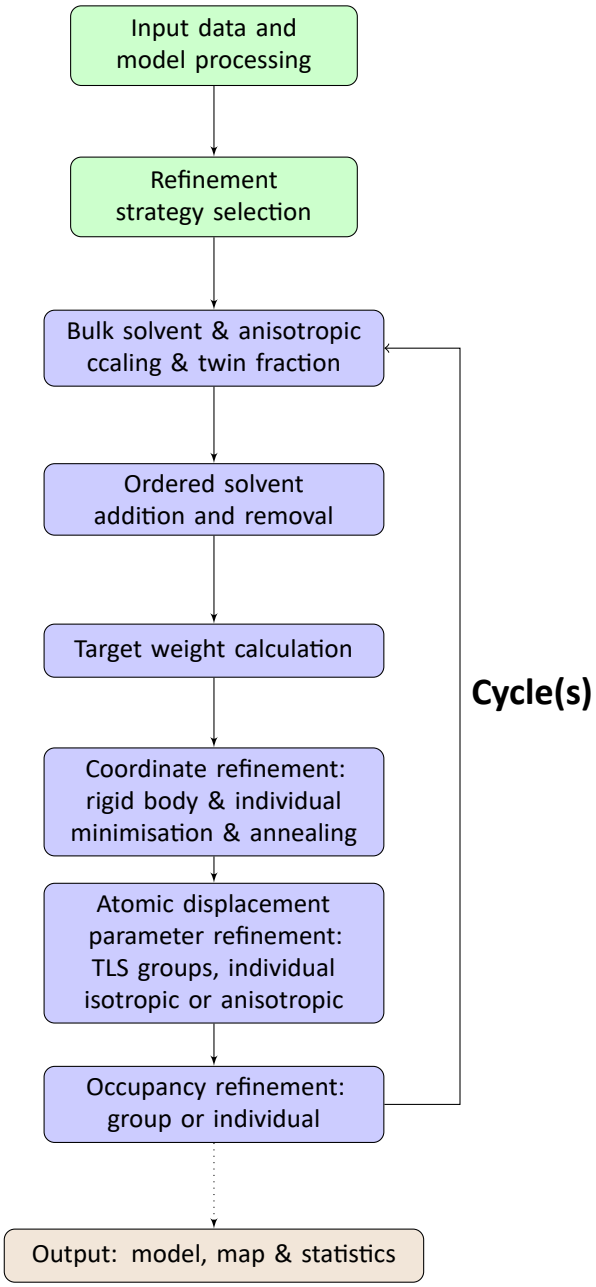



**Figure 1.7:** Flowchart describing the steps involved in the refinement of macromolecular structures using `phenix.refine`. Typically, refinement occurs in a series of iterative cycles, with different steps to improve the model's fit to the data. This is adapted from the description of the `phenix.refine` program [223].

#### 1.3.13.7 Refinement of occupancy

Occupancy refinement is typically applied to small portions of high resolution structures. For alternate side chain conformations, the total occupancy of all possible states should sum to one. For bound ligands or metal ions occupancy refinement should be done when unitary occupancy would lead to an excess of density beyond which there is evidence for. At lower resolutions there is a too low data to parameter ratio to justify refining occupancies.

With the possible exception of metal ions (as these could be involved in the substructure used to define phases, due to their increased number of electrons), these features of a protein are modelled near the end of the refinement process. The protein backbone should be well modelled before modelling side chains, and these should be modelled before fitting ligands or alternate side chain conformations.

By default, `phenix.refine` refines the occupancies of atoms in alternate conformations, as describe by altloc identifiers in the structure's PDB file, and any atoms which have a non-zero occupancy in the provided input structure. By default, constraints are setup such that alternate conformations of the same residue sum to unitary occupancy. Occupancy can be defined to be refined for groups of atoms, such as to be the same across a ligand and the alternate conformation.

## 1.4 Detecting weak binders using the PanDDA algorithm

Crystallographic fragment screening produces datasets that differ only by potential compound binding. Pan Dataset Density Analysis (PanDDA) [1] models ligand binding events as perturbations to a known ground state structure. This allows detection of a larger number of ligand binding events.

The PanDDA method utilises a reference dataset, by default this is the highest resolution dataset. The model used for the reference dataset is usually the apo model provided to DIMPLE [6] to produce initial models and maps for each dataset.

In the PanDDA method the first step is to align the datasets for analysis. The structures and importantly the corresponding maps need to be aligned to allow for

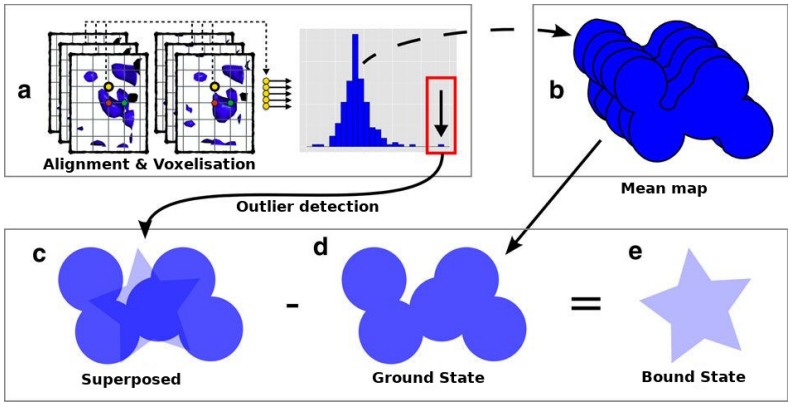



**Figure 1.8:** PanDDA is used to detect weak binding ligands. (**a**) Aligned datasets are voxelised, in each voxel, the distribution of electron density over datasets is calculated. Voxels with datasets with outlying density are clustered. (**b**) Averaging the ground state datasets into a mean map. (**c**) shows the superposition seen in the crystal, of the ground (**d**) and bound (**e**) states. The amount of electron density correpsonding to the ground state(**b**) to take from electron density of the dataset (**c**) is determined by contrast maximisation and called the background density correction (BDC). Adapted from [1].

comparison locally in real space of electron density maps of different crystals. There is non-isomorphism between cryo-cooled crystals, as cryocooling perturbs the local conformational heterogeneity differently for each crystal [151], therefore the global conformation of protein varies between datasets. Local alignment in PanDDA uses the $C_\alpha$ atoms which form the backbone of the protein chain, as these should be well evidenced if there is local motion altering the side chain positions. This may be the case when the structure is perturbed by ligand binding. For each $C_\alpha$, all residues with a backbone atoms within 10 Å 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, to produce a rotation-translation matrix for each $C_\alpha$. To align a dataset to the reference, the 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 centred on the $C_\alpha$. This alignment reduces global variation in the structures.

The variation in alignment matrices was shown to be small, with the alignment matrices for adjacent $C_\alpha$ atoms varying smoothly over the protein structure [203]. There are regions of proteins in which the variation in alignment matrices is higher across datasets, these are typically the more crystallographically disordered regions.

The crystallographic maps need to be aligned to be compared locally. A isotropic cartesian grid is used in the reference dataset to cover the aligned structures. The coordinates of the grid are transformed using the alignment matrix of the nearest $C_\alpha$. The electron density in the datasets native coordinate frame is then sampled at these points, and are presented as maps on the isotropic grid in the refence frame.

The crystallographic maps are voxelised, and the distribution of electron densities at each voxel is compared across maps. A Z-score at each voxel is calculated. This allows for the identification of individual dataset maps which deviate from the mean ground-state map. Criteria for significant differences in Z-score were identified as a minimal peak value of Z=3, and minimal volume of $10\,\text{Å}^3$ (note that the volume of a water molecule is 30 Å$^3$). These Z scores can be represented on the aligned grid as Z-maps, see figure 1.5(b). These can be clustered to determine whether a dataset has an interesting event.

Datasets that do not contain clusters of high Z-score are used to parameterise the ground-state model. A mean value of electron density for each voxel of these datasets is referred to as the ground state mean map. It is possible to use this ground state mean map to model features such as alternative conformations which represent the flexibility of the unperturbed model. This model can then be reprocessed through the PanDDA algorithm, often reducing the number of false positive events. Manual rebuilding of ground state density has been implemented in XChem Explorer [115], as discussed in section 1.6.1.

Future work on PanDDA could include parameterising the error in alignment, such that confidence intervals could be provided on the Z-scores.

Once an interesting event, a spatially close cluster of high Z scores, has been determined, the next step is to determine the fraction of the electron density representing the dataset which is well explained by the ground state model. This is known as background density correction (BDC). For each dataset different fractions of the mean ground state map are subtracted from the dataset's electron density map. The correlation between the resultant map and the averaged ground state map is calculated locally to the event (within 1 Å of the blob identified in the Z-map) and globally. The BDC value is obtained when the difference between the local and global correlation is maximised. An event map:

$$[\text{Event map}] = [\text{Observed map}] - \text{BDC} \times [\text{Mean ground state map}] \tag{1.53}$$

is used to determine what changes to the protein model explain the locally outlying electron density. The PanDDA algorithm allows efficient identification of datasets that are perturbed on ligand binding. It identifies the centroid of events of binders which could have been identified in $mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$ difference maps, as well as weaker binders where single dataset analysis would only show evidence for solvent.

The refined ligand occupancy detected using PanDDA events is correlated with the BDC. As the BDC is an estimate of how much density can be associated with the ground state versus the bound state, this is expected, and shows that the contrast maximisation to estimate BDC is appropriate. However, the pseudo-occupancy (1-BDC) from the contrast maximisation technique, event map density for changes appear somewhat stronger than density for unchanged atoms (i.e. surrounding protein). It was shown that refined occupancy of bound ligands was a factor of approximately two times greater than 1-BDC [203]. Updating this value is discussed in section 4.3.7.

The contrast maximisation to find the proportion of the ground and bound states can be affected by systematic errors; inconsistency in the ground state datasets can lead to large number of false positive events. This can include regions that were originally unmodelled, leading to events which correspond to features which are present in some or all of the datasets without a newly identified ligand, appearing as events. Future work to correct this is looking at both recognise whether the event is appropriately ligand shaped and sized, using machine learning on the existing set of events to select out features that identify an event as more confidently a ligand. Furthermore, analysing the similarity of electron density (using principal component analyse and t-distributed stochastic neighbour embedding) to group sub-populations of datasets will allow the filtering of datasets prior to `pandda.analyse` (private correspondence, Conor Wild, SGC).

### 1.4.1 Group deposition of PanDDA analysis using XChemExplorer

The PDB [117–119] has implemented a group deposition interface for the deposition of many related protein structures. These structures are linked by a common identifier, and relate to a single experiment such as a fragment screening campaign. For a XChem fragment screening campaign in which PanDDA [1] has been used, two deposition groups

are made aided by a group deposition interface in XChemExplorer [115, 227]. The first group contains the ground state structure and the corresponding mean map, and include the structure factors of every dataset used to undertake the PanDDA analysis in a single mmCIF. The ligand bound structures are deposited under the second group deposition ID, alongside the dataset structure factors and the event map in which that bound ligand was identified.

## 1.5 Modelling multiple states in crystals

Proteins exist in many conformations around their native folded state, which can be characterised by an energy landscape [228]. Understanding protein function, requires understanding which states a protein can access, and how a protein transitions between different conformational states.

A protein can sample different conformational states [229]. Stable sub-states are separated by energy barriers, and the transitions between these states occurs based on statistical thermodynamic potential. The conformer with the most complementary binding site for a ligand will be selected from an existing conformation.

There are many techniques to explore conformational flexibility in proteins: MD [230], NMR (especially relaxation dispersion [231]), single molecule spectrometry [232] and single particle imaging in cryo-EM [231]. Discussion of non-crystallographic techniques is beyond the scope of this thesis.

Constraints on a protein to form a lattice, mean that large scale motions are often not detectable in a protein crystal. Small scale motions can be detected in many crystals, and as crystallography provides atomic detail on the protein, these small changes can be modelled.

As a macromolecular crystalline lattice is formed from approximately $10^{13}$ molecules, which can be in subtly different conformations from one another, the resulting electron density map can show evidence of multiple conformations of the protein. However, a single structural model is typically built into the highest peaks of electron density [233], if the experimental data are of sufficiently high resolution, otherwise the model may be built to prefer continuity of density rather than build into erroneous peaks of density.

To better understand the protein it is possible to model multiple conformational states for high resolution datasets [234]. The main challenge is that there is a relatively low data to parameter ratio in crystallography, so fitting more than a single model without increasing other restraints causes over-fitting.

Limitations to using the single conformation models with B factor to describe conformational variety include a misleading degree of accuracy [235], and systematic underestimation of the rmsd from average coordinates [236]. B factors can model the magnitude and sometimes direction of the protein motion, they are limited to a Gaussian description of the probability density function for each atom position and cannot accurately capture anharmonic or multimodal motion [237]. Ensemble models can capture anharmonic motion. Analysis of MD trajectories shows that ensemble models better capture residue probability distributions [237], thus can better describe the position of residue over the course of the simulation,as shown by an improved fit to density for structures re-refined with an ensemble model.

### 1.5.1 Independent refinement of multiple structures

Multiple independent refinements from slightly perturbed initial models can reveal conformational heterogeneity [238]. To refine a structure multiple times independently, a diverse set of starting points for refinement is required. This is usually generated from an existing structural model. Starting diversity of structures can be achieved by simulated annealing, Monte Carlo sampling or random rebuilding [239]. Independent refinement gives un-coupled alternate models which explain the data with similar quality of fit. In some cases, independent refinement of multiple structures has been shown not to sample unknown conformations, but instead it gives an estimate of the imprecision of the coordinates of the main conformation [239].

### 1.5.2 Multi copy ensemble

Averaging the structure factors from the independent refinement of multiple structures can lead to a lower $R_{free}$, and therefore a better model than any of the individual refinements [239]. Therefore, by refining multiple conformations in an ensemble together [240]

a further improvement in model quality can be achieved by sampling the conformational heterogeneity of the protein. Studies into combining up to 16 copies of the protein shows an $R_{free}$ improvement (1.9% average of 50 targets) [237]. Typically these methods set the occupancy of a full copy of the protein to $\frac{1}{\text{models}}$ and then combine all models into a single ensemble model. Part of this improvement may be due to the ability of an ensemble to model anharmonic motion, and not that the protein occupies the states distinctly. Multi copy ensembles require many restraints to keep the data to parameter ratio reasonable.

### 1.5.3 Utilising molecular dynamics to produce an ensemble of structures

Single structural models derived from X-ray data do not adequately account for the dynamics of protein molecules. The inability to model anharmonicity and anisotropy leads to systematic model errors [236]. By sampling local molecular vibrations of the protein in the crystal using MD, it is possible to achieve a better fit to the crystallographic data [241]. MD simulations can be carried out with an energy term related to the fit to electron density. By biasing the MD to sample electron density, simulations tend to sample under-sampled regions (those with positive $F_{obs} - F_{calc}$) and reduce sampling in already over-sampled regions (those with negative $F_{obs} - F_{calc}$). To reduce over-fitting the global motion of the protein can be modelled with TLS parameters [242], allowing MD to sample only local atomic fluctuations [241].

Structural ensemble models informed by MD trajectories can fit the X-ray data better, whilst retaining a sufficiently small, and reproducible ensemble. As with all ensemble modelling techniques it can be relatively difficult to rationalise which differences between the multiple structural models are most informative. Furthermore, any increase in number of model parameters needs to be justified, otherwise the model may be subject to over-fitting.

### 1.5.4 Multi-conformer modelling

Multiple conformer modelling represents conformational diversity without creating multiple copies of the protein. Instead, local regions that are not well modelled by

a single conformer have a low number of alternate conformations introduced. This modelling can be done manually, but as model comparison metrics which account for over-fitting are global (such as $R_{\mathrm{free}}$) it can be difficult to know when a model is being improved locally. Therefore, manually modelling alternate conformations is time consuming, and as some are difficult to determine whether appropriate by visually inspecting electron density maps. This usually leads to models only having at most few alternate conformations modelled in regions where the crystallographer was most certain of conformational heterogeneity.

qFit is an automated routine for building multiple conformer models [243, 244]. This tool selects the optimal conformational heterogeneity at each residue, allowing 1-4 conformers that can explain the electron density. It can represent the model with fewer conformers at a residue if that is most appropriate. Arguably this makes qFit models more interpretable than ensemble models.

Superposed modelling discussed in section 1.6 can be consider a semi-automated multi-conformer model. The conformers are generated based on the differences between ground (apo) protein state and the model where a ligand has been fitted within the event detected by PanDDA.

### 1.5.5 Limitations of ensemble modelling and fitting multiple states in crystallography

Ensemble models are more complex to interpret than a single structural model. Furthermore, as no technique has proven to be the most appropriate description of conformational heterogeneity through multiple models, there is no consensus on how to deposit, store or present such models. Alternate conformer models with limited conformations (i.e. qFit models) improve interpretability, as it is easier to interpret a limited number of distinct conformations at only some positions in the model. However, such models are less well equipped to model anharmonicity without distinct conformations. Analysis tools outside of crystallography, such as prediction tools for homologous structures are designed to take a single dominant model, and will require adaptions to make best use of ensemble models. Presenting the ensemble model may also lead to an individual

inexperienced in ensemble modelling techniques to make a preferential selection of models which are members of the ensemble which better would support a hypothesis, potentially ignoring the dominant conformations.

As the PDB [117–119] has moved to mmCIF for structure deposition and introduced version history of structures, there is an opportunity to further develop the databank to allow for the release both single structural models and ensemble models within a PDB entry. Care should be taken to ensure the end user of the structure can interpret the model, whether that be a downstream program, or an individual.

### 1.5.6 Sampling side chain density using Ringer

Ringer [245] samples electron density by rotating side chains about their dihedral angle in the electron density map. It has been shown to detect lowly populated conformations in $> 18\%$ of uniquely modelled residues. This technique is used in chapter 6 to sample electron density to detect correlated motion. It is advantageous for this, as the results are relatively easy to visually compare between many datasets, and outliers can show ligand binding or uncommon alternate conformations.

## 1.6 Modelling protein ligand complexes as superposed states

Superposed modelling creates separable conformer models for the whole chain of the protein. This is required as the ubiquitous PDB file format, used in refinement, can only represent local conformational diversity as alternate positions of atoms or residues using altloc identifiers.

An ensemble or superposed modelling paradigm [202] was introduced with PanDDA that has a consistent ground state in every dataset. Structural perturbations are modelled as a bound (changed) state away from that ground state. The superposition is a model where the bound and ground states have been combined for refinement, as we propose they both simultaneously exist in the crystal.

PanDDA can be used to detect both ligand binding events and larger scale protein motions. If the protein movements result in multiple partly occupied conformations of multiple residues, it can be difficult to interpret the electron density.

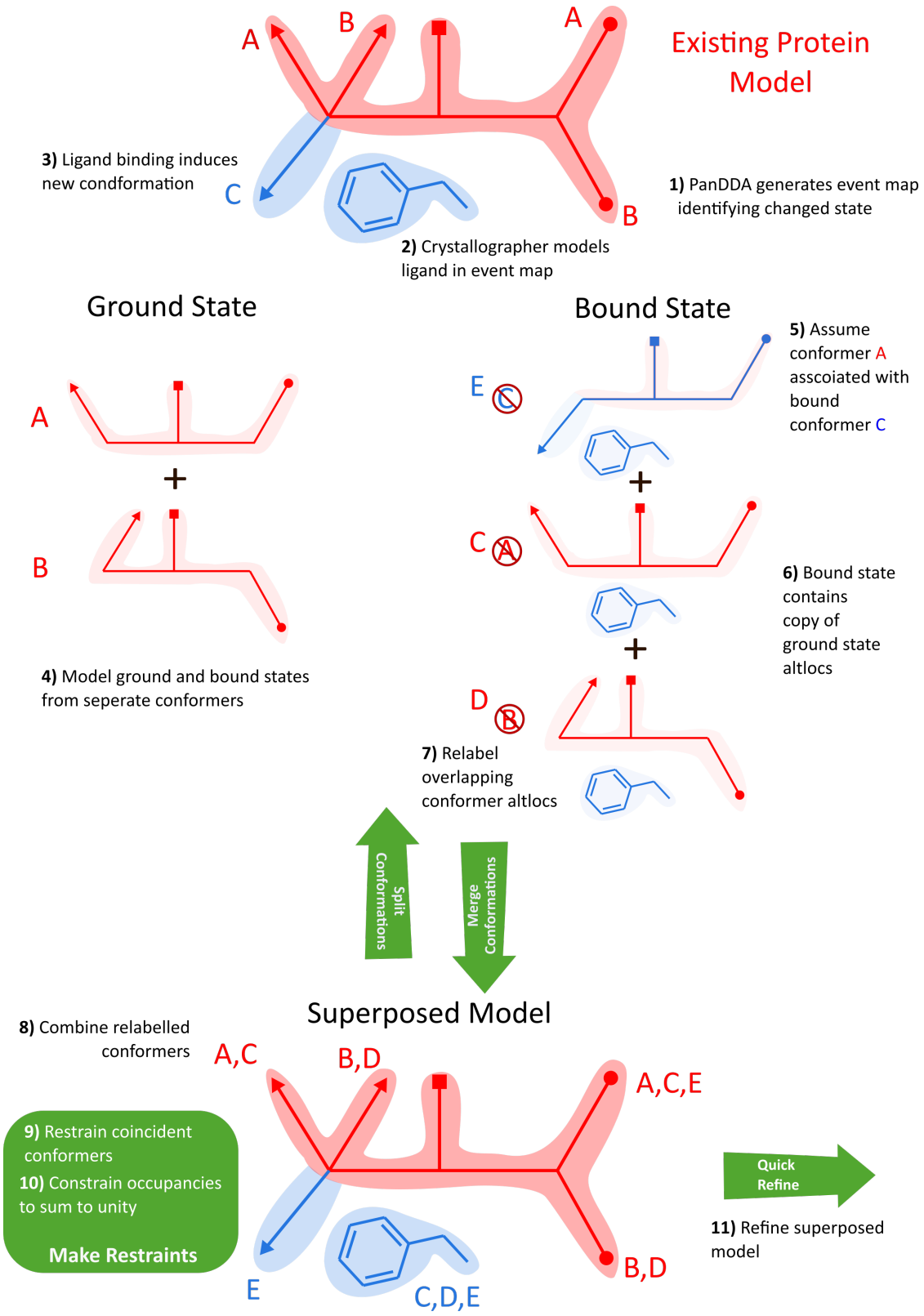



**Figure 1.9:** Superposed modelling framework (continued on next page).

**Figure 1.9:** (continued) The superposed model creation process for a three residue section of protein, where a binding ligand induces a new conformation in the first residue (triangles). Statistical analysis of local electron density, suggests an outlying region of density, modelled as a Z-map. `pandda.analyse` generates an event map using contrast maximisation at this location (**1**). The proposed ligand (shown in blue) fits the event density, and is modelled then real space refined to that event map, using the `pandda.inspect` interface (**2**). In fitting the ligand, a new induced conformation for the first residue of the protein is evident (**3**). The fitted model is split into ground and bound models as separate conformers (**4**). As there is a nearby residue with multiple conformations, the correct conformer to associate with the bound state is unknown, so a state is assumed, in this case that of conformer A (**5**). The bound state contains a copy of the ground state conformations (**6**). To achieve consistently labelled conformers the alternate location identifier are incremented to C,D,E for the bound state and A,B for the ground state (**7**). `giant.merge_conformations` is used to join the models into a single model (**8**). `giant.make_restraints` is used to produce restraints which tie conformers which are the same in the bound and ground state (such as A and C) to the same location (**9**). The occupancies are constrained so that the ground and bound states sum to 1.0 (**10**). `giant.quick_refine` is used to refine this superposed model (**11**), this provides the restraints and model suitable for refinement in `REFMAC5` or `phenix.refine`.

A superposed model represents the protein as a superposition of a ground and altered (bound) state. The ground state of the protein as determined by either an earlier apo protein model, or from modelling the mean electron density map across multiple datasets generated by `pandda.analyse`. The altered (bound) state represents the changes to the protein conformations and addition of any ligands that are identified in the protein.

Figure 1.9 shows how the superposed model and the corresponding restraints are generated. Firstly, the ligand is modelled into the event map generated by PanDDA. In general ligand binding is to be associated with some change in the existing protein model, such as the removal of water molecules from the site, or the movement of the protein chain to accommodate the ligand. Figure 1.9 illustrates a case of side chain movement, this is modelled as a new conformation (conformer E, first residue) associated with the ligand binding. The ground state of the protein is expected to be present in the proportion of the crystal where the ligand does not occupy the binding site, a ground state is modelled, separating each conformer. As it is not known whether the bound state is exclusively associated with the new conformation, i.e. the ligand may be able to bind without displacing the residue, copies of the ground state conformations are added into bound state (conformers C, D). These are relabelled to be distinct from those in the ground state, and then are combined into a single model.

One unexplored feature of superposed models is whether there would be an improvement in model quality by allowing each model to have a different bulk solvent mask. This effectively would allow bulk solvent to be in the site in the apo state, if appropriate, and not be in the site in the bound state. This may lead to a more physically accurate representation of the crystalline state in the model. Further testing is required to determine whether this would improve the refinement of the superposed model.

### 1.6.1 Using the mean of ground state electron density maps to improve apo-model

In some cases where there is more variation in the ground state, than can be seen in a single dataset it can be advantageous to use the mean map output from PanDDA, to model a more representative ground state. This has been integrated into XChemExplorer, as a COOT [3] interface which allows a user to fit the model into the mean map, and then refine this model into a representative dataset. Usually features such as alternate protein side chain conformations can be evidenced using this technique. This can be useful in reducing the amount of false positives found in a `pandda.analyse` run, as motifs common to many datasets do not appear as outlying density. However, care must be taken to ensure that such changes truly represent a common ground state, and not a sub-population of datasets, perhaps associated with slightly different crystallisation conditions. It can also be useful when resolution increases when many similar datasets are collected, as this can allow building a better structural model.

### 1.6.2 Refinement of superposed models

A superposed model is refined using the `giant.quick_refine` script. Any duplicated atoms are refined using tight external interatomic distance restraints, keeping the copies of the atom coincident. Duplicated atoms need to be grouped such that the occupancy of the state can be refined. For example, the third residue in figure 1.9 is restrained such that: occupancy $A + B + C + D + E = 1.0$ and occupancy $A = C = E$ and occupancy $B = D$. Extensions to the `giant.quick_refine` script to allow superposed refinement using Buster is discussed in section 3.2.1.5.3.

The script `giant.split_conformations` takes a superposed model, extracts the ground and bound state models. This split conformation model is currently the model deposited as the bound state in the PDB deposition of XChem fragment screening campaigns (historically the superposed model was deposited). It is also presented to the user as the bound state in the refinement interface within XChemExplorer, and can be used to fit further changes associated with the changed state.

## 1.7 Intact mass spectrometry of proteins

Intact MS was used to show that crystallographic occupancy can be correlated with an alternate technique to detect the amount of ligand present in the crystal.

Protein MS can be divided into top down approaches, in which the proteins are analysed intact and bottom up approaches in which the protein is digested into fragments. The proteolytic digestion, alongside techniques such as tandem MS process allows identification of peptide sequences which make up the protein, and can be useful for protein identification, identification of post translational modification and quantification of expression yield [246].

MS of proteins requires the protein to be ionised in the gas phase before they are analysed in an electric or magnetic field. Two ionisation techniques are prevalent; matrix-assisted laser desorption-ionization (MALDI) and electro-spray ionisation (ESI).

In MALDI, the sample is embedded within a matrix, which can absorb ultraviolet light and convert it to heat energy. The sample and matrix are vaporised and can be analysed in the gas phase. Ionisation by MALDI is usually coupled with a time of flight (TOF) spectrometer, and it general provides singly charged species and is thus suitable for the study of many proteins, such as from cell extract.

Electro-spray ionisation is a soft ionisation technique which can isolate intact ions from proteins in solution [248]. The transfer of ions from solution into gas phase involves three steps, first the dispersal of a fine spray of charged droplets, followed by solvent evaporation and ion ejection from the highly charged droplets, shown in figure 1.10. After ionisation the sample ions are accelerated into a mass analyser. ESI typically imparts

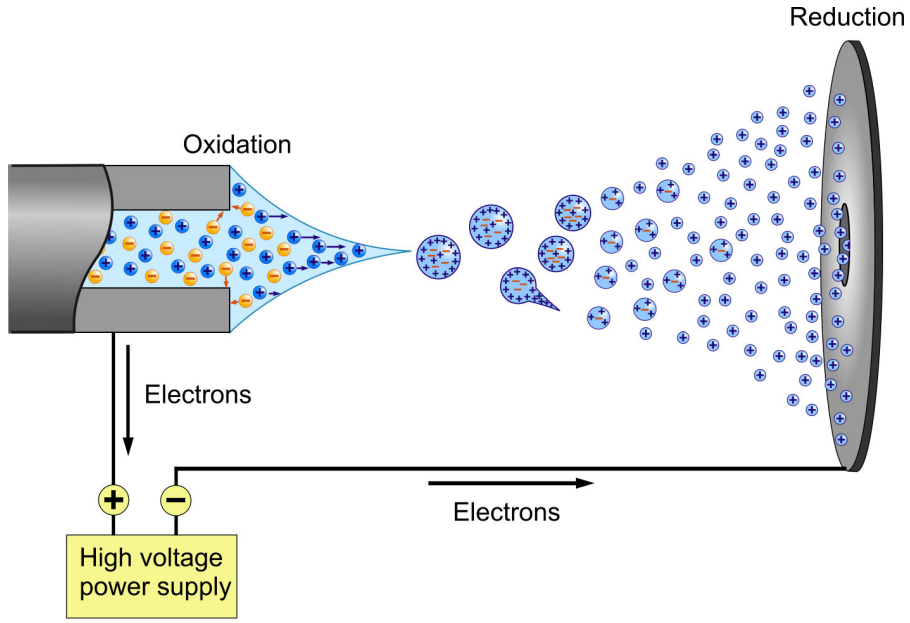



**Figure 1.10:** Schematic diagram of electrospray ionisation in positive mode. Within an ESI the sample is passed as a continuous stream through a capillary maintained at a high voltage (2.5-6.0kV) relative to the chamber wall, generating a mist of highly charged droplets. A nebulising gas (typically nitrogen) shears around the sampler to increase flow rate. The charged droplets pass down a pressure and potential gradient towards the analyser. Increased temperature and/or drying gas reduces the size of the droplet by evaporation. This leads to an increased surface charge density, within the droplet at this reduced size it is possible for ions at the droplet surface to be ejected into gaseous phase [247]. Reproduced from ESI positive mode (21589986840) produced by Andreas Dahlin released under CC BY 2.0.

more charge per protein, thereby generating multiple charged species usually in the mass to charge (m/z) range of 800 to 2000 Da $e^{-1}$.

Different mass analysers can be used. This work uses quantitative time of flight detectors. These work by accelerating ions through a high voltage. The velocity of the ions, and thus the time taken to travel through the detector is proportional to their mass to charge ratio. By pulsing the acceleration of ions, a mass/charge spectra can be obtained.

When MS is coupled to a HPLC system, a mass/charge spectra can be obtained at each time point. As over the time course of a chromatographic separation on a HPLC system the analyte can change. Coupling a MS system to a HPLC system allows separation of different analytes. Typically we average over a chromatographic peak of the analyte of interest.

The mass/charge spectra can be de-convoluted into a mass spectra, the maximum entropy (MaxEnt) algorithm is typically used [249, 250]. It can find the molecular masses

of the components, and has a high mass resolution, typically within 1 Da for typical protein masses (10-80kDa).

### 1.7.1 Ultra high performance liquid chromatography

Many MS methods, including ESI are preceded by liquid chromatography. This physically separates different components of the analyte. The liquid mixture containing the analyte(s) is distributed between immiscible stationary and mobile phases.

Reversed phase partition chromatography is often used prior to ESI [251]. In this the stationary phase is hydrophobic, in our use cases a column packed with silica particles functionalised with alkyl chains (either $C_4$ or $C_{18}$). This work uses gradients of acetonitrile and water in the mobile phase. Separation occurs due to the difference in path length of different size, charge and polarity species in the column due to sorption and desorption onto the silica beads. A chromatogram represents the separation of analytes over time. Ideally chromatographic peaks should be distinct.

Challenges in HPLC include ensuring adequate separation of analyte(s) of interest which is influenced by many factors. The column material, functionalisation and packing can provide different degrees of separation. The solvent in the stationary phase can either be kept constant or varied in a gradient. A solvent gradient can provide more separation but with increased time. Flow rate and pressure parameters can also be tuned to improve separation. High viscosity species such a polyethylene glycol can cause issues with separation [252]. Chapter 5 discusses issues this caused for analysing NUDT7 crystalline samples.

### 1.7.2 Intact mass spectrometry of crystals

Undertaking intact MS on crystals is valuable, as the crystalline state may be different to the solution state of the protein. For example, the crystallisation may occur after proteolytic degradation of the sample.

Gel electrophoresis can be used on crystalline material to get an approximate mass, this is often used to determine the presence and stoichiometry of protein complexes

present in the crystal. However, it is not sensitive enough to determine smaller changes to individual domains.

MALDI can be used to determine the complex state of an individual crystal [253]. For example, a study to determine the ternary complex of transcription factor binding protein, TATA box binding protein and DNA containing the TATA element undertook MALDI on individual crystals to determine whether the complex in the crystal contained all three components [254].

Preparing crystals for ESI intact MS, involves wicking the mother liquor from the crystal, washing the crystal in crystallisation reservoir solution, then wicking away excess solution. If the crystallisation condition contains compounds such as polyethylene glycol which are problematic as they can both damage the MS equipment and reduce signal by ion suppression, it is also advisable to wash the crystal in a solvent such as acetonitrile a few times [255]. Note that in this preparation, non covalently bonded ligands would not be expected to remain bound the protein.

Intact MS of crystals was used in chapter 5 as an orthogonal method for looking at crystallographic occupancy of a covalently bound ligand.

## 1.8 Thesis statement

In recent years the available data generated by crystallographic fragment screening has increased massively. Efficient analysis and interpretation of this data remains a challenge. This work aimed to determine the most appropriate modelling strategies for variable loop regions, weakly binding fragments and associated metal ions. Also aiming to understand whether crystallographic occupancy was being correctly parametrised and whether it could be used to rank ligand binding.

Existing work has shown a relationship between crystal soaking time and occupancy [256]. However, an outstanding question was whether fragment sized ligands have repeatable occupancies and thus could crystallographic occupancy can be used to rank the relative binding of compounds. Chapter 3 demonstrates that there is variation in ligand occupancy for repeated datasets of protein-ligand complexes. Furthermore, low occupancies (below 0.4) were not sampled frequently. As B factor and occupancy are

correlated, analysing how these parameters were handled in refinement, occupancy had a higher variance suggesting that the B factor is currently more stable in refinement.

As crystallographic occupancy was not reproducible for the repeat crystals, the aim was to understand whether the refinement was sensibly sampling potential values. Therefore, a brute force search of different ligand B-factors and occupancy values was developed. Given that this showed a smoothly varying function near the minima of difference density over the ligand region, refinement appeared to be performing sufficiently, but the parameterisation of the model was insufficient.

Chapter 4 focused on analysing weakly binding ligands from XChem fragment screening, to explore the distribution of ligand occupancies. As in repeat datasets of the same protein ligand complex low occupancies were not commonly sampled across the existing XChem hits . To ensure this wasn't due to the superposed modelling framework, or the use of any particular refinement method, refinement was tested without the superposed model, as well as using `REFMAC5`, `phenix.refine` and Buster `refine`. The lack of sampling of low occupancies held, and again the different refinement methods reached different occupancies for the same datasets.

To explore whether there was a detection limit for low occupancies, or whether the lack of sampling of low occupancies was due to a poorly defined model, an orthogonal experimental method to determine occupancy was devised. Using a protein system amenable to intact MS of crystals, increasing amounts of an efficiently binding covalent compound were mixed with protein, and the proportion of compound that could be detected after crystallisation was measured. This proportion of labelling was then used to calibrate the refinement of crystallographic occupancy of similar crystals. Low occupancies are only sampled if the ligand model is altered to have a B factor similar to its environment. These datasets can be used to calibrate future methods developed to parameterise occupancy and B factor of weakly binding ligands.

During the deposition of a fragment screening campaign (*T. cruzi* FPPS) metal ions in the active site, near some fragment binders were being poorly modelled (chapter 2). Adding distance and occupancy restraints for the metal ions which were not related to those which restrain the superposed models of the ligands, was necessary to ensure

the metal ion had a reasonable coordination sphere. However, when deployed with magnesium ions, there was still positive $F_{\mathrm{obs}} - F_{\mathrm{calc}}$ density over the region. As prior work shows other divalent ions had bound in the site, zinc ions were modelled, which refined consistently across multiple datasets.

The `pandda.analyse` algorithm can detect both small and large perturbations. In a case study where large perturbations were identified in a variable loop region, there were many events which were unexplained by ligand binding alone. As the region was known to be flexible, it was posited that the loop may have adopted an alternate position. In a dataset with a covalent ligand bound it was possible to model a different loop conformations based on a homologous structure. The crystallographer suspected that the other events were related to the loop rearrangement. By refining 3 models for each dataset (the initial loop position, the rearranged loop position and a superposition of both loop positions) then presenting the RSCC of the loop region to the crystallographer, it was possible to identify which loop model was appropriate. This showed that the loop rearrangement was present stochastically in datasets without ligands bound, but was also triggered by most covalently binding compounds. This allowed the crystallographer to prioritise which ligands to do follow up chemistry on.

By building models into the mean electron density map across different crystals it is possible to determine alternate side chain conformations which could not be modelled using a single dataset. Using a model with these additional conformations reduces the number of false positives detected in `pandda.analyse`. It was hypothesised sampling the electron density of multiple datasets would detect weakly occupied alternate conformations. Utilising Ringer [245] across multiple datasets there were some residues which had outlying electron density. In a case study (chapter 6) the detection of alternate conformations of multiple residues was related to the binding of an allosteric ligand, and movement of a distal residue in the active site, which coordinates metal ions. Some occupied conformers will be related to the conformational heterogeneity between cryo-cooled samples [151], and some may be related to ligand binding. Therefore it was difficult to determine which conformations were correlated. It was not possible to algorithmically identify correlated conformational changes. However, viewing the

electron density for each residue across all datasets remains a useful tool for identifying outlying conformations when model building.

# 2

# Utilising superposed modelling

## Contents


**2.1 Introduction . . . . . 79**
**2.2 Adaptations to the superposed modelling paradigm . . . . . 80**
**2.3 Metal ions near the ligand binding site of *T. cruzi* required additional distance and occupancy restraints . . . . . 81**
2.3.1 Introduction . . . . . 81
2.3.2 Methods . . . . . 84
2.3.3 Results . . . . . 85
2.3.4 Discussion . . . . . 90
**2.4 Local real space correlation coefficient determined appropriate switch I loop models in a GEF/GTPase complex . . . . . 91**
2.4.1 Introduction . . . . . 91
2.4.2 Methods . . . . . 93
2.4.3 Results and discussion . . . . . 97
**2.5 Discussion and future work . . . . . 100**


## 2.1 Introduction

The superposed modelling framework introduced with PanDDA [1] has been shown to be appropriate in modelling the binding of fragment sized ligands [257]. However, it had not been shown to be useful for modelling larger motions, such as loop rearrangement, which can also be detected with PanDDA. This chapter shows by using metrics such as RSCC alongside a superposed model, the most appropriate model of the loop region

can be determined. This allowed a crystallographer to classify datasets and understand the effects of compound binding.

Another case study, of a protein ligand binding with metal ions bound in the active site, required restriants related to the superposition of the ground and bound states, as well as distance restriants to ensure coordiantion of the metal ion by water molecules and surrounding residues.

Before discussing these case studies, improvements made to the superposed modelling framework since the publication of PanDDA are outlined. These improvements have been applied on all XChem fragment screening campaigns since 2017.

## 2.2 Adaptations to the superposed modelling paradigm

The superposed modelling framework is outlined in section 1.6. This section discusses adaptations made to improve the superposed modelling framework since publication [202].

For a more consistent user experience within XChemExplorer [115] minor changes were developed to the original protocol (`giant.merge_conformations`, `giant.split_conformations` and `giant.quick_refine`) for refining superposed states outlined in [202]. To test these changes the DCP2 fragment screening data presented in appendix B.1 was used.

Firstly, assuming the only change to a protein was to add a bound ligand, without inducing new protein conformations or displacements of existing water molecules , a superposed model could not be determined. As no change in the ground state to contrast the ligand against could be observed, and the superposed refinements cannot be initialised with constraints to sum the ground and bound (changed) state to unitary occupancy. In this case the refinement of an incomplete occupancy group for the ligand was tested. This would allow the ligand occupancy to vary between 0 and 1.0, and not be constrained to sum to unity. Based on the refinement results it was decided that the introduction of an unconstrained parameter, did not aid model interpretation, so the constrained refinement model was kept as default behaviour. If there was no modelled change in protein conformation or change in water molecules when a ligand is bound,

some method to contrast the bound and altered state is required. In chapter 4 this was achieved by adding dummy water molecules, at the centroid of the ligand. This allowed comparison of the superposed refinement and non-superposed refinement across the XChem hits.

Whilst assisting individuals modelling ligand bound structures using the XChem platform, it was noted that often the changes to the model were introduced during the refinement stage. Therefore, the default behaviour of the XChemExplorer refinement interface was set to use a superposed model, after the initial position of the ligand model was determined in the `pandda.inspect` interface. Technical changes were made, mostly around package import errors for cluster job submission, to allow refinement by `giant.quick_refine` with `phenix.refine` from this refinement interface.

With help from Nicholas Pearce, issues with the early implementation of `giant.merge_conformations` and `giant.split_conformations` were tested and fixed. Blank altloc identifiers (for residues that have a single conformer) are now reset to unitary occupancy after merging. This reduced the number of free parameters in the model being introduced, avoiding unnecessary occupancy refinement of some atoms. Initially the merging and splitting of conformations made assumptions about the relative occupancy of states, and used rescaling after refinement. This rescaling was removed, so that the refined occupancy did not change after merging and splitting conformations.

These changes resulted in the protocol outlined herein and have been applied across all XChem fragment screening campaigns since late 2017.

## 2.3 Metal ions near the ligand binding site of *T. cruzi* required additional distance and occupancy restraints

### 2.3.1 Introduction

*Typanosoma cruzi* Farnesyl Pyrophosphate Synthase (*T. cruzi* FPPS) underwent XChem fragment screening to identify non-bisphosphonate inhibitors. Two metal ion binding sites were identified close to the active site. The initial ground state model contained

two magnesium ions, but this model was poorly supported by the electron density across multiple datasets. Zinc ions modelled at partial occupancy provided a more consistent fit to density across multiple models. To fit these metal ions to the superposed model close to ligand binding sites, additional occupancy and distance restraints for the metal ions were required in order to retain correct distances to coordinating residues and water molecules.

#### 2.3.1.1 Biological relevance of *T. cruzi* FPPS

*T. cruzi* is the causative parasite [258] for Chagas disease. Inhibiting *T. cruzi* FPPS is known to inhibit proliferation of parasites *in vivo* and *in vitro* [259]. Bisphosphonates are active site directed FPPS inhibitors [260–262], and are used clinically for bone resorption therapy [263]. Crystallographic analysis of bisphosphonates bound to FPPS, have shown that they mimic the FPPS substrate DMAPP [264, 265], inhibiting the enzyme by reducing access to the site for DMAPP, and restricting the conformational change the protein undergoes during catalysis. However, bisphosphonates are rapidly removed from circulation via bone mineral binding [266], and thus are not ideal for treating Chagas disease. Therefore fragment screening was undertaken to look for non-bisphosphonate inhibitors.

FPPS is a key enzyme in the isoprenoid biosynthesis pathway [267]. FPPS catalyses the condensation of isopentenyl diphosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) to form geranyl diphosphate (GPP) [258]. A second molecule of IPP is condensed to form farnesyl pyrophosphate (FPP). FPPS has two conserved aspartate rich motifs (DDXXD). These are present on opposite walls of the protein's central cavity approximately 12 Å apart, with the carboxylate groups pointing into the cavity. The first motif can bind DMAPP, GPP or FPP, whereas the second is involved in binding IPP [264].

Non-bisphosphonate inhibitors that bind to an allosteric pocket were found for human FPPS in a recent NMR fragment screening campaign [268, 269]. It has also been shown that the product of FPPS, farnesyl pyrophosphate can bind to this allosteric pocket and can lock the enzyme in an open and inactive state [270].

#### 2.3.1.2 Metal ions in existing structures of FPPS

In existing structures of FPPS there are three metal ion sites key to both the catalytic function of the enzyme, and the binding of inhibitors. Thus, accurate models of the metal ions, and any local protein movement are required for identifying new active site inhibitors found in fragment screening.

Structures of human FPPS in complex with bisphosphonate inhibitors involve interaction with both conserved DDXXD motifs through a trinuclear metal site. Different divalent cations ($Mg^{2+}$, $Mn^{2+}$ or $Zn^{2+}$) have been observed in the structures [264]. These structures all have octahedral coordination of the ions, involving the bisphosphonate oxygen atoms and the side chain carboxylate groups of aspartic acid 103, aspartic acid 107 (first DDXXD motif) and aspartic acid 243 (second DDXXD motif) and water molecules.

The bisphosphonate binding stabilises the closed conformation of human FPPS. In comparison to the apo-protein, this conformation involves the rigid body motion of the final 130 C-terminal residues. The bisphosphonate binding brings the two DDXXD motifs closer together; two loops close the entrance to the binding site resulting in a partially closed conformation. This partially closed state is required for the catalytic cycle, as then substrate binding of the IPP can follow. This involves the basic C-terminal tail (lysine350, arginine351, arginine352, lysine353) becomes ordered, closing the IPP binding site.

Structures of *T. cruzi* FPPS in complex with bisphosphonate inhibitors show similar coordination of the metal ions [258] to when FPPS is in complex with alendronate (or Risedronate) a magnesium ion is coordinated by the aspartic acid 98 and aspartic acid 102 of the first DDXXD motif, an oxygen atom from each phosphate group for the bisphosphonate inhibitor and a water molecule. The second magnesium ion is coordinated by two phosphate oxygen atoms, the carboxylate of aspartic acid 250 (second DDXXD motif) and three water molecules. The final magnesium ion is coordinated by three water molecules, aspartic acid 98 and aspartic acid 102 of the first DDXXD motif and one phosphate oxygen from the bisphosphonate. Figure 2.1(a) shows the coordination of the ions when *T. cruzi* FPPS is in complex with a bisphosphonate inhibitor.

#### 2.3.1.3 Metal ion modelling

Metal ions are present in many macromolecules, and are often required to maintain the structural stability of the molecule or act as co-factor in catalysis reactions [271]. Often it is necessary to determine the most likely metal ion from a selection of possible ions that were present in the cell, the purification buffers, or the crystallisation condition. Furthermore, the evidence for the metal ion(s) in a structure must be assessed against the quality of the diffraction data and any local effects of radiation damage [272]. The lower resolution of macromolecular datasets in comparison to small molecule datasets results in higher uncertainties in bond length and bond angle. Therefore, these properties cannot be relied upon to uniquely determine the identity of the bound ion [273].

In macromolecular crystallography, metal ions often cannot be uniquely identified from one another, or in some cases from water molecules, especially when present at less than unitary occupancy. Further experimental work and careful analysis of metal coordination chemistry and geometry, is required to confirm which metal ion species is most likely to explain the electron density.

If the absorption edge of the target metal falls within the tuneable wavelength of the synchrotron beamline or source used to collect diffraction data, anomalous diffraction data collected above and below that absorption edge can be used to identify the metal [272]. Other techniques for metal identification include florescence scans [272] or X-ray absorption spectra [274]. Validation tools such as CheckMyMetal [273] can use local binding environment and metal coordination to assess whether a macromolecular structure has appropriate modelled metal ions.

### 2.3.2 Methods

#### 2.3.2.1 Fragment screening to identify non bisphosphonate inhibitors of *T. cruzi* FPPS

Fragment screening by NMR and crystallography of *T. cruzi* FPPS was undertaken by Joy Kristin Petrick. NMR fragment screening of 1818 fragments revealed 109 diverse fragment hits. Counter screening by NMR against human FPPS and *T. brucei* FPPS, showed selectivity of the fragments at this early stage of screening.

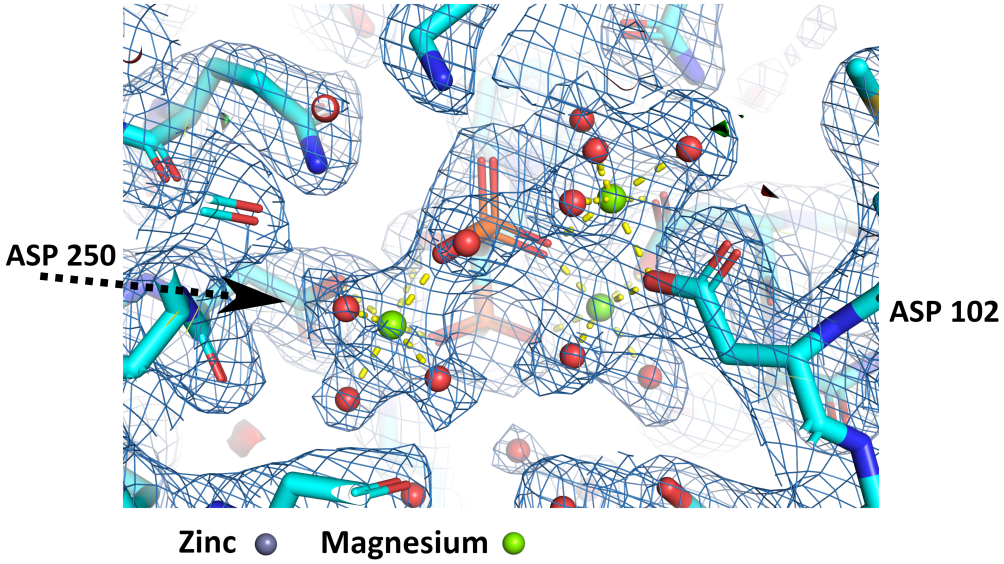
Crystallographic screening was undertaken at the XChem facility, detailed in appendix section B.2. We prepared the crystal system, screened compounds, collected diffraction data and prepared final models of *T. cruzi* FPPS. 35 datasets were deposited in the PDB [117–119], all of which required restraints not normally created for a superposed model to ensure metal ions were correctly refined.

### 2.3.3 Results

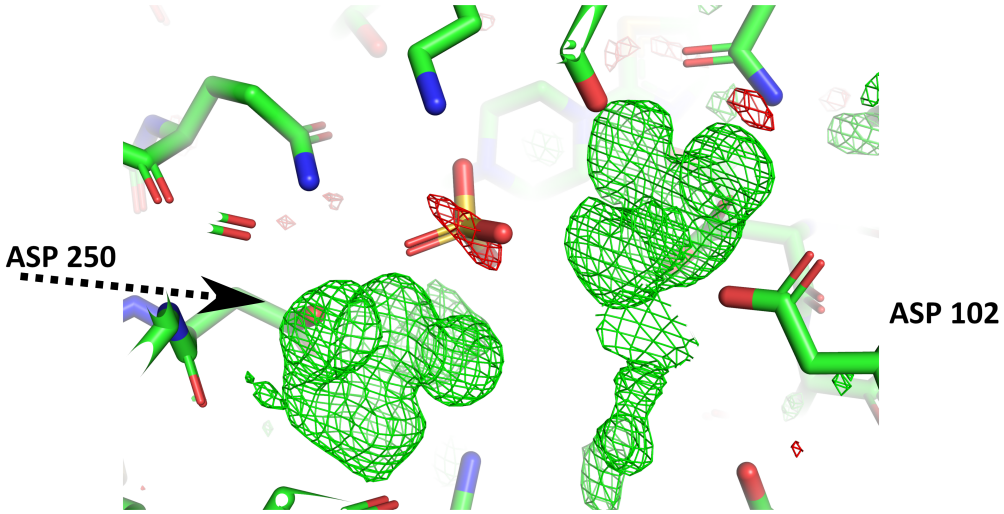
The structure of *T. cruzi* FPPS used for fragment screening has two identifiable metal ion sites, which were best modelled by partially occupied zinc ions with local restraints to ensure correct coordination of the ion by water molecules and nearby residues.

These sites correspond to two of the three magnesium ion ($Mg^{2+}$) sites discussed in [258], however significant differences were observed. The crystallisation condition contained zinc sulphate ($ZnSO_4$). In solution this dissociates into zinc ($Zn^{2+}$) and sulphate ions ($SO_4{}^{2-}$). As the purified protein may have carried magnesium ions ($Mg^{2+}$) through from the cell, there was initial uncertainty whether the sites were occupied by magnesium or zinc ions. Refinement of *T. cruzi* FPPS with magnesium ions yielded reasonable $2mF_{obs} - DF_{calc}$ maps in some datasets. However, in all cases the sites still had a large amount of unexplained density evident in the $mF_{obs} - DF_{calc}$ difference density maps, despite being fully occupied by the magnesium ion, see figure 2.1(c). Although it is possible that the structure had a superposition of magnesium and zinc ions present at the same time, this was considered unlikely, and the structure was modelled to have partly occupied zinc ions, as shown in figure 2.1(d).

To determine the most appropriate metal ion model the positions of metal ions in previously determined structures of *T. cruzi* FPPS were compared. These structures included a structure without inhibitors bound (PDB accession code: 1YHK [258]). As well as structures of *T. cruzi* FPPS in complex with Risedronate, DMAPP and $Mg^{2+}$ (PDB accession code: 1YHL [258]), *T. cruzi* FPPS in complex with alendronate, Isopentyl diphosphate and $Mg^{2+}$ (PDB accession code: 1YHM [258]) and *T. cruzi* FPPS in complex with isopentenyl pyrophosphate and 3-Fluoro-1-(2-hydroxy-2,2-bis-phosphono-ethyl)-pyridinium (PDB code: 3ICN [276]). The conformation of aspartic acid 98 present in

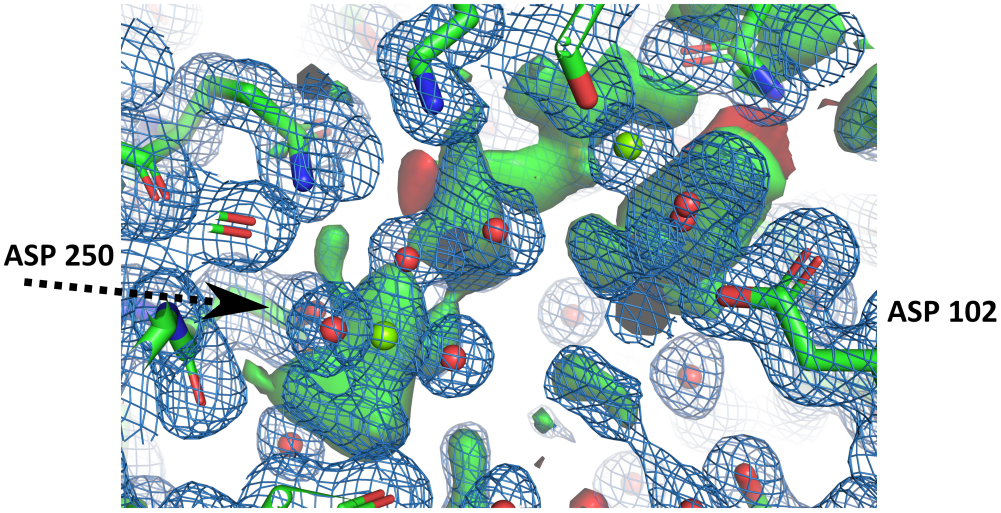



**(a)** 3ICN with added waters

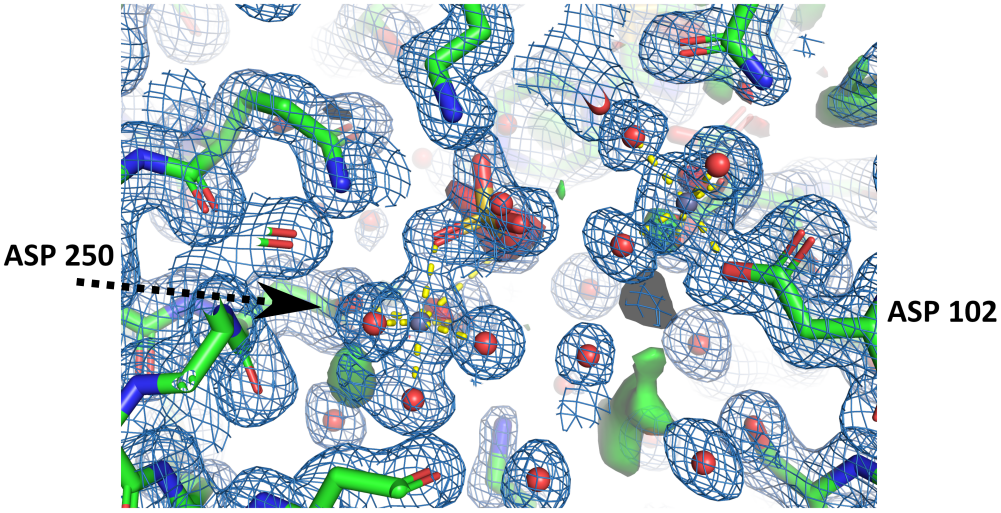



**(b)** TCRUFPPS-x0064 polder map omitting metal ions

**Figure 2.1:** Two zinc ions at partial occupancy are the most appropriate model the active site of *T. cruzi* FPPS in XChem fragment screening. This differs from the three divalent ions found in other FPPS structures. For all figures $2mF_{obs} - DF_{calc}$ electron density maps are countered at $1\sigma$ and shown as a blue mesh, and the $mF_{obs} - DF_{calc}$ electron density maps are contoured at 3 $\sigma$ and shown as green (positive) and red (negative) surfaces. **(a)** shows the site of *T.Cruzi* FPPS with isopentenyl pyrophosphate and 3-Fluoro-1-(2-hydroxy-2,2-bis-phosphono-ethyl)-pyridinium (PDB code: 3ICN). The model of three magnesium ions explains the site well. Note that two water molecules have been added which were not included deposited structure, appendix figure C.1 shows the model without these added waters. **(b)** shows a polder omit map, contoured at $3\sigma$ where the metal ions and coordinating waters have been omitted.




**(c)** TCRUFPPS-x0064 from fragment screening modelled with magnesium ions




**(d)** TCRUFPPS-x0064 from fragment screening modelled with zinc ions

**Figure 2.1:** (continued) **(c)** shows the fragment screening structure with the magnesium ions refined at full occupancy. This gave the ion a low B factor of 5 Å$^2$, and leaves a large amount of unexplained difference density. Note that the collapse of waters into the density for the magnesium is indicative that an ion with more electrons such as zinc was bound. **(d)** shows the same structure modelled with two zinc ions with their occupancy refined, note the large reduction in difference density. Alignment of the structures is achieved via Pymol `align` method [275].

the apo-protein (1YHK) closely matched the position of the ground state conformer present in the fragment screening models. On binding of active site ligands, the occupied conformer of aspartic acid 98 was closer to that of the structures in complex with bisphosphonate inhibitors (1YHL, 1YHM and 3ICN), in which it was rotated by 90° about $\chi_1$ to accommodate metal binding. In the fragment screening structures aspartic acid 98 coordinated a single zinc ion, this is behind the right-hand zinc in figure 2.1(d), instead of two ions in the previously determined structures. In some copies of the structures there is weak density evidence for a third zinc ion, however this was not considered sufficient evidence to change the ground state model. This ion was also coordinated by aspartic acid 102, from the conserved first aspartate rich (DDXXD) motif.

The second zinc ion was coordinated by aspartic acid 250, and in some datasets by a free sulphate ion. This closely matched the magnesium ion site (1YHL, 1YHM and 3ICN) near aspartic acid 250, involved in the conserved second aspartate rich (DDXXD) motif. In the previously determined structures the magnesium ion was coordinated by the phosphates of the bisphosphonate, figure 2.1(a).

In figure 2.1(d) negative electron density was observed in the difference density map over the sulphate ion, this may be due to local radiation damage, as the sulphurs atoms have a higher number of electrons and are susceptible to local damage.

#### 2.3.3.1 Applying occupancy groups and restraints

By default, the zinc ions, as they are close to the ligand binding site, they are placed in an occupancy group with the ground state of the protein, which will vary in opposition to the ligand bound occupancy. Instead, the occupancy of the zinc ions was refined between 0 and 1 without coupling the occupancy to that of the nearby ligand. This required the introduction of a further occupancy group. This incomplete group stops the zinc ion occupancy being constrained to that of the ground state atoms in superposition with the ligand. Protein residues with conformations only detected in the ligand bound state, and displaced water molecules constrained occupancy groups were used to produce an appropriate superposed model. The metal ions required additional distance restraints to maintain coordination with the surrounding atoms. Their occupancy was refined so

**Table 2.1:** Geometry and coordination of ions in alternate refinements of *T. cruzi FPPS*. Note that the occupancy of magnesium ions was not refined in these models. Values are shown are mean (standard deviation).

| | **Refinement 1** | | **Refinement 2** | |
|---|---|---|---|---|
| | **Zinc 1** | **Zinc 2** | **Magnesium 1** | **Magnesium 2** |
| **Occupancy** | 0.63 (0.14) | 0.53 (0.15) | 1.0 (0.0) | 1.0 (0.0) |
| **B factor** [Å$^2$] | 22.88 (7.71) | 30.44 (10.11) | 16.10 (7.03) | 25.19 (9.20) |
| **Mean distance to coordinating water molecules** [Å] | 2.04 (0.17) | 2.10 (0.18) | 1.79 (0.31) | 1.99 (0.20) |
| **Mean distance to coordinating residues** [Å] | 2.04 (0.05) | 2.16 (0.19) | 2.01 (0.04) | 2.13 (0.18) |
| **B factor ratio. Ion to coordinating residues** | 1.20 (0.23) | 0.88 (0.18) | 0.79 (0.26) | 0.72 (0.23) |
| **B factor ratio. Ion to water molecules** | 0.67 (0.17) | 0.835(0.17) | 0.44 (0.18) | 0.68 (0.26) |

that it could vary from 0-1, rather than be related to that of the nearby ligand. In cases where `phenix.refine` has been used over REFMAC5, the default behaviour of refining the occupancy of any atom which has non-unitary occupancy was used.

Maintaining the appropriate distance relationship, for the residues and water molecules coordinating the metal ions necessitated the introduction of further distance restraints for the zinc ions. As the zinc ions were coordinated by water molecules, aspartic acid residues and phosphate groups, a 2.1 Å distance restraint was used, this is the median distance for a zinc ion coordination by aspartic acid residues [277]. These restraints had a variance of 0.2 Å to account for the different zinc coordination distances for differing bond types.

Table 2.1 shows that the refinements with zinc ions have a mean distance to the water molecules much closer to the ideal distance (2.10$\pm$0.13Å [277]) and zinc to aspartic acid (2.10$\pm$0.19Å [277]) than when magnesium ions are modelled: magnesium to water molecules (2.10 $\pm$0.08Å [278]) and magnesium to aspartic acid (2.05$\pm$0.1Å [279]). The inability to stably refine reasonable magnesium-water molecule distances along with the presence of positive electron density in the electron density difference map indicates, that the zinc ion models are likely more appropriate.

The B factor of the refined magnesium ions were significantly lower than the surrounding residues, whereas the zinc ions B factors were closer to that the of the surrounding

residues. The coordinating water molecules have higher B factors than the ions in all refinements.

Angle restraints could also be used to retain a good geometry of the coordinating atoms. In these examples these were not used, as the current restraints provided sufficiently acceptable geometry (visual inspection, Elliot Nelson & Joy Kristin Petrick).

Due to the initial uncertainty whether magnesium or zinc ions were bound, these restraints were applied by adding the local restraints to the restraints file generated by `giant.make_restraints`. This allowed for quick exploration as to whether the restraints and ion placement resulted in an appropriate structure. Further work to automate the application of metal ion restraints alongside the generation of the superposed model of a ground and bound state should explore adding angle restraints.

### 2.3.4 Discussion

Altering the refinement behaviour of metal ions allowed the superposed model to well represent the fragment screening datasets.

Metal ions in some of the fragment screening structures were still detected as outliers in some properties when utilising the CheckMyMetal server [273]. This was likely due to weak electron density evidence for some coordinating water molecules, therefore these were not modelled in some datasets. The CheckMyMetal server identified magnesium as a likely alternative, although the large disparities in electron density were not taken into account with this validation, which would give a strong preference for partly occupied zinc ions. Adding refinement of the occupancies of the coordinating water molecules may produce a more appropriate model.

## 2.4 Local real space correlation coefficient determined appropriate switch I loop models in a GEF/GTPase complex

### 2.4.1 Introduction

In a crystallographic fragment screen of a guanine exchange nucleotide factor (GEF) in complex with its corresponding GTPase, PanDDA event maps identified structural changes which could not be well explained by ligand binding alone.

Screening of a covalent electrophile library against the target, identified a dataset with a well resolvable ligand with an associated loop rearrangement. It was hypothesised that this loop rearrangement may be occurring in datasets where the electron density and PanDDA event maps could not be explained by a ligand model alone. The RSCC was used to assess the fit to density for different structural hypotheses. For each dataset three models were refined; one with the original loop position, one with the loop rearrangement, and one where these were combined in a superposed model. In the superposed model the occupancy of both loop models was constrained to sum to unity. It was confirmed that the loop rearrangement occurs with and without ligand binding but was most often seen when covalent electrophiles were bound nearby. Furthermore, it was shown that the superposition of two alternate loop states in some cases provides a better representation of the electron density than a model with a single loop position.

By identifying the relative occupancy and fit to density of the protein rearrangements, the model of the crystal became more interpretable. This allowed the crystallographer (Janine Gray) to prioritise hits and determine which further compounds to soak or attempt to co-crystallise.

#### 2.4.1.1 Targeting GTPase-GEF complexes

GEF/GTPase complexes can be targeted by interrupting the protein-protein interactions that support the formation of the complex. Alternatively, when GTPase is in complex with a GEF, the affinity for the nucleotide is reduced, allowing for direct targeting of the site. Other targeting strategies are discussed in detail in [280].

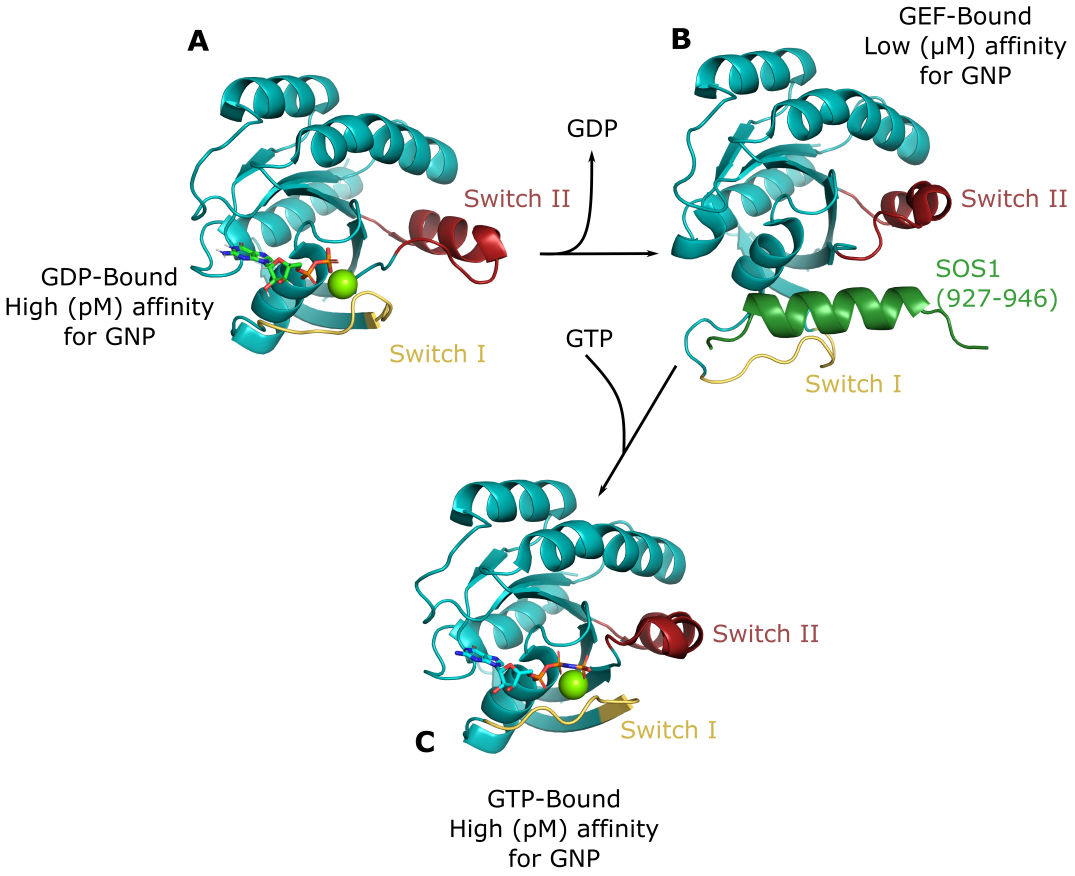



**Figure 2.2:** Switch I and switch II regions of guanine nucleotide exchange factor (GEF)/GTPase complexes are flexible regions which adopt different conformations during nucleotide exchange. **A** shows Ras GTPase bound to GDP (PDB code: 4OBE), this complex has a high affinity for guanine nucleotide phosphates (GNP). **B** shows nucleotide free Ras GTPase bound to SOS1 GEF. Note that the switch I loop has moved significantly compared to **A** (PDB code: 1NVU). **C** shows Ras GTPase bound to a GTP analogue GPPNHP (PDB code: 3L8Z). The switch I loop moves back. Switch I region shown as yellow, switch II region shown as red, GEF shown as green, magnesium ion shown as green sphere. Adapted from [280].

The Ras GTPases are a superfamily (over 150 family members) of guanine nucleotide dependent switches used in a large number of cellular processes [281]. Guanine nucleotide exchange factors (GEF) bind to the GTP bound form of the protein [282]. GTPase activating proteins stimulate hydrolysis of GTP to GDP [283], returning the GTPase to the inactive form. Guanine nucleotide dissociation inhibitors bind to the inactive form of the GTPase, changing its cellular location away from the cellular membrane so that GDP cannot dissociate [283].

Due to the large number of GEF and GTPases, there exists a variety of binding modes

between the partner proteins, however there are some core features, as shown in figure 2.2. The GEF associates with the GDP bound form of the GTPase forming a low affinity complex. The nucleotide dissociates leaving a high affinity GTPase/GEF complex. GTP binding converts the complex back into a low affinity state, subsequently releasing the GEF . The binding of the GEF to the GTPase induces conformational change of the switch I loop as well as stabilising the region by interacting with the switch II region [280].

The binding of the GTPase to the GDP or GTP is typically at pico-molar affinity, leading to classification of the interaction as undruggable. As opposed to ATP dependent kinases for which ATP analogues can out compete the typically micro-molar affinity ATP binding. Instead it may be viable to target the GEF/GTPase complex, either identifying chemical matter to interrupt the protein-protein interaction between the GEF and GTPase, or to directly target the guanine nucleotide binding site whilst in the complex.

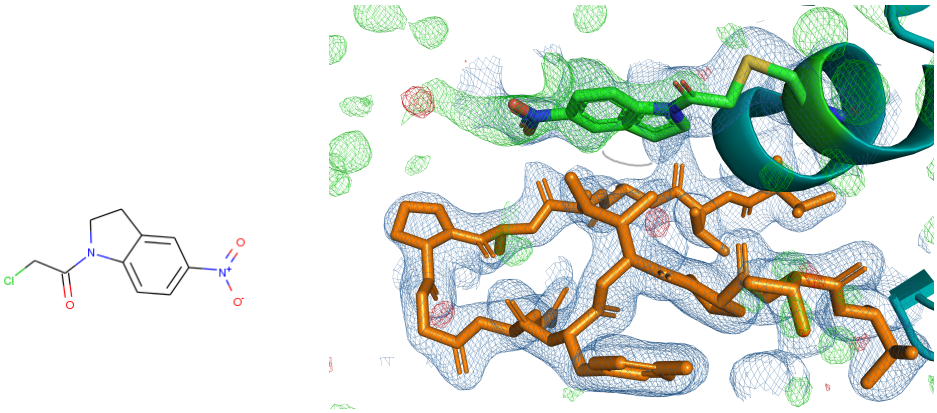
## 2.4.2 Methods

### 2.4.2.1 GTPase-GEF complex: Detecting a shift in the ensemble of states upon ligand binding

The target of interest was the guanine nucleotide exchange factor (GEF) Kalirin Dbl-homology 1 (DH1) domain in complex with the small GTPase Rac1, specifically holo-GDP structure (PDB accession code 5O33). The fragment screening campaign and follow up chemistry was undertaken by Janine Gray.

The aim of screening was to identify initial chemical matter against interactions mediated only when the complex is formed. Understanding any changes in the structure of the complex on ligand binding was essential to choosing between initial hits. Generally, the switch I and switch II regions were known to be flexible, and to adopt different conformations. This fragment screening confirmed that the Kalirin/Rac1 complex could adopt different structures with or without perturbation by a ligand.

A single hit in the guanine nucleotide binding pocket, which displaced the GDP, was identified from screening of the DSiP library. A collaboration with Nir London (Weizmann Institute of Science), was undertaken to screen covalent electrophiles binders using



**(a)** PCM-0102403-001 (**1**) **(b)** Covalent binding

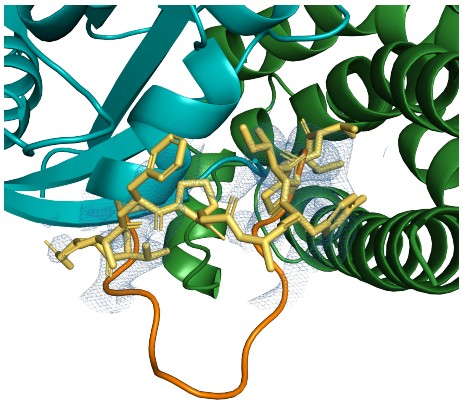
**Figure 2.3: (a)** A covalent binder (**1**) results in **(b)** loop rearrangement of switch I region in Kalirin/Rac1 complex. Rac1 is in teal, rearranged switch I loop position in orange. $2mF_{obs} - DF_{calc}$ electron density maps, in blue, contoured at 1 $\sigma$, within 2.5 Å of (**1**), cysteine 18 and the rearranged loop. $mF_{obs} - DF_{calc}$ electron density maps contoured at 3 $\sigma$, positive density is green, negative density is red.

intact MS. Crystal structures of covalent hits were obtained via co-crystallisation and were found to bind to cysteine 18 in the GDP binding active site.

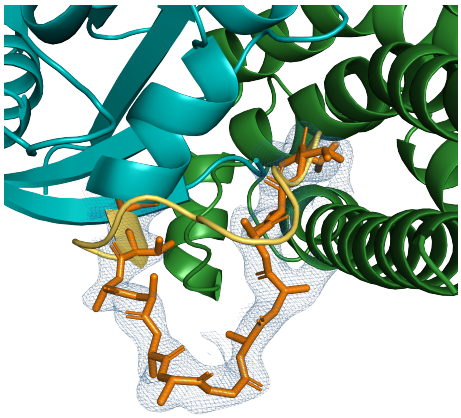
In this fragment screening campaign, an alternate switch I loop conformation was first noted on the binding of the covalent electrophile (**1**) as shown in figure 2.3(b).

The uninterpretable electron density, and PanDDA event density was in the region this loop rearrangement occurred. Therefore, it was hypothesised the loop rearrangement could explain the electron density. To determine if loop movement occurred, three models were refined: each loop position and a superposed model including both loop positions in a single structure.

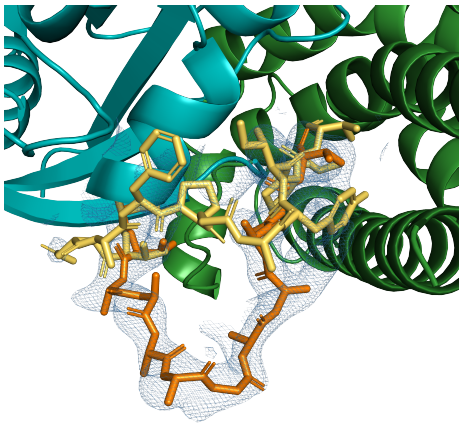
Refinement was undertaken using the `giant.quick_refine` script using `REFMAC5`. A complete occupancy group was defined to include all residues of the loop model. This set the occupancies of all the residues in the loop to be the same, and overall occupancy was constrained so that the occupancy of the residues in the original loop position and those in the rearranged loop position summed to unity. It was non-trivial to confidently classify whether the superposed refinement or each loop state refined individually was the more appropriate crystallographic model, using just the electron density and fitted model.



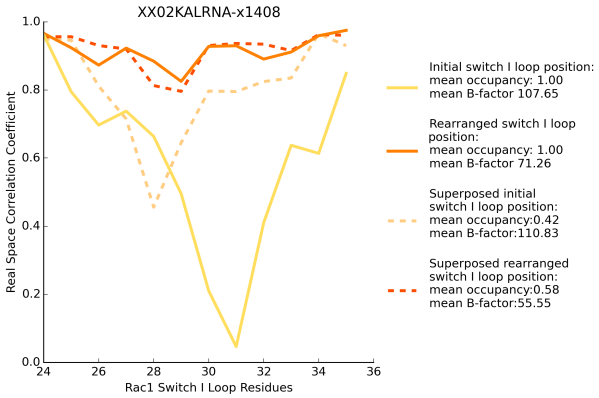
**(a)** Original loop position. Cartoon of rearranged loop shown, not in the refined model.



**(b)** Rearranged loop position. Cartoon of original loop shown, not in the refined model.

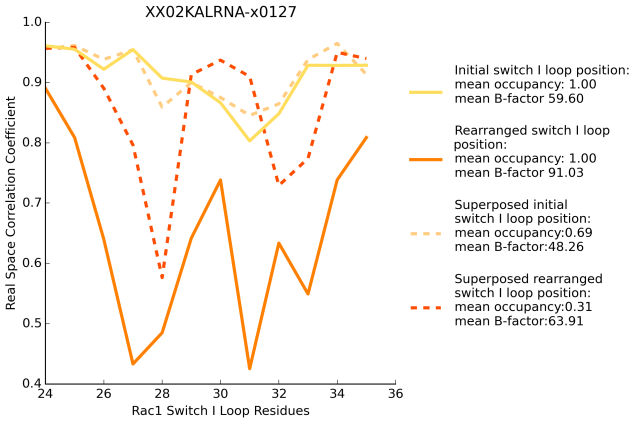


**(c)** Superposed model.

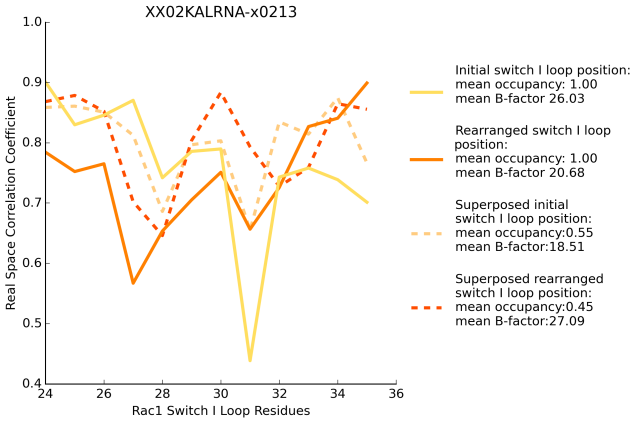



**(d)** Real space correlation coefficient of switch I loop residues.

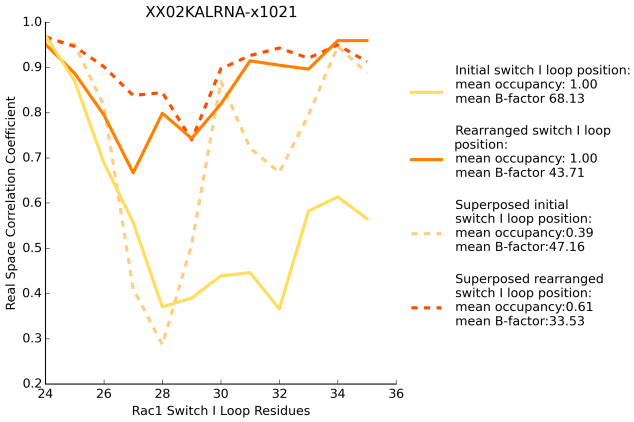
**Figure 2.4:** RSCC per residue identifies whether a model containing only the original loop position, containing only the rearranged loop position, or a superposition of both loop positions provides the best fit to the electron density local to those loop positions. Model colouring agrees with figure 2.2, where Rac1 is in teal, original switch 1 loop position in yellow, rearranged switch I loop position in orange, and Kalirin (GEF) is green. $2mF_{obs} - DF_{calc}$ electron density maps contoured at 1 $\sigma$ shown for XX02KALRNA-x1408. **(a)** shows the poor fit to density for the original switch I loop position when only the original residue positions are modelled. **(b)** shows the improved fit to density when a rearranged model for the switch I loop was fitted. **(c)** shows a superposed model, where the occupancies refined to 0.6 for the rearranged model, and 0.4 for the original model, showing that this dataset was most appropriately modelled with loop in the rearranged position only. **(d)** summarises the RSCC showing that refinement of a model with only the rearrangement loop position gives the best fit to density.




**(a)** Holo-GDP position of the switch I loop considered dominant by crystallographer.




**(b)** Neither switch I loop rearranged state or the holo-GDP position of the switch I loop considered dominant by crystallographer.




**(c)** Rearrangement of switch I loop considered dominant by crystallographer.

**Figure 2.5:** Assessing model quality of fit: RSCC of loop residues on a per residue basis. Colours correspond with figure 2.4, with rearranged loop in orange, and original loop position in yellow. The dashed lines represent the loop in superposition. Occupancies of the loop residues are shown in the legend. Note that the occupancy was refined using occupancy groups, such that the occupancy between residues is the same, and that sum of occupancies of both loop positions occupancy is one.

A tool for quantitative assessment of fit to density was required to aid the crystallographer. A tool was developed which presented the RSCC of each residue for the hypothesised models. The model included the switch 1 loop either in the original position, the rearranged position or a model where both positions were allowed but they were constrained to sum to an occupancy of 1 (the superposed model), as in figures 2.4(d) and 2.5.

**Table 2.2:** Mean R values for different refinements of switch I loop region. $R_{\mathrm{Work}}/R_{\mathrm{Free}}$

| **Model used in refinement** | **Crystallographers classification** | | |
|---|---|---|---|
| | Initial | Rearranged | Further information required |
| Initial | 0.250/0.0304 | 0.246/0.311 | 0.280/0.335 |
| Rearranged | 0.278/0.331 | 0.235/0.294 | 0.306/0.355 |
| Multiple | 0.251/0.323 | 0.230/0.301 | 0.290/0.349 |

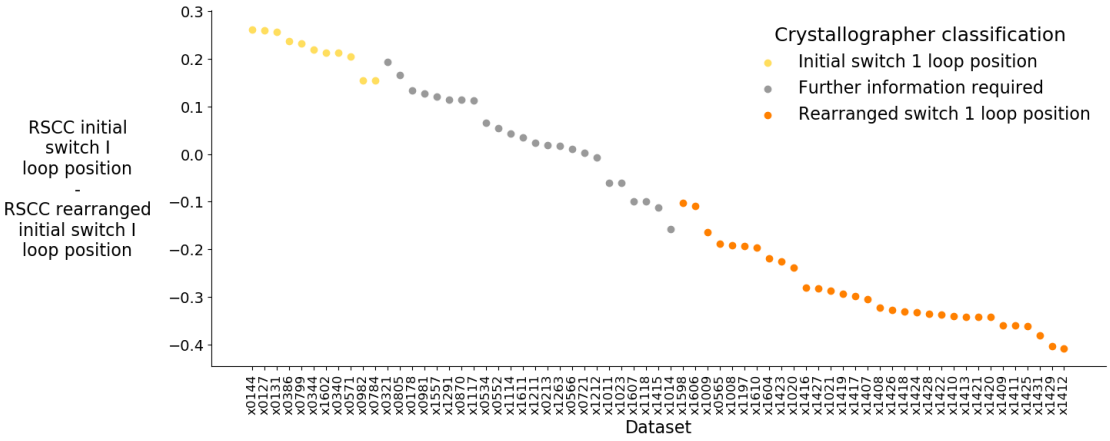
### 2.4.3 Results and discussion

The RSCC information presented per residue (as seen in figures 2.4(d) and 2.5) allowed the crystallographer to assign the models into three categories; original loop position dominant, rearranged loop position dominant or unable to classify the dataset.

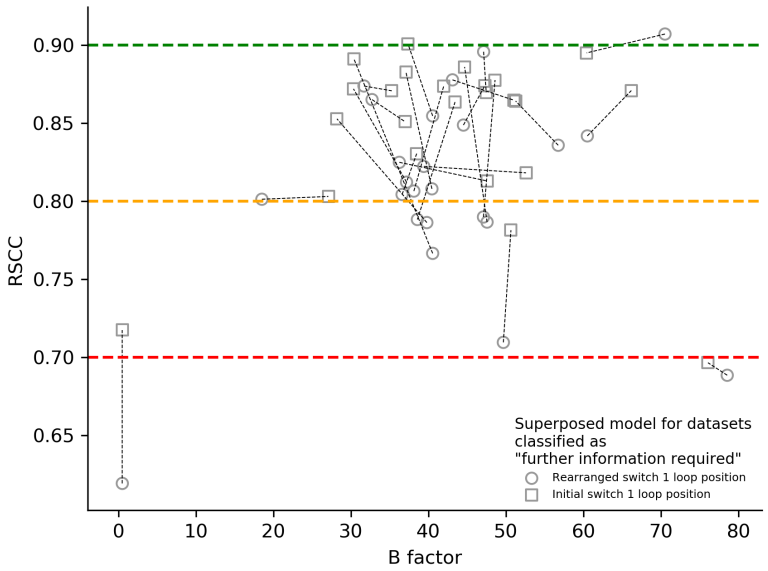
The RSCC information was summarised by taking the difference between the mean value of RSCC for refinement of the initial switch I loop position and the rearranged loop position, shown in figure 2.6(a). This single metric would have been sufficient to discriminate datasets which are well modelled by either non-superposed model.

In the cases where it was not possible to classify the dataset, the relatively high RSCC and the low B factor as shown in figure 2.6(b) indicates superposed models could be a reasonable modelling strategy in most of these intermediary cases. Figure 2.6(b) shows that in most cases both loop models have a similar RSCC and B factor, suggesting that both can explain the proportion of the electron density the occupancy implies that conformer explains.

However, figure 2.7 which shows the difference in RSCC between the loop positions, against the $R_{\mathrm{work}}$ value for the refinement which corresponds to the crystallographer's classification. Some of the models where the crystallographer was unsure which loop conformations dominated there is a $R_{\mathrm{work}} > 0.3$ implying that the datasets are of relatively poor quality. Table 2.2 shows that the mean $R_{\mathrm{work}}$ value is lowest when the classification matches the model refined, for models with either the initial loop position or the rearranged loop position. Utilising the superposed model sometimes results in a large $R_{\mathrm{free}} - R_{\mathrm{work}}$ value, indicating over-fitting. Therefore, the superposed model is not appropriate for some of these datasets.




**(a)** Classification of datasets difference in RSCC between refinements of each loop model.




**(b)** RSCC and B factor of datasets where the crystallographer was unsure whether a single loop position well explained the electron density.

**Figure 2.6: (a)** The difference between the mean RSCC of the loop residues, when modelled with either the initial loop position, or the rearranged position agrees with the crystallographer's classification from viewing the RSCC per residue and electron density of datasets. **(b)** shows the datasets where the crystallographer was unsure which loop model fitted the density, which are shown in grey in **(a)**. **(b)** shows the RSCC and B factor of the superposed model with the conformer corresponding to the original loop conformation (square) and rearranged loop conformation (circle). In most of these cases the RSCC of both models are close to one another, and above 0.8, a threshold for reasonable quality of fit to the electron density.

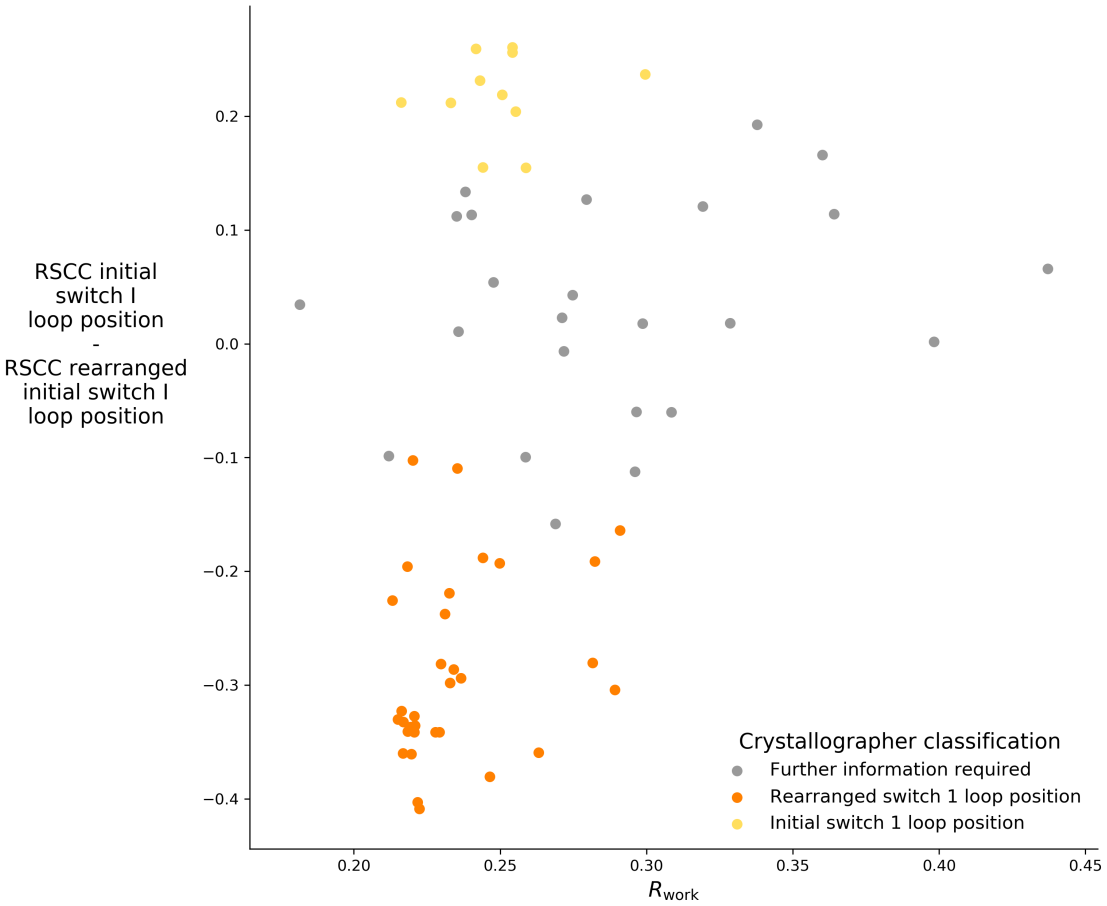



**Figure 2.7:** RSCC difference between loop positions versus $R_{Work}$ in refinement. The $R_{\text{work}}$ value for the refinement which corresponds to classification, such that a dataset classified as initial loop position is refined with that loop position. Datasets where the crystallographer was unsure which loop conformation was most appropriate, are shown with the superposed refinement.

The classification using RSCC allowed the crystallographer to focus on datasets where the ligand binding could be well modelled. Without explicitly modelling the rearranged loop, both covalent and non-covalent binders would not have been detected. It also directed experiments to identify for which compounds the loop rearrangement was consistently detected.

Using the superposed model as the ground state reference model for a PanDDA run, can be carried out if sufficient datasets which cannot be well modelled by either loop conformation are identified. This will likely require 20-30 datasets with loop rearrangement for a ground state model from which to detect outliers from using `pandda.analyse`.

## 2.5 Discussion and future work

Tools to help crystallographers to quickly assess hypotheses (such as loop rearrangement) across multiple datasets are essential for progressing fragment screening projects especially where it is clear a single model does not sufficiently explain the protein structure.

Metal ions can be well represented in the superposed modelling framework, but do require extra restraints when in the local environment of the bound ligand. Future work should include automation of the restraint generation for metal ions that are close enough to perturb atoms in the superposed state. This would require study of further test cases, including metal ions which are at different distances from the ligand or perturbation to the protein model. Tools such as `phenix.metal_coordination` should be tested and integrated with the superposed modelling framework.

As PanDDA is unbiased to whether the change to the protein is due to ligand binding, or stochastic changes in the protein, it is useful for detecting large scale protein motion that occurs only in a fraction of datasets. However, when these changes are detected they do not present as clearly to the crystallographer as a bound ligand can, as the event map density usually overlaps the protein model. RSCC was sufficient for classification of appropriate loop model, initially as a per residue metric, but also as an overall fit for the loop. This facilitated work towards identifying causes of the rearrangement. The real space z-observed difference may be preferred for checking fit to density, as it measures the difference density in that region of the model [201].

These methods of testing alternate hypothesis should be integrated into a user workflow: allowing rebuilding of models, and the ability to efficiently re-run `pandda.analyse` with a newly hypothesised model, with a subset of datasets to determine whether outliers to the new model correspond to ligand binding in an alternate protein conformation.

# 3

# Crystallographic refinement programs did not converge on occupancy for repeat datasets of protein-ligand complexes

## Contents


**3.1 Introduction . . . . . 101**
**3.2 Methods . . . . . 102**
3.2.1 Repeat datasets of protein-ligand complexes of fragment sized ligands . . . . . 102
3.2.2 Exhaustive search algorithm . . . . . 106
3.2.3 Sampling coordinate space . . . . . 108
**3.3 Results . . . . . 111**
3.3.1 Refinement programs do not converge on occupancy for weakly binding ligands . . . . . 111
**3.4 Discussion and future work . . . . . 115**


## 3.1 Introduction

As measuring binding affinity is not directly possible in crystallography, correlating crystallographic occupancy with affinity measured in biophysical techniques could allow prioritisation of hits. For ligands detectable in crystallographic fragment screening it can be difficult to measure the binding affinity in orthogonal biophysical techniques, as weak

affinities of fragments are often outside their detection limits [104].

Before any correlation between binding affinity and crystallographic occupancy can be determined, it is necessary to determine whether crystallographic occupancy is variable across repeat datasets of protein-ligand complexes. High variability in occupancy would not allow for any correlation to affinity. To determine the variability in occupancy of weakly binding fragment sized ligands, repeat datasets of protein-ligand complexes were collected using the XChem fragment screening pipeline.

Initially variability in occupancy was seen in repeat protein-ligand datasets of DCP2 refined by *REFMAC5* as can be seen in figure 3.6(a) . To understand whether the model used, refinement method or underlying crystal variability was the cause of occupancy variation, an algorithm that systematically searched all B factor and occupancy values for the ligand was developed. This exhaustive search algorithm determined which occupancy and B factor values minimise the difference density in the region of the ligand. This method gave similar distributions of occupancy as that seen when the models where refined in different refinement programs. The difference density in the region local to ligand bidning was normally a smoothly varying function of B factor and occupancy.

## 3.2 Methods

### 3.2.1 Repeat datasets of protein-ligand complexes of fragment sized ligands

To determine variability in crystallographic occupancy, repeat crystals of the same protein-ligand complex were prepared in as similar crystallisation conditions as possible. Three NUDIX hydrolase targets (DCP2, NUDT7 and NUDT22) within the SGC pipeline were chosen as model systems. Although there are structural similarities, the binding poses and chemical structures of the fragments are diverse. Therefore, no major bias due to choice of the same protein family for this study was expected.

#### 3.2.1.1 Protein preparation

NUDT7 purification was carried out as outlined in section 5.2.2.1. Apo-crystals for soaking were obtained as outlined in section 5.2.2.3.

**Table 3.1:** Nominal concentration of compound, measured as relative to the crystallisation drop volume, presuming no loss of volume since dispense used in repeat datasets experiments. Confidence is the assignemnt of confidence in the binding ligand pose attributed to the protein-ligand complex when first identified in crystallographic fragment screening.

| Compound | Protein | Concentration | Number of Crystals | Confidence |
|---|---|---|---|---|
| FMOPL000435a (**2**) | DCP2 | 100 mM | 24 | High |
| FMOPL00622a (**3**) | NUDT22 | 52.2 mM | 47 | Medium |
| N13725a (**47**) | NUDT22 | 52.2 mM | 15 | High |
| N13369a (**48**) | NUDT22 | 52.2 mM | 18 | Medium |
| N13663a (**49**) | NUDT22 | 52.2 mM | 29 | Medium |
| OX210 (**50**) | NUDT7 | 100 mM | 23 | High |
| NUOOA000181a (**51**) | NUDT7 | 20 mM | 15 | High |

Purification of DCP2 was carried out as outlined in appendix section B.1.1.1. Crystallisation was carried out as outlined in appendix section B.1.2.

Purification and crystallisation of NUDT22 are outlined in appendix section B.3.

#### 3.2.1.2 Crystal Soaking

Crystal soaking was performed by acoustic transfer using a Labcyte Echo 550 [108], with the compounds/solvent targeted away from the crystals and towards the edge of the crystallisation drop using the TeXRank image viewer [114]. 10-20% DMSO concentrations were used to give between 20 and 100mM concentration of fragments. The concentration for the compound series is given in table 3.1. DCP2 soaks had a mean soak time of 3hrs 32 minutes, NUDT22 had a mean soak time of 2hrs 22mins, NUDT7 had a mean soak time of 3hrs 33 minutes.

#### 3.2.1.3 Data collection

Data was collected at I04-1 at DLS, usually in automated queued data collection [109] following typical screening parameters for XChem fragment screening (at time of collection: 100% transmission, Omega Oscillations 0.12 °, 1500 images, 0.040 seconds per image, 1.5Å (maximal resolution determining distance from detector)). Diffraction data were automatically processed by software pipelines at DLS [109]. Preparation of ligand geometries (CIF generation) was carried out with AceDRG [2], eLBOW [5] or GRADE [4]. Initial refinement and map calculation were carried out with DIMPLE [6]. PanDDA [257]

was used for hit identification and model building was performed with Coot [3]. Initial structure determination steps were carried out within XChemExplorer [115].

#### 3.2.1.4 Compound selection

The compounds were selected based on the confidence that the hit was correctly modelled in the original fragment screen of that protein ligand complex. Hits were chosen that had both low and high confidence. The electron density can be seen in figure 3.6 and appendix figures D.1 and D.2. Table 3.1 gives the confidence assigned to the protein-ligand complexes with repeat datasets.

#### 3.2.1.5 Refinement with superposed models

The superposed model discussed in section 1.6 was used. The refinement of superposed models is undertaken with the `giant.quick_refine` script. The number of refinement cycles was increased to improve convergence of occupancy, as discussed in section 4.3.2. Refinement of superposed models by Buster had not previously been undertaken. This was implemented by extending the restraint generation framework within the `giant` library distributed with PanDDA.

##### 3.2.1.5.1 `REFMAC5` refinement with superposed models

Default settings for `giant.quick_refine` script using `REFMAC5` were used, except for extending the length of refinement to 50 cycles, ensuring occupancy refinement convergence, as discussed in section 4.3.2.

##### 3.2.1.5.2 Phenix refinement with superposed models

The `giant.quick_refine` script was used, with 20 cycles, ensuring occupancy convergence in refinement as discussed in section 4.3.2.

##### 3.2.1.5.3 Buster refinement with superposed models

To compare occupancy refinement in commonly used refinement programs, I implemented superposed refinement using Buster. This was achieved by extending the `giant.quick_refine` script with an implementation of the restraint generation script

`giant.make_restraints`. For this occupancy group restraints and distance restraints, were introduced. The distance restraints are between identically positioned conformers of the changed state atoms introduced to make the model consistent. These are set to be coincident with a distance of zero, and allow the multi-conformer model that is generated using `giant.split_conformations` to be used with Buster refine.

The occupancy group restraints define groups of atoms which should have the same occupancy, and how those groups of atoms should add to unitary occupancy. i.e. the ground (un-changed) state is represented by conformers `A` and `B`, the bound (changed) state is represented by the ligand, and any other alternate conformations, or additional water atoms belonging to conformers `C` and `D`. As the bound and unbound states should represent all possibilities in the crystal, their occupancy should sum to unity. Therefore an occupancy group involves a constraint such as:

$$\underbrace{\text{Occupancy}(A) + \text{Occupancy}(B)}_{\text{Ground State}} + \underbrace{\text{Occupancy}(C) + \text{Occupancy}(D)}_{\text{Bound State}} = 1.0 \qquad (3.1)$$

In Buster *refine* this can be achieved by defining the set of conformations that any residue or ligand can take using a parameter file in which the occupancy of the atoms in each conformer are defined using the `BUSTER_SET` keyword. To constrain the sum of occupancy of multiple groups to 1.0 the `BUSTER_OCCSUM` keyword is used. A representative atom from each group is placed with the `BUSTER_OCCSUM` constraint, this is then applied across all atoms of the groups.

Refinement was carried out using the methods outlined for superposed refinement as in section 3.2.1.5, and non-superposed refinement as in section 3.2.1.6.

#### 3.2.1.6 Refinement with non-superposed models

Protocols for non-superposed refinement were implemented using `REFMAC5`, `phenix.refine` and Buster `refine`. The model used for refinement in the non-superposed method is the single conformer model that can be produced from the superposed refinement methods using the `giant.split_conformations` script.

#### 3.2.1.6.1 REFMAC5

Occupancy restraints were generated by placing the ligand in an incomplete occupancy group. This allows its occupancy to vary from 0-1 in an unconstrained manner. The number of refinement cycles was set to 50, to ensure occupancy convergence as discussed in section 4.3.2.

#### 3.2.1.6.2 Buster

As there were atoms at below unitary occupancy which need refinement, the autoBuster script `pdb2occ` was used to generate suitable occupancy restraints. Otherwise the refinement was undertaken as per default parameters in Buster refine. Convergence of refinement was achieved according to the internal metrics reported in the Buster log. Therefore, the number of refinement cycles was kept as default, 4 cycles. Note the number of cycles is not directly comparable between programs, as each refinement program defines the work done in a refinement cycle differently.

#### 3.2.1.6.3 Phenix

As `phenix.refine` implements occupancy refinement of non-unity occupancy atoms by default, this refinement was applied as default except extending to 20 cycles to ensure refinement convergence, as discussed in section 4.3.2.

### 3.2.2 Exhaustive search algorithm

Initial repeat datasets of DCP2 with compound (**2**), figure 3.2(d), refined using `REFMAC5` gave a wide distribution of occupancy values, as seen in 3.6(a), salmon line. To identify whether the wide distribution was caused by the refinement method or crystal variation, a method which systematically searches B factor and occupancies for the ligand was developed, which is outlined in figure 3.1.

The superposed model of the protein-ligand complex, as described in section 1.6, is used within the exhaustive search algorithm. Residues which are involved in the superposed state are defined to be those within 5 Å of the ligand atoms and have an

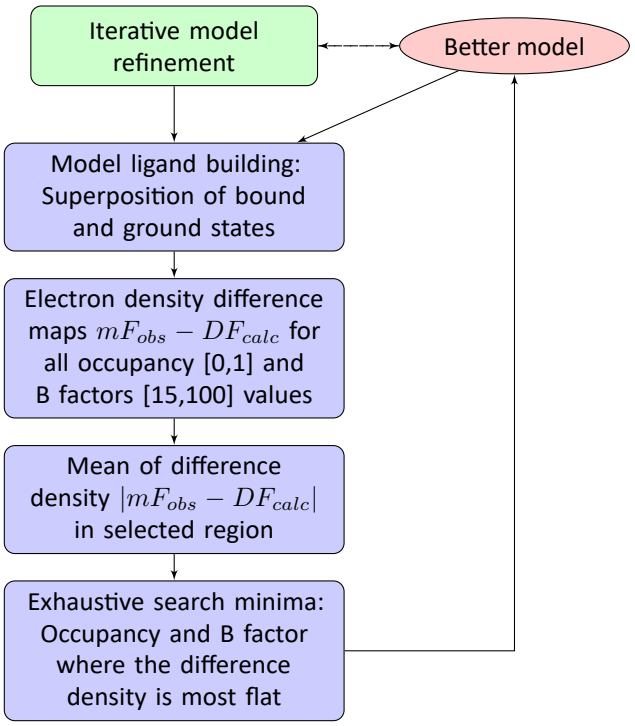



**Figure 3.1:** Flowchart describing exhaustive search algorithm.

alternative conformation within the superposition. Newly added or displaced hetero-atoms within this cut-off distance are placed in the ground state of the superposed model. For example, water molecules that are displaced by ligand binding.

Multiple separated regions which vary in occupancy, i.e. two distal copies of a ligand, these are currently considered to be involved in the same superposition and combined to create a single ground state and bound state.

The search of parameter space involves looping over occupancy, from 0-1, and B factor from 15-100Å$^2$ (using $u_{iso}$). This range of B factor is chosen as this covers the reasonable range for most well refined ligands, as at medium resolution, atoms with B factors above 100Å$^2$ have small or negligible contributions to scattering. A lower B factor value can be searched if the local protein environment has rigid and well-defined residues with a low B factor. For each combination of B factor and occupancy, the occupancy of the atoms in the ground state was set to $X$ and the occupancy of the changed state to $1 - X$. The B-factor of all atoms in the superposed state was set via $u_{iso}$. At each increment, a difference electron density map, $mF_{obs} - DF_{calc}$, is generated.

A minimum boundary (convex hull) is then defined around each set of atoms involved in the superposed state: the volume around each atom coordinate is sampled within 0.77 Å, the atomic radius of a carbon atom. For all the atoms within a local superposition, points within the volume around the sampled atoms are used to form a convex hull. This sampling method ensured that an isolated atom, such as a water molecule is sufficiently sampled, without sampling an overly large proportion of the coordinate space. The points in any separated convex hulls were all considered together.

Over the points defined to be in the convex hull(s), the average flatness of the difference density is calculated:

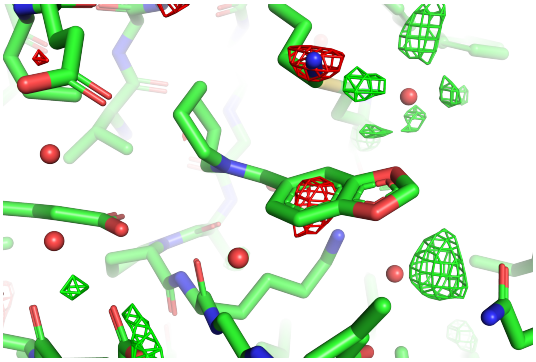
$$\text{Average flatness of difference density map} = \frac{\sum |mF_{obs} - DF_{calc}|}{\text{Number of points in convex hull}} \quad (3.2)$$

Figure 3.2(a) shows the difference density map for a protein-ligand complex at the minima as determined by the exhaustive search algorithm. Comparing this to the maps generated away from the minima in figures 3.2(b) and 3.2(c) it was visually evident that there is a much lower amount of difference density across the ligand and surrounding residues, when at the minima. Whether fixing the B factor distribution to a single value is appropriate is discussed in section 3.4.

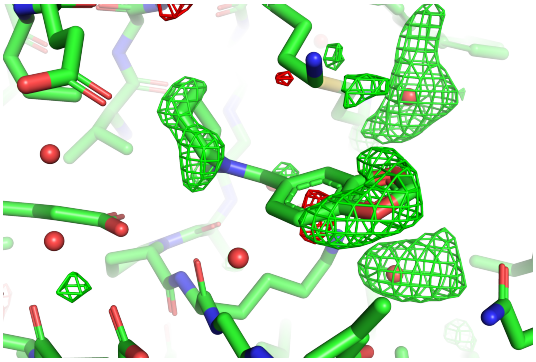
The occupancy and B factor values where the mean absolute difference density was minimised was taken to be the optimal B factor and occupancy for the atoms involved in the superposed state. As the function is normally smoothly varying near the minimum point, as shown in figure 3.3 the exhaustive search algorithm is run in larger increments, followed by sampling near the minima. By default, the step size in $u_{iso}$ (equivalent to varying the isotropic B factor) and occupancy was 0.05. This was then followed by running within $\pm$0.2 of the minima in steps of 0.01 for both $u_{iso}$ and occupancy.

### 3.2.3 Sampling coordinate space

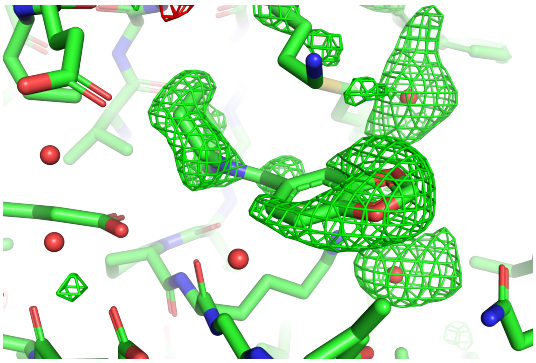
Choosing where to sample the flatness of difference density is important, as this determines whether just the ligand, or the local environment of that ligand is also investigated. Sampling methods included: just sampling the atom coordinates that are specified,



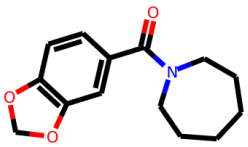
**(a)** At minima: Bound occupancy: 0.61, B factor: 32 Å$^2$



**(b)** Away from minima: Bound occupancy: 0.3, B factor: 43 Å$^2$

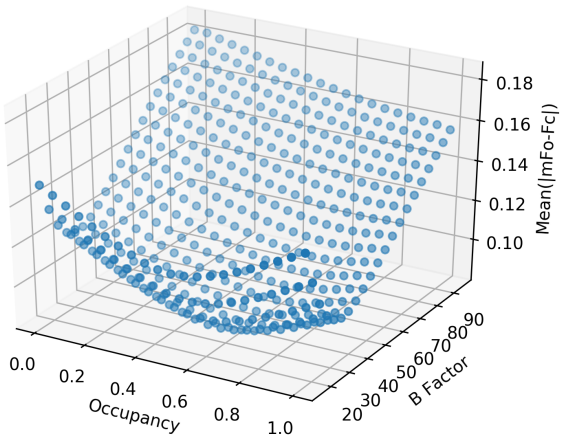


**(c)** Away from minima: Bound occupancy: 0.1, B factor: 71 Å$^2$

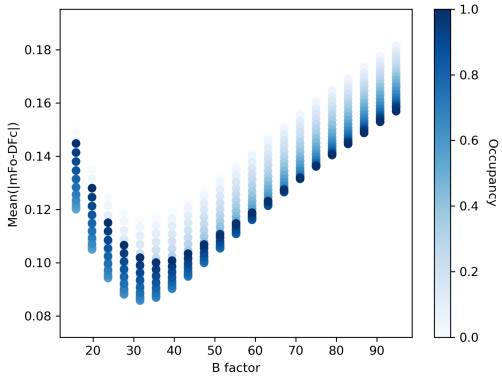



**(d)** FMOPL000435a (**2**)

**Figure 3.2:** Examples of how occupancy affects visible evidence in difference density maps ($mF_{obs} - DF_{calc}$). Contoured at 1 $\sigma$ near ligand complex DCP2 in complex with (**2**) (DCP2B-x0146), for different occupancy and B factor values at and away from the minima found by exhaustive search.




**(a)**




**(b)**

**Figure 3.3:** Exhaustive search produced a smoothly varying function around the minima. Minimum for DCP2B-x0146, shown in three-dimensional plot (a), and the corresponding 2d plot (b). The minimum of difference density was when the bound state was occupied at 0.6 with a B factor of 32 Å$^2$

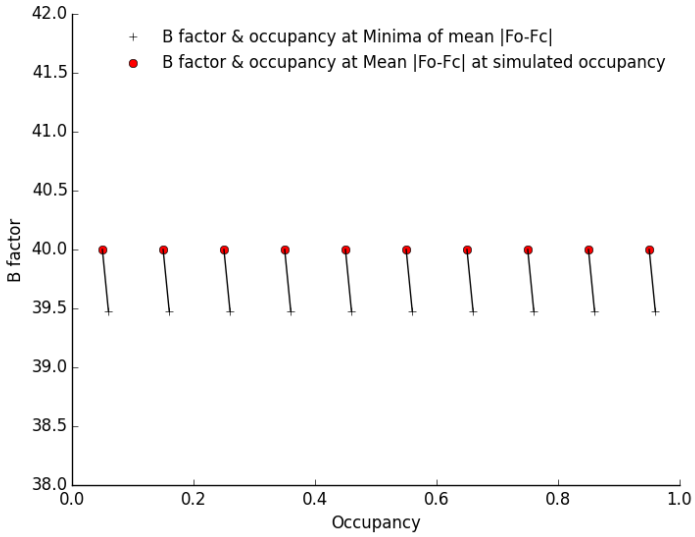



**Figure 3.4:** Exhaustive search algorithm was validated against synthetic data. For each occupancy a run of the exhaustive search algorithm was undertaken. The small disparities were likely due to the program internally representing B factor in terms of $u_{iso}$ and taking finite steps in both $u_{iso}$ and occupancy.

sampling at all points within a box bounded by the maximal extent of the superposed state, and various buffer distances around convex hulls around the atoms involved in the superposed state. These were tested with the validation script (section 3.2.3.1). The current convex hull parameters were chosen as they represented a good sampling of the regions of interest in the electron density, without sampling unimportant areas.

#### 3.2.3.1 Validation of the exhaustive search algorithm

Validation on simulated data was performed to demonstrate that the minima can correctly identify the occupancy and B factor, given the supplied density is calculated from the model rather than experimental data. The underlying structure was a known protein structure from fragment screening, built as a superposed model. The occupancy and B factor of the ligand and changed ground state atoms were set giving a test set with many similar structures to minimise B factor and occupancy to different values. For each occupancy and B factor value, an electron density map was generated. The exhaustive search method was run for each simulated occupancy and B factor and it was checked that the correct occupancy and B factor values were returned. Figure 3.4 shows the

exhaustive search algorithm can recapitulate the occupancy and B factor of the simulated data, (only one B factor value is shown for clarity). The small disparity was due to the step size in sampling.

## 3.3 Results

### 3.3.1 Refinement programs do not converge on occupancy for weakly binding ligands

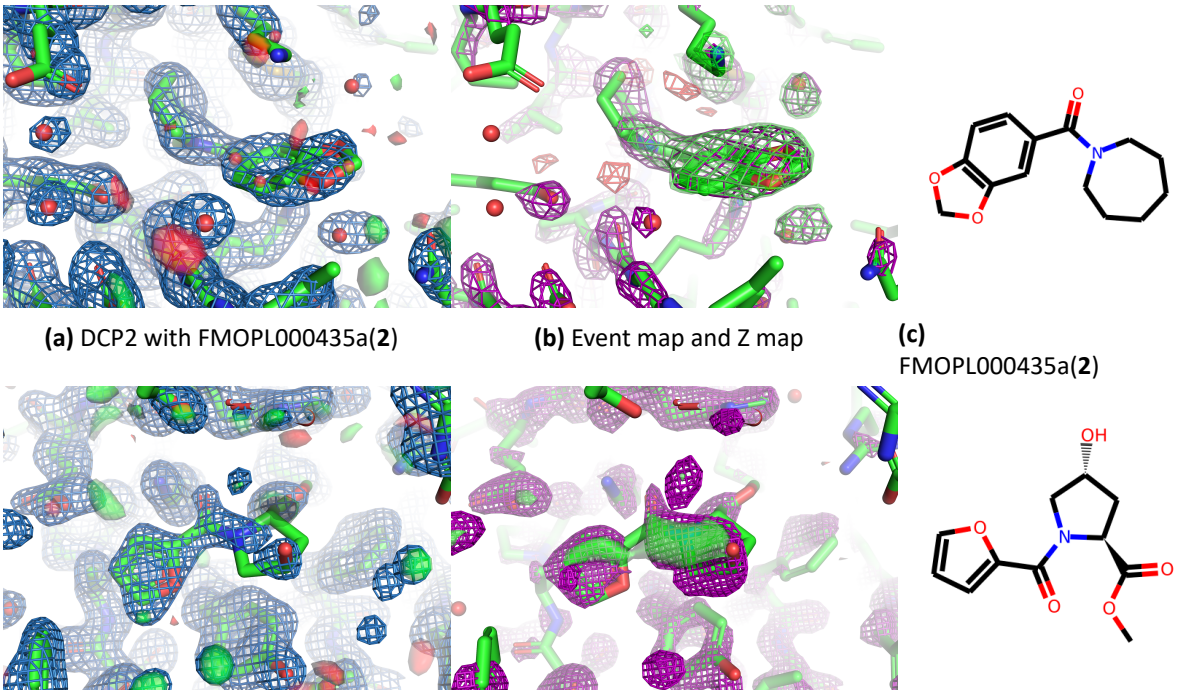
There was little consensus between refinement programs on refined occupancy for repeat datasets of protein-ligand complexes.

Results from three targets (NUDT22, NUDT7 and DCP2) with 6 ligands are discussed. Note that one of the protein-ligand complexes (NUDT7 with OX210 (**50**)) had two copies of the ligand bound close together.

Both high and medium confidence ligands were investigated, it was hypothesised that the higher confidence ligands would have a tighter distribution of occupancies. Figure 3.5(a) shows an example of a high confidence ligand where there is contiguous electron density for the whole ligand in the $2mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$ map. This example has clear event map density, as shown in figure 3.5(b), present for the whole ligand. The outlying electron density across multiple datasets, as shown by the Z map also covers the majority of the ligand, a sign of high confidence. A lower confidence example is given in figure 3.5(d) and figure 3.5(e), there the Z map and event map are not contiguous for the whole ligand.

Note that confidence was assigned by visual inspection of the event maps, as well as the corresponding refined $2mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$ maps. Table 3.1 summarises the confidence of the protein-ligand hits. Appendix figures D.1 and D.2 show the electron density evidence of an example dataset for each protein-ligand complex outlined in table 3.1.

For each protein-ligand complex, refinement was carried out using both a superposed model, and with a model of the protein ligand complex which does not use ground and bound states. When the superposed model is used, the occupancy of the bound ligand and any new conformations of nearby protein residues are constrained to sum to unity, with the water molecules or ligands which the ligand replaces. As throughout this thesis,



**(a)** DCP2 with FMOPL000435a(**2**) **(b)** Event map and Z map **(c)** FMOPL000435a(**2**)

**(d)** NUDT22 with FMOPL00622a(**3**) **(e)** Event map and Z map **(f)** FMOPL00622a(**3**)

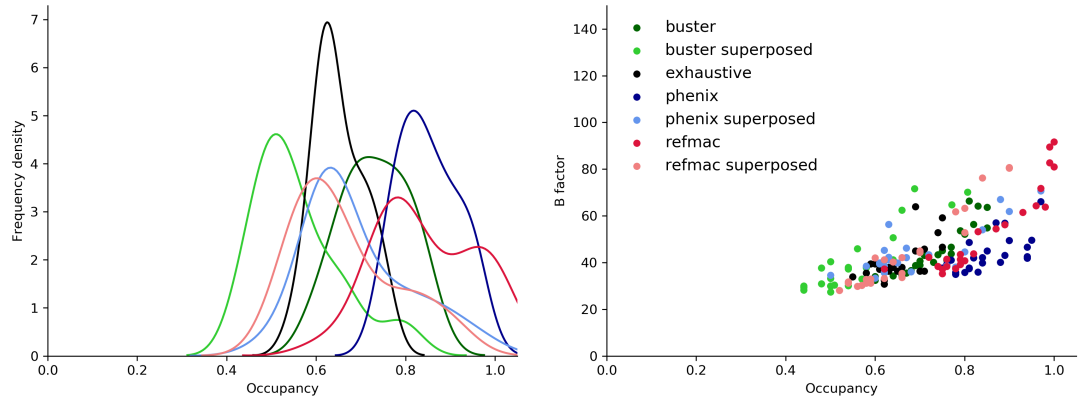
**Figure 3.5:** Electron density evidence for repeated datasets of protein ligand complexes. Compound (**3**) has a lower confidence than compound (**2**) as the event map is less well defined. $2mF_{obs} - DF_{calc}$ electron density maps contoured at 1 $\sigma$ shown as blue mesh, $mF_{obs} - DF_{calc}$ electron density maps contoured at 3 $\sigma$, shown as surfaces, green positive density, red negative density. Event maps contoured at twice the background density correction value, Z maps countered at $\pm 4.0\sigma$, shown as surface (**e**) and mesh (**b**), green positive density, red negative density.

`REFMAC5`, `phenix.refine` and buster `refine` have all be used, to determine whether there are any systematic differences between refinement programs.

Figure 3.6(a), shows the distribution of ligand occupancy values, there was a relatively wide distribution of occupancies for DCP2 in complex with compound (**2**). This was the case for all refinement methods, with both the superposed and non-superposed models. Therefore, there was no consensus agreement on an occupancy value. Values below 0.4 occupancy were not sampled by all methods. As evidenced in figure 3.6(b), higher occupancy refinements had higher B factors.

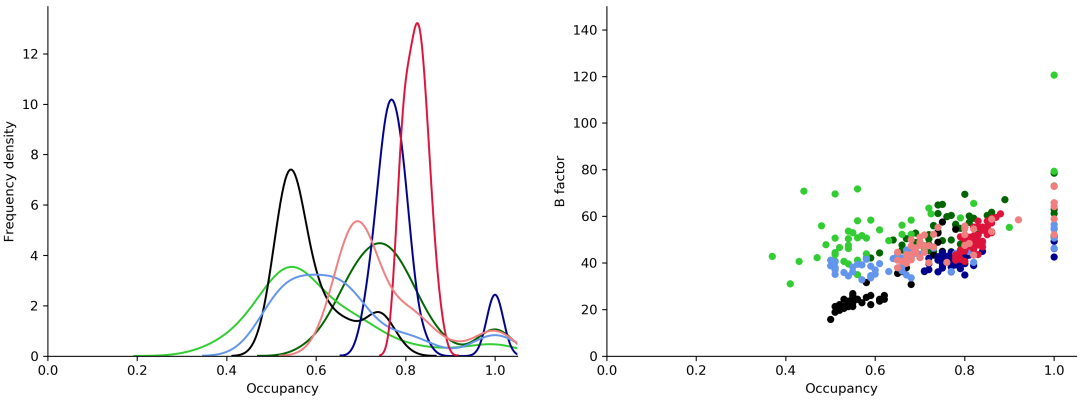
In figure 3.6(a), the non-superposed methods have a higher occupancy for the ligand than the superposed method, as the superposed methods are constrained to occupy the ground state structure.

In general, the ligand occupancy was widely distributed for the repeat datasets




**(a)** Occupancies: DCP2 with FMOPL000435a. 24 samples (**2**)

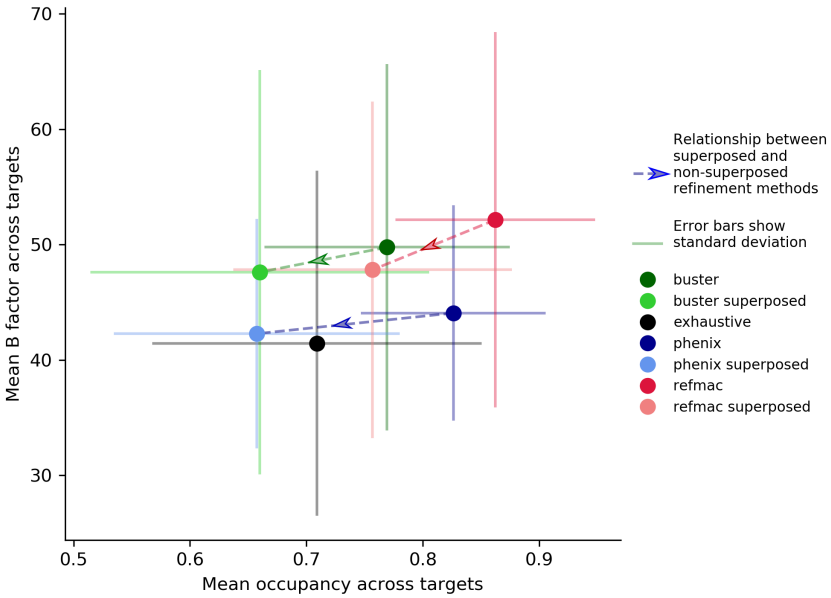
**(b)** Occupancy and B Factor: DCP2 with FMOPL000435a (**2**)



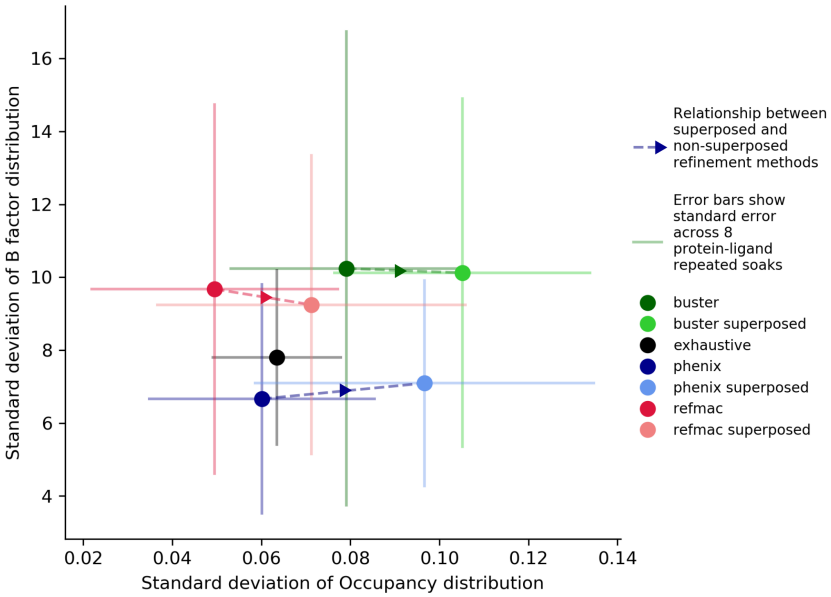

**(c)** Occupancies: NUDT22 with FMOPL00622a. 47 samples (**3**)

**(d)** Occupancy and B Factor: NUDT22 with FMOPL00622a (**3**)

**Figure 3.6:** Refinement programs did not agree on occupancy for repeat datasets of protein-ligand complexes. Ligand occupancies and B factors compensate for one another in repeat datasets of protein-ligand complexes. Occupancies for the ligand were quite widely distributed. NUDT22 with compound (**3**) had tightly distributed occupancies in Phenix and REFMAC5 but was broader in other refinement methods. NUDT22 also had datasets which did not refine away from unitary occupancy. Superposed models had a lower occupancy for the ligand, as the ground state is being occupied. **(a)** and **(c)** show the occupancy distributions. **(b)** and **(d)** show the occupancies and B factors. Colours indicating refinement programs are consistent across figures.




**(a)** Mean Occupancy and B factor across targets




**(b)** Mean width of occupancy and B factor distributions across targets

**Figure 3.7:** Superposed refinements yield lower occupancy for the ligand, but the actual values and spread of parameters vary greatly by program. Lower B factors compensated for the reduced occupancy in the superposed methods. Datasets with higher occupancies had a narrower distribution of occupancies. Superposed models had wider distributions of occupancy. Exhaustive search had the lowest standard error, suggesting these distributions are most similar across targets.

of protein-ligand complexes (figure 3.6 and appendix figures D.3 and D.4). This is summarised in figure 3.7(a), which shows for all refinement methods there is wide distribution of occupancies and B factors. Note refinements which utilise the superposed model typically have a lower mean ligand occupancy across all targets. This is not surprising as there is always the presence of a ground state model. Figure 3.7(a) also shows that the refinements with superposed models also have a lower B factor than the equivalent non-superposed counterpart. It appears that, lower B factors compensate for the reduced ligand occupancy in the superposed methods. This trend can be seen in figure 3.6(b), for most refinement methods the B factor of the ligand increases as the occupancy increases. Note that the strength of this trend varies between targets and method, as can be seen by the relatively wide spread of B factor values for Buster refine applied to the superposed model in figure 3.6(d).

For NUDT22 with compound (**3**), the distributions of occupancies were relatively narrow for `REFMAC5` and `phenix.refine`, but remain wider for the other methods, as in figure 3.6(c). This protein ligand complex had 47 samples, so the relatively narrow distribution was indicative that the crystals were quite similar despite the differences between refinement programs. A few datasets had unitary ligand occupancy, which indicates that refinement did not occur due to a poorly defined model, or incorrectly appliedestraints in those datasets.

Figure 3.7(b) shows the standard deviations of the occupancy and B factor distributions. This was narrowest for `REFMAC5`. Methods which had higher mean occupancy had narrower distributions of occupancy, likely as they assign less uncertainty to the occupancy parameter. Superposed models had wider distributions of occupancy. The exhaustive search method showed the most consistency in distribution width across targets, as shown by the lowest standard error in both B factor and occupancy.

## 3.4 Discussion and future work

Applying multiple refinement methods to repeated datasets of protein ligand complexes, results in widely varying occupancy distributions between refinement programs. This is likely due to the ability for most refinement programs to attribute uncertainty in the

position and occupation of the site into both occupancy and B factor parameters, as they are correlated. Some methods did show narrow distributions of occupancy suggesting that the variability in occupancy and B factor may be due to a poor understanding of the best structural model and corresponding model parameters to use rather than crystal variation.

Other factors which affect the electron density maps when soaking crystals with compounds include compound solubility and crystal permeability. As this study looked to explore behaviour of weakly binding ligands with any crystal system these parameters were not intentionally varied.

The exhaustive search method has confirmed that in most cases the B factor and occupancy for a ligand creates smoothly varying function near the minima. In the exhaustive search method it was chosen to fix the B factor across the ligand to the same value. Therefore, the B factor distribution across the ligand is different to that which is obtained during refinement. This was initially done to reduce the number of samples to check, as this is a brute force sampling procedure and check every B factor would have been computationally intensive. This shortcoming was not addressed as it became evident that the B factor and occupancy curve was smoothly varying, and that the results from exhaustive search were similar to refinement methods. Therefore I believed it more important to determine whether crystallographic occupancy could be measured in an orthogonal technique (chapter 5).

Although a further brute force search of the ratio between B factors within the ligand could be done, continuity of local B factors would imply that B factors should be kept relatively close to the atoms neighbouring B factor. If instead the refined distribution of B factors for the ligand was used, and then exhaustively searched different occupancies with this new B factor distribution, one could then see if this leads to a further improvement in the flatness of difference density. If this was not considered sufficient, a search of similar but slightly different B factor distributions could then be carried out.

One difficulty is that the ligand likely should have a higher flexibility than the surrounding protein, but the extent of this increased flexibility is not easily estimable. This is partly because it is difficult to decouple flexibility and positional error. Modelling which degrees

of freedom the ligand would lose when constrained in the protein pocket is important for determining its overall flexibility. Estimating this may require techniques such as locally constrained MD, not currently implemented as part of determining stereochemical information supplied in the ligand CIF.

Applying the knowledge gained to determine an ideal refinement procedure for occupancy one strategy would be to apply B factor continuity restraints between the protein and ligand. These could be introduced as continuity restraints to atoms in closest contact, or those acting as H-bond acceptors or donors. This would be applied in later stages of refinement, when the ligand's position is less likely to change. The other regions of the ligand could then be allowed to vary in B factor, using restraints of riding motion, as outlined in section 1.3.13.5. The existing formulation of restraints on close (bonded, angle-related or proximal) atom-pairs could be used to this. The ability to define these for proximal atoms was not known to the author when the work was carried out, and thus they were not applied.

A study relating crystallographic occupancy to the size of crystals found that crystal size and soaking time were major factors for the crystallographic occupancy of the ligand [256]. In their study the B factor of the ligand was fixed to obtain a consistent refinement of occupancy. Crystals size controls could be applied more carefully to the repeat datasets of protein-ligand complexes outlined here. However, as the study showed that full occupancy was reached within a maximum of 15 minutes, and as we incubated with the protein for greater than an hour, we expected maximal occupancy to be reached in these experiments.

The work outlined in this chapter did not determine whether variation in occupancy was due to crystal variation or inadequate model parameters being provided to refinement programs . I hypothesised it would be possible to validate crystallographic occupancy using intact MS. Chapter 5 explores the relationship between crystallographic occupancy and intact MS of crystals to measure the amount of compound in the crystals. The amount of compound determined by intact MS correlates with occupancy in all refinement methods. But as in this chapter refinement methods do not agree, and low occupancy values are not sampled. However, by restricting the ligand B factors to

be similar to the B factors of the surrounding protein residues, then crystallographic occupancy more closely matches the ratio of compound determined by MS.

Given that the exhaustive search algorithm did not uniquely identify the occupancy and B factor of a ligand, an alternative target value, such as the RSZD metric may be more appropriate. This would more harshly punish missing atoms, or over occupied atoms in the model, due to its maximal sub-sampling method.

In general, finding an appropriate refinement strategy, should focus on developing appropriate metrics, such as an information criterion to assess whether more complicated models with an increased number of parameters for the protein model are justified, as outlined in section 4.4.

Overall, there is currently enough variation in occupancy refinements for weakly binding fragment sized ligands that a crystallographer should only be concerned if a surprisingly low occupancy (below 0.4) is obtained during refinement. Although the occupancy should be reported for the ligand, in the context of its refinement, the value will not be comparable between methods, and may not be that informative. Chapter 5 shows that the occupancy value does become informative when the occupancy is calibrated with an orthogonal experimental technique and the B factor of the ligand is fixed.

# 4

# Crystallographic refinement did not model weakly binding fragment ligands with a low occupancy

## Contents


**4.1 Introduction . . . . . 120**
**4.2 Methods . . . . . 121**
4.2.1 XChem database . . . . . 121
4.2.2 Refinement methods . . . . . 122
4.2.3 Refinements were run using distributed task scheduling . . . 122
4.2.4 Preparation of datasets and refinement by multiple methods . 123
**4.3 Results . . . . . 127**
4.3.1 Existing refinements showed that low occupancy states are uncommon . . . . . 127
4.3.2 Longer refinement increased the convergence of occupancy . 127
4.3.3 Occupancy distribution varied drastically by refinement method, unlike the B factor distribution . . . . . 129
4.3.4 Superposed methods consistently had lower ligand occupancy 132
4.3.5 Exhaustive search showed similar behaviour to refinement of superposed models . . . . . 133
4.3.6 Ligands were sufficiently supported by density to be a representative subset . . . . . 134
4.3.7 PanDDA background density correction calibrated by refined occupancy of XChem hits . . . . . 138
**4.4 Discussion and future work . . . . . 139**

## 4.1 Introduction

As reported in chapter 3 there was a relatively wide distribution of occupancies across repeat datasets of protein ligand complexes. Additionally, refining the models in different refinement programs led to different distributions of occupancies. Occupancies below 0.4 were rarely modelled. Therefore, to discover whether models are parametrised such that low occupancy values are infrequently sampled across in many protein-ligand complexes, the existing XChem fragment screening data is analysed.

Initially the occupancies and B factors of existing refinements were analysed. As this only represented `REFMAC5` run on superposed models re-refinements using `giant.quick_refine`, and non-superposed refinement and the exhaustive search algorithm, outlined in section 3.2.2 were applied to all datasets.

Initially, the occupancies and B factors of existing refinements were analysed. However, this only represented *REFMAC5* run on superposed models. For each dataset superposed models, and non-superposed models were generated, and refinement was carried out using `REFMAC5`, `phenix.refine` and Buster `refine`. The exhaustive search algorithm was also applied to all datasets.

The occupancy of the ligand in the protein-ligand complex was not normally refined to the same value when using different refinement programs. All tested refinement programs had very few protein models with bound ligands that refine with low occupancy (1st percentile $0.30 \pm 0.15$ across 7 combinations of model and refinement method). Despite a high variation in distribution of occupancy, the B factor distribution of ligands varied much less. This is likely due to restraints for adjoined or nearby atoms to have smoothly varying B factors.

By comparing the occupancy of ligands in models which were refined with and without an explicit ground state, explicitly modelling the ground state consistently reduced the occupancy of the ligand. This reduction is similar in magnitude across different refinement programs suggesting the ground state is stably refined.

## 4.2 Methods

### 4.2.1 XChem database

The XChem fragment screening platform at DLS has collected over 100,000 crystals mainly from primary fragment screening campaigns. This data therefore includes a large number of fragment binding hits, which have been at least initially modelled using the superposed modelling framework introduced alongside PanDDA [202]. Work by Rachael Skyner, (code accessible at `https://github.com/xchem/pipeline`), to agglomerate the data processed by XChemExplorer into a PostgresSQL database provided the basis for the analysis outlined here.

The code used to undertake this analysis is provided at `https://github.com/nelse003/parse_xchemdb`.

#### 4.2.1.1 Preparation of protein ligand models in the XChem database

As the data used in this chapter comes from fragment protein-ligand soaking experiments at the XChem facility, the modelling and refinement process is briefly outlined here. At the XChem facility, an initial model processed by DIMPLE [6] is displayed to the user alongside the event map determined by `pandda.analyse` to allow a crystallographer to model the expected ligand into the density. If there is sufficient evidence for the ligand the crystallographer saves the model, by fitting and merging the ligand.

The ligand has been placed into the PanDDA event map, and fitted using real space refinement in COOT [3] to the statistical event map. The superposed protein-ligand model will have been generated using `giant.merge_conformations` from the existing ground state model of the protein without the ligand bound, and the model where the protein has been fitted. Ideally changes in the local protein environment will be modelled, such as the displacement of water by the ligand, or movement of side chains. As constructing the superposed model relies on some change between the protein in the ground and bound state, for this analysis models without changes to the protein between the ground and bound state were altered, as discussed in section 4.2.4.4

By default, in existing datasets the models were refined using `giant.quick_refine` which uses the `REFMAC5` refinement of the superposed model. After an initial round of refinement, the models of both the ground and bound states are presented to the user in a modified COOT [3] refinement interface in XChemExplorer. The user can alter the bound state model, whilst the ground state model is kept common to all analysed datasets. Further iterative cycles of refinement with `giant.quick_refine` were undertaken between model building steps in the modified COOT [3] refinement interface.

Note that until the crystallographer is presented with the XChemExplorer refinement interface, they have not seen electron density maps corresponding to the refined model of the protein-ligand. Before this a crystallographer has only seen the PanDDA event map evidencing the position of the ligand. Therefore, it is at this first refinement that the crystallographer can determine whether there is sufficient evidence to support the ligand model. As the common practice in XChemExplorer is to flag datasets for progression if there is sufficient evidence, in some cases datasets flagged as 'in refinement' which were used in this analysis will have been rejected at this stage. Therefore, some of these models are likely to be of low quality, such that a crystallographer would not continue to refine them. Therefore, the models were assessed for their fit to the electron density using metrics such as RSCC, RSZD and RSZO.

### 4.2.2 Refinement methods

Refinement with superposed models follows the method outlined in section 3.2.1.5. Refinement with non-superposed modelsfollows the method outlined in section 3.2.1.6.

### 4.2.3 Refinements were run using distributed task scheduling

To compare refinement methods, all methods were run on many datasets; approximately 2000 different unique protein ligand complexes with modelled events. `phenix.refine`, Buster `refine` and `REFMAC5` were applied to superposed and non-superposed models, as well as the exhaustive search method, this led to approximately 14,000 independent refinement tasks undertaken.

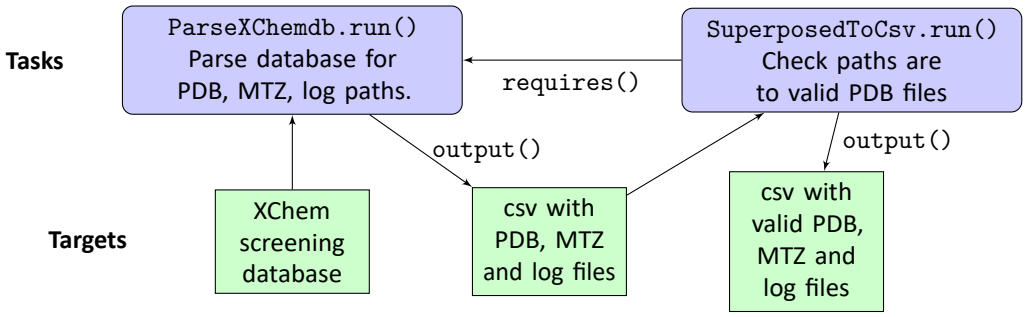



**Figure 4.1:** Luigi tasks structure facilitate large batch processes. A task is defined by its `run()` method, which indicates the actions take. A `requires()` method which indicates what needs to be run beforehand. A `output()` which defines the output file or target to be checked to determine whether the task ran successfully.

Luigi, a python task scheduling platform was used to prepare and schedule the refinement tasks. Each task is carried out using an instance of `luigi.Task` class, as shown in figure 4.1, which has three main required methods: A `run` method which carries out the action of the task, an `output` method which defines the output produced by the run method, and a `requires` method which defined the immediately preceding task in the task hierarchy. This organisation is useful for running a series of short task each with an output which precedes the next. The output target is usually a file, or a line within a database. Computationally intensive jobs are run on high performance computing (HPC) resources at DLS, other tasks are organised by the scheduler to be run in parallel on a local machine. This also had the benefit of sensibly handling failed refinement tasks.

### 4.2.4 Preparation of datasets and refinement by multiple methods

#### 4.2.4.1 Existing datasets and refinements

Figure 4.2 describes the overall refinement pipeline, which was set up to allow comparisons between multiple refinement programs.

From the XChem database, each dataset which had a correctly recorded refined model (PDB file) and reflection data (MTZ file), as well as a log file describing the existing refinement by `REFMAC5` utilising a superposed model were refined.

The log file is used to determine the convergence behaviour of the most recent iterations of refinement, as discussed in section 4.3.2. From this and the PDB file the occupancy of the superposed state can be determined

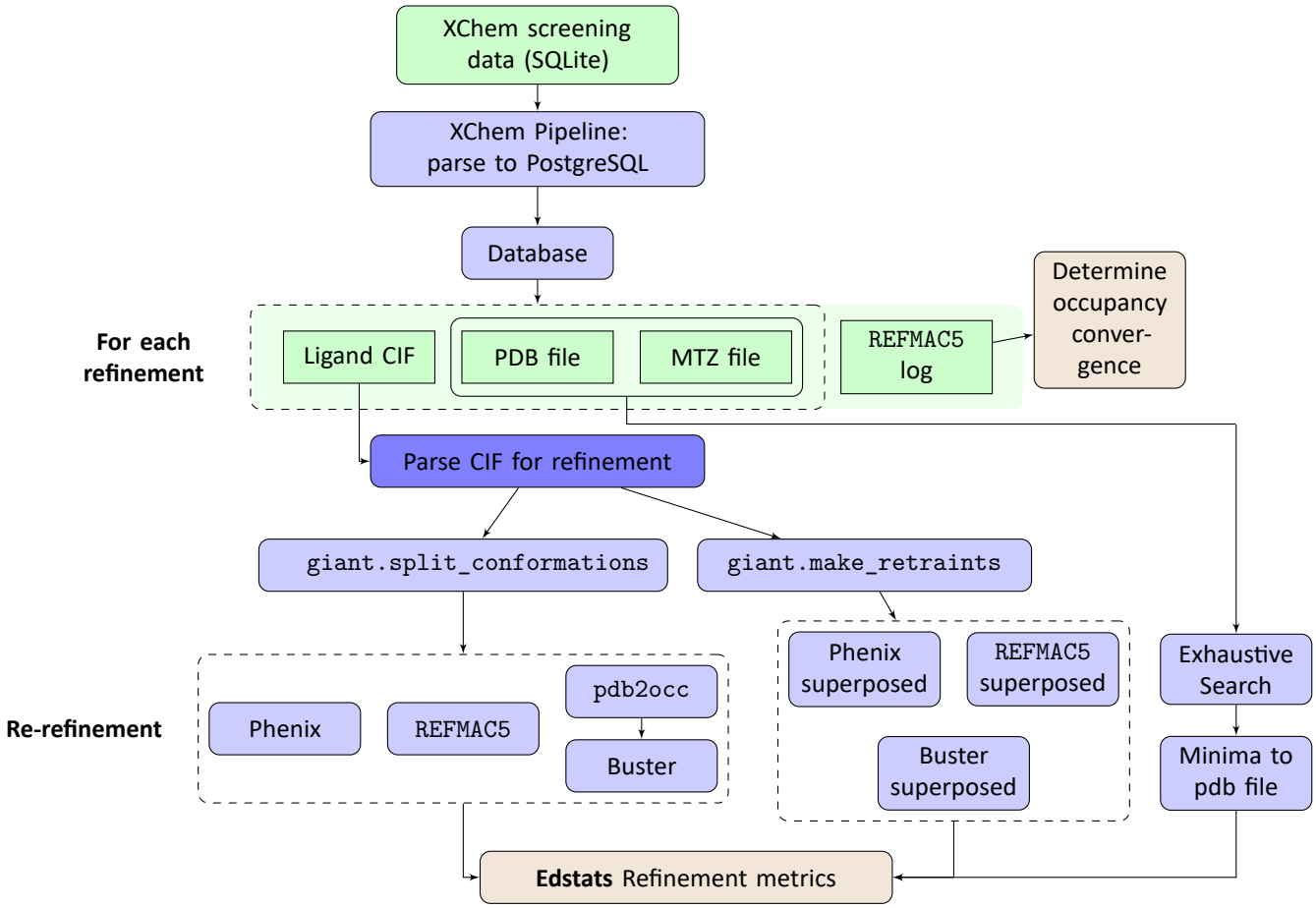



**Figure 4.2:** Re-refinement of XChem hits is suited to a task based pipeline. Green corresponds to data, blue represents tasks, with parsing CIF for refinement discussed in detail in figure 4.3 and beige represents the ends of the pipeline.

#### 4.2.4.2 New refinements

For new refinements, all datasets with an existing refinement were re-refined. A batch task submits a list of refinement tasks to the HPC cluster. Non-superposed refinements had their model prepared by running `giant.split_conformations` on the existing superposed model to obtain a current changed state model.

Superposed refinements used the existing superposed model as an initial point from which to refine. New copies of restraints are generated using `giant.make_restraints`. These restraints are modified in the case of `phenix.refine` and `REFMAC5` to increase the default number of cycles to 20 and 50, respectively.

The exhaustive search algorithm utilises the same model prepared for the superposed refinements, although no restraint generation is required.

#### 4.2.4.3 CIF preparation

Alongside ensuring the correct PDB file was obtained, care must be taken to include an appropriate CIF defining the ligands and co-factors present in the model. The process used is described in figure 4.3.

If the path to initial CIF was missing in the database, the SMILES string could be used to regenerate a CIF using AceDRG [2]. As `phenix.refine` and `REFMAC5` share the common monomer dictionary [216], most co-factors which had been handled well in the initial fragment screening refinement were processed by both these techniques. Buster `refine` uses an alternate monomer library, therefore there were some co-factors which required explicit provision of a CIF for that ligand. After discussion with the Buster developers, they provided a pre-processing script which could identify which ligands or co-factors required an explicitly provided CIF.

The existing CIF descriptors of ligands were used if possible, as this would keep refinement between different methods as similar as possible. The Buster pre-processing script was used to trigger a search for suitable existing CIF(s) of ligands/co-factors if they were determined not be in the Buster monomer library.

If previously a manual intervention to combine CIF had been created by the crystallographer (likely, in the case that `REFMAC5` monomer dictionary did not contain a co-factor), these may also be found using this recursive search. If the ligand was found in the `REFMAC5` monomer dictionary the CIF was not updated for refinement by either `REFMAC5` or `phenix.refine`. If appropriate CIF were not found, an attempt was made using `phenix.ready_set` to create ligand/co-factor CIF(s) from the refined PDB file. The missing co-factor CIF(s) would then be combined with existing CIF(s) for known ligands using `phenix.elbow` [284]. If `phenix.elbow` failed then failed Libcheck [285] was used.

#### 4.2.4.4 Adjusting superposed models

If the bound (changed) state contained only a ligand and no changes to the nearby protein residues or hetero-atoms, its occupancy would not be refined within a superposed model, as there is no difference between the ground and bound (changed) state of the protein. Therefore, in these cases a water molecule was at the centroid of the ligand in the ground

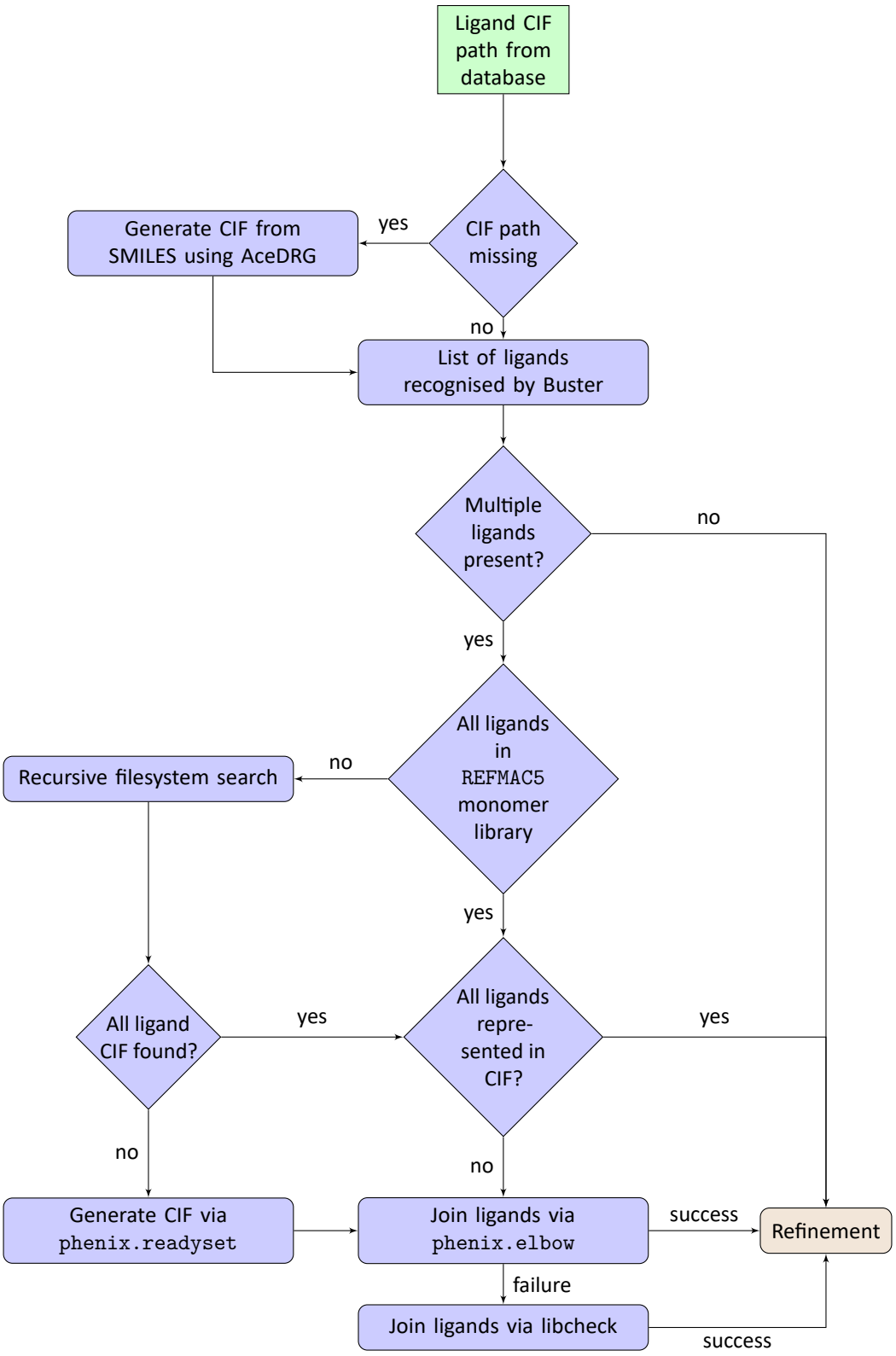



**Figure 4.3:** Flowchart for methods to prepare ligand CIF. Due to differences in monomer libraries, each method needed CIFs for different ligands. Green corresponds to data, blue are tasks, with parsing CIF for refinement discussed in detail in figure 4.3 and beige represents ends of the pipeline. Decision are represented by diamonds, whereas rectangles represent tasks.

state. Although this is a poor estimate of what may have changed in the vicinity of the bound ligand, it should be sufficient to obtain non unitary occupancies for these ligands in the superposed method.

#### 4.2.4.5 Summarising refinement information

After refinement a series of tasks are run to summarise the occupancy and B factor information from each dataset. For `REFMAC5` the per cycle convergence of occupancy information was extracted from the refinement log file. Edstats [201] is run via `giant.score_model` to generate metrics per residue.

Despite due diligence, some refinement routines work more consistently on datasets, leading to disparity in the number of refined datasets. The major causes of these disparities are discussed, but due to large volume of data processed some causes of refinement or data processing failures currently not been considered.

## 4.3 Results

### 4.3.1 Existing refinements showed that low occupancy states are uncommon

In existing refinement data very few ligands refined with a low occupancy: 95% of datasets had an occupancy above 0.38 and 99% had an occupancy above 0.3. Figure 4.4 shows the distribution of occupancy and B factor of the refinements. The occupancy was close to normally distributed around the median of 0.63. The B factor distribution was relatively constant across all hits, suggesting that uncertainty in the ligand was not being heavily compensated by overly high or low B factors.

### 4.3.2 Longer refinement increased the convergence of occupancy

The refinements recorded in the XChem database were refined using the default behaviour of `giant.quick_refine`, which uses `REFMAC5` with 10 cycles. To determine whether the occupancy of the ligand is appropriately converged, the ligand occupancy

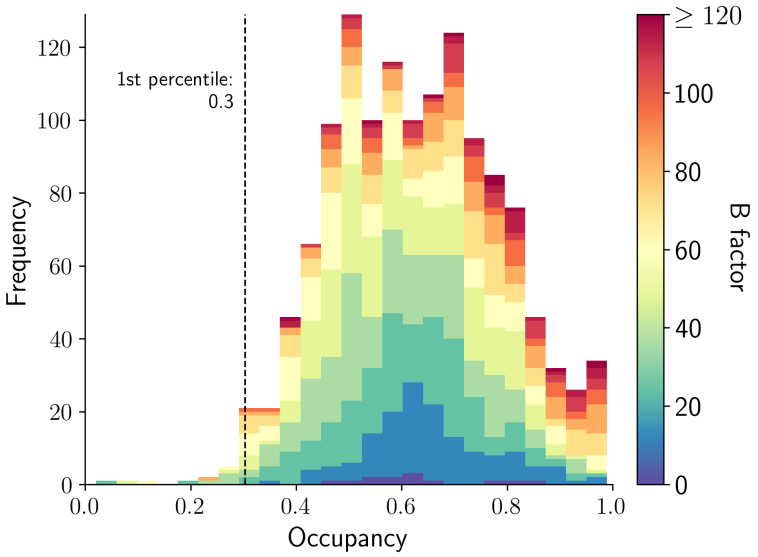



**Figure 4.4:** Existing refinement of datasets using REFMAC5 showed that low occupancies were not sampled by refinement. Histogram of crystallographic occupancy of refinements done during fragment screening. The B factor of the ligand is grouped into 10 Å bins, and plotted as a stacked histogram, B factors above 120 are grouped into the same bin.

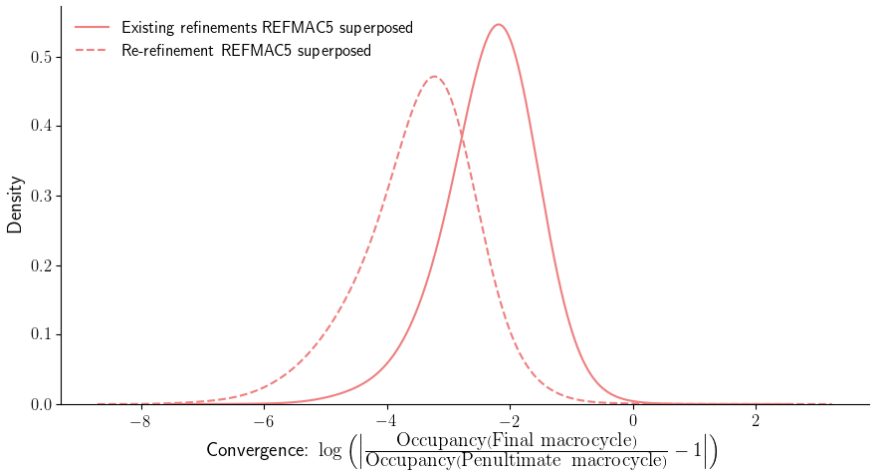



**Figure 4.5:** Longer refinements were needed for the convergence of occupancy. Convergence of occupancy after 10 cycles was notably lower than after 50 cycles. Kernel density plots of the convergence of occupancy of refinements by REFMAC5.

values in the final and penultimate cycles are compared:

$$\text{Log Convergence Ratio} = \log\left(\left|\frac{\text{Occupancy(Final cycle)}}{\text{Occupancy(Penultimate cycle)}} - 1\right|\right) \tag{4.1}$$

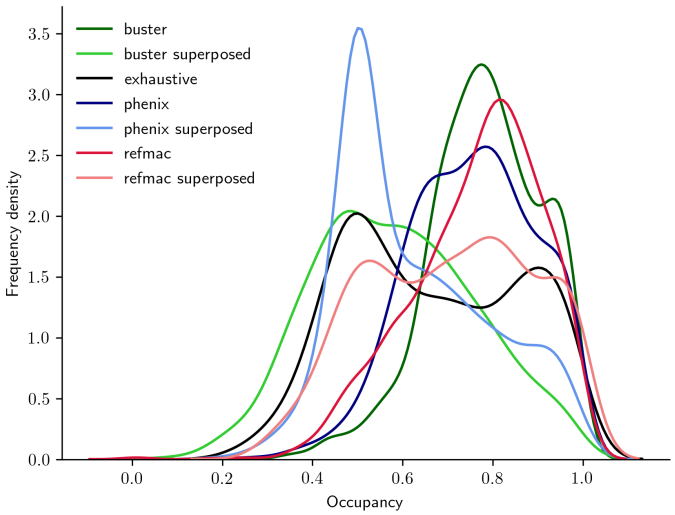
Figure 4.5 shows that increasing the number of cycles did increase the convergence of occupancy, showing the potential benefit of re-refinement of the existing datasets. Note that the number of cycles is not fully indicative of the total of refinement as previous stages of refinement stages are not currently recorded. Some datasets will have undergone multiple rounds of refinement by XChem users.

Similar results in `phenix.refine` indicated that convergence was sufficient with 20 cycles. The convergence metrics in Buster `refine` limit refinement to a given number of cycles.

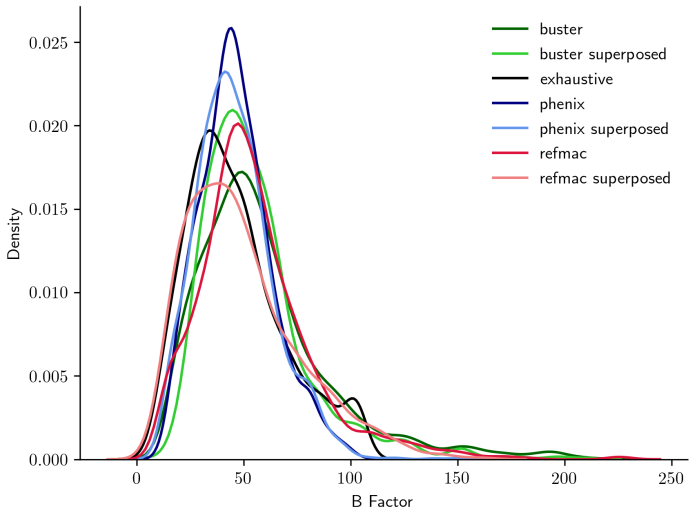
### 4.3.3 Occupancy distribution varied drastically by refinement method, unlike the B factor distribution

Occupancy refinement using different methods, showed that low occupancy states of ligands are uncommon to be reached in refinement. The 5$^{th}$ percentile of occupancy is 0.45$\pm$0.09, and 1$^{st}$ percentile is 0.30$\pm$0.15 across all refinement methods. Refinement methods had dissimilar distributions of occupancy, see figure 4.6(a). However, the B factor distribution of ligands was remarkably similar between the refinement methods, see figure 4.6(b).

Ligands refined using `REFMAC5` more often had unitary occupancy, as shown in figure 4.7(a). These unitary occupancy refinements often had high B factors. Figure 4.7(d) shows that `phenix.refine` applied to superposed models often refined to an occupancy of 0.5. This may be due to an initial starting point close to 0.5 for datasets where the superposed state was introduced as part of the preparation process. The constraint to sum to 1.0, gives the ligand and new model an occupancy of 0.5. The refinement then did not refine far from this value. Applying `phenix.refine` to superposed models also resulted in a few datasets with near 0 occupancy, therefore the 1st percentile is 0.0, figure 4.7(d), this is likely to be an artefact of the method.

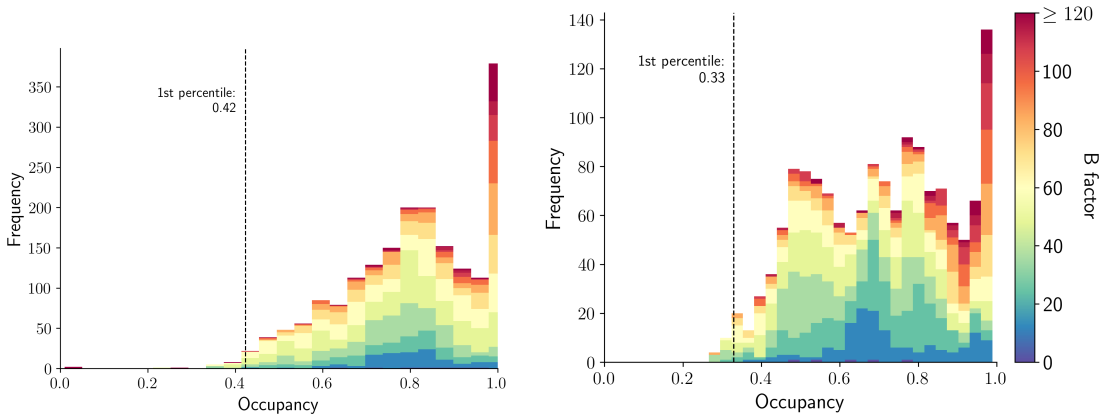



**(a)** Refinement programs did not converge on occupancy.



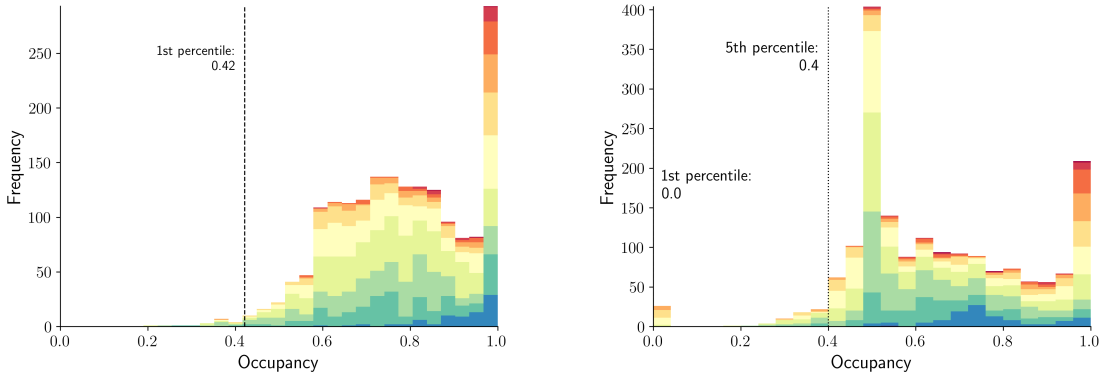

**(b)** B factor distributions are similar between refinement programs.

**Figure 4.6:** Refinement programs did not converge on ligand occupancy, but the ligand B factor distributions of the ligand were similar. All refinement methods, for both the superposed and non-superposed models were skewed towards higher occupancies. Kernel density plots for each refinement method. Data from approximately 200 hits, numbers for each method are listed in figure 4.7.



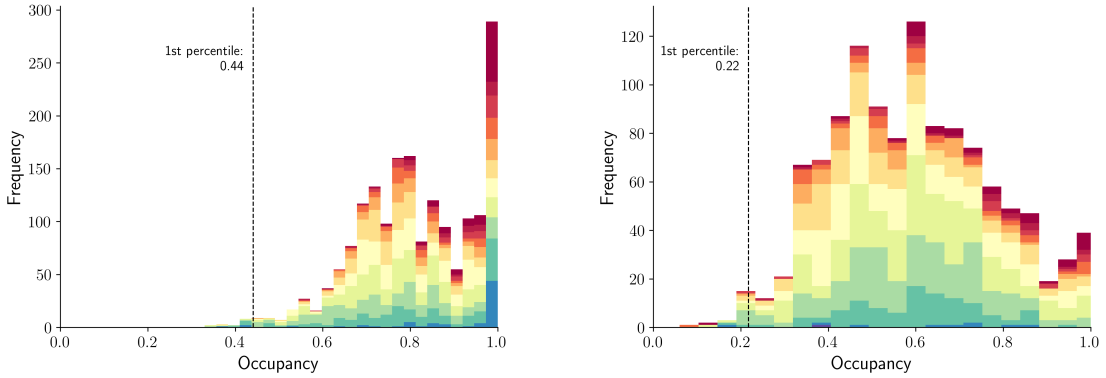

**(a)** Refmac: 1934 modelled events

**(b)** Refmac superposed: 1510 modelled events




**(c)** `phenix.refine`: 1814 modelled events

**(d)** `phenix.refine` superposed: 1798 events




**(e)** Buster refine: 1811 modelled events

**(f)** Buster refine superposed: 1179 events

**Figure 4.7:** Occupancy and B factor distributions of re-refined protein ligand complexes from existing XChem fragment screening campaigns. The B factor of the ligand is grouped into 10 Å bins, and plotted as a stacked histogram, B factors above 120 are grouped into the same bin. All refinement methods have a skew to higher occupancies for both superposed and non-superposed models. Most methods had a surprising number of ligands with unitary occupancy. This is likely an artefact of poor models (i.e. ligands fitted to density without sufficient evidence), or the presence of poor quality data. Also note that when `phenix.refine` is used on superposed models it resulted in an increased number of models refining with 0.5 occupancy for the ligand. This likely originated from ligands where the no change to the protein had been modelled. In the pipeline, the addition of a water atom at 0.5 occupancy at the centroid of the ligand likely led to this artefact.

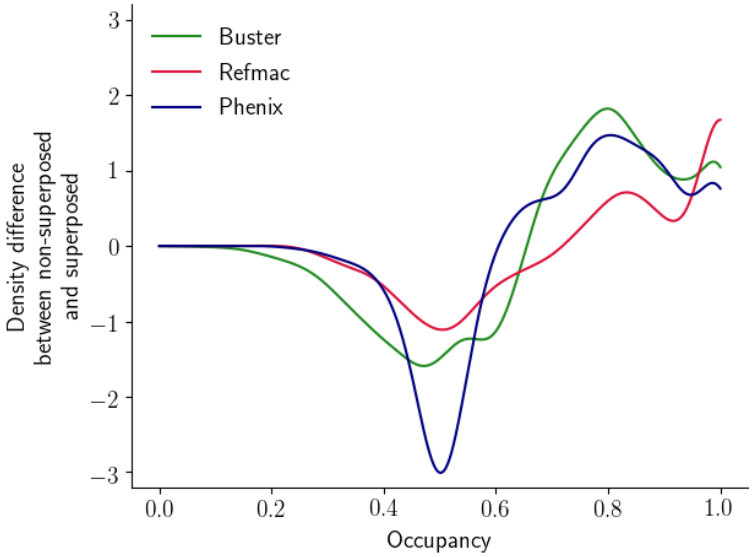



**Figure 4.8:** Difference between the kernel density plots for refinements using superposed models with both a ground and bound state, and refinements that allow that only have a single state, with an occupancy that can vary freely between 0-1. When two states are refined the occupancy of the bound ligand is lower than when allowed to freely vary, this can be seen for refinement with Buster refine, `phenix.refine` and `REFMAC5`.

For non-superposed models refined by `phenix.refine` a large number of hits had unitary occupancy, figure 4.7(d). There was not as strong a bias to high B factor at unitary occupancy when compared to `REFMAC5`.

For non-superposed models refined by Buster `refine` a wider spread of high B factors across a wider occupancy range than in the other refinement methods was observed. Figure 4.7(e) shows for refinement of non-superposed models using Buster `refine`, there were many near unitary occupancy ligands, with a wide spread of B factor. Furthermore, a peak at unitary occupancy, was not observed for Buster `refine` with superposed models, as seen in figure 4.7(f).

### 4.3.4 Superposed methods consistently had lower ligand occupancy

Figure 4.8 shows the difference between occupancies for refinements of datasets using a superposed model, and those without a superposed model. When there is a ground state the occupancy of the bound states is lower, this is the case for refinement with Buster refine, `phenix.refine` and `REFMAC5`. It was hypothesised that the superposed model may be appropriate in many cases, as it could better represent the different states

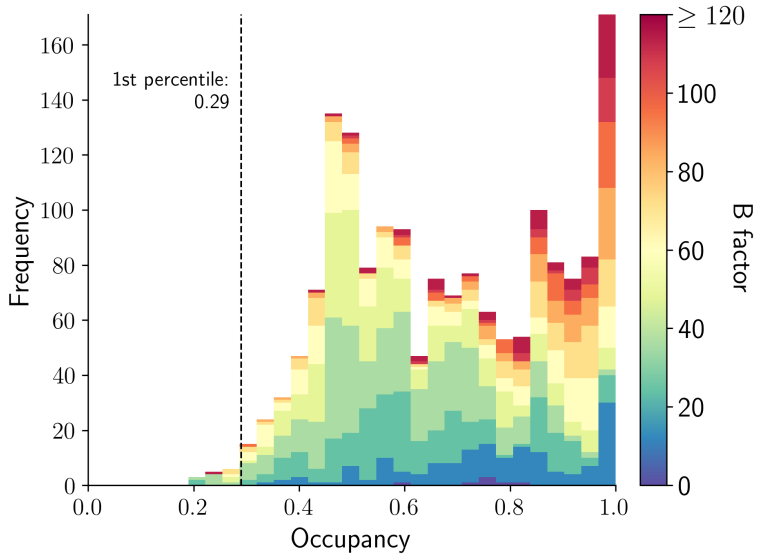



**Figure 4.9:** Occupancies from exhaustive search were more uniformly distributed than other methods. The B factor of the exhaustive search minima for the ligand is grouped into 10 Å bins, and plotted as a stacked histogram.

that can be adopted in the binding pocket of the ligand in a single model, by explicitly modelling the ones that are known before binding. However, these results only show that different models, with different constraints can represent a local environment and does not inform which is most appropriate.

### 4.3.5 Exhaustive search showed similar behaviour to refinement of superposed models

The exhaustive search algorithm tends to skew to higher occupancies similar to other refinement programs. Figure 4.9 shows the occupancy and B factor at the minima of difference density found using the exhaustive search method, discussed in chapter 3. A lower B factor was more likely than in other methods, this may be partly due to the cut off of B factor at 103 Å$^2$. There was a peak near 0.5 although this was not as large as that for `phenix.refine` and thus is less likely to be solely an artefact.

In general , the distribution was quite similar to that of the other refinement methods, suggesting that the exhaustive search is not exploring drastically different model parameters to refinement methods. The smoothness of the exhaustive search minima suggests that different refinement methods may compensate variations in occupancy with B factor.

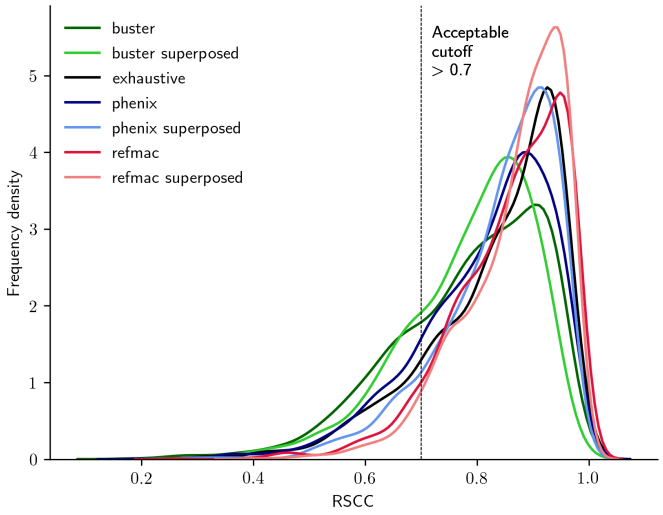



**Figure 4.10:** RSCC of the ligand atoms for all XChem hits, in each refinement method, with both superposed and non-superposed models. Most ligand models (66±9% (standard deviation over refinement methods) have an RSCC above 0.8) have a partial fit to the observed density. A large proportion have at least a part of the ligand which is well described by the density (86±7% have an RSCC over 0.7).

### 4.3.6 Ligands were sufficiently supported by density to be a representative subset

As these datasets were those which crystallographers have attempted modelling a ligand, it would be expected that there is reasonable evidence based on fit to electron density for most ligands. 66 % of the ligands modelled have an RSCC over 0.8, suggesting that the ligand fits the density in a substantial proportion of its substructure, as shown in figure 4.10. With 86% of ligands having an RSCC > 0.7, the majority of ligands have at least some part of the ligand is evidenced. As most of the datasets are not classified as ready for deposition, there may be further model building improvements which would improve the RSCC. Furthermore, as some of these ligands would be not progressed by the crystallographer to deposition, a subset with a higher RSCC would be considered final models.

Figure 4.10 shows that there is little difference between the RSCC distributions for superposed and non-superposed models when refined with `phenix.refine` or `REFMAC5`. This suggests that the coordinates are not being significantly changed when the ground

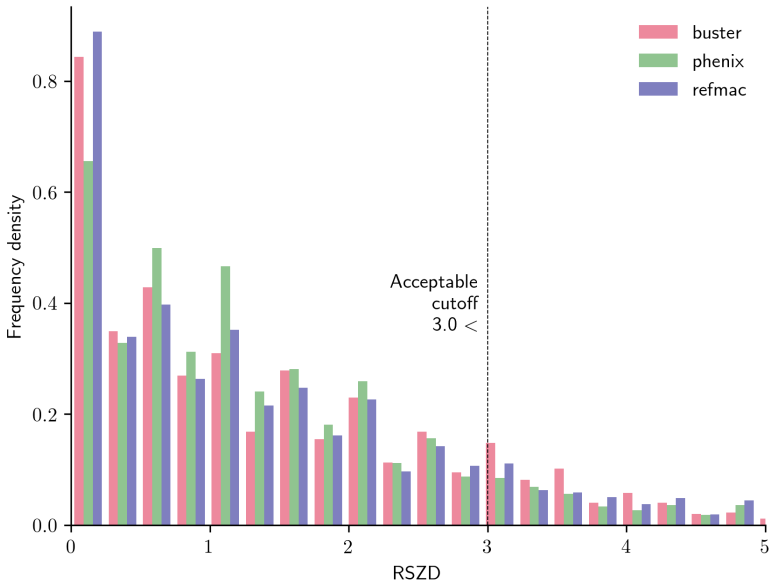
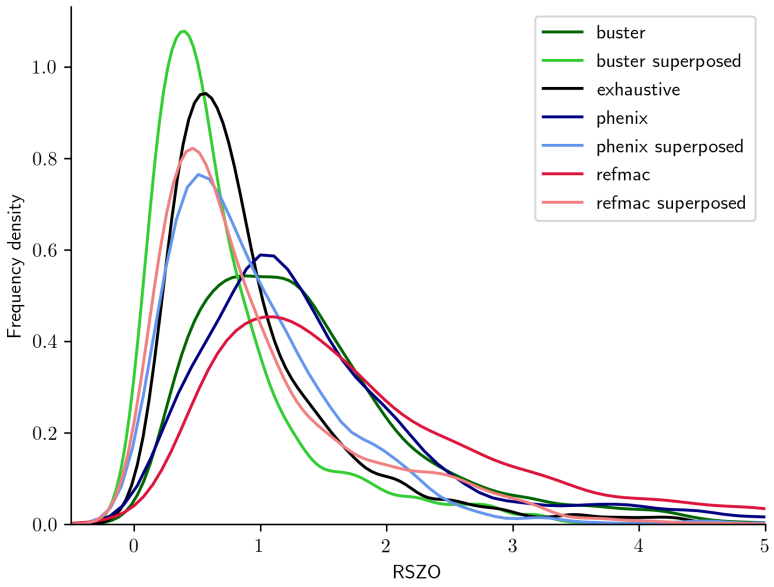



**Figure 4.11:** Real space Z-difference of XChem hits, shows that there was not an excess of difference density over the model and atoms were not over or under represented. Note that there were compatibility issues between superposed models and Edstats [201] the only the RSZD scores for the non-superposed models were reported.

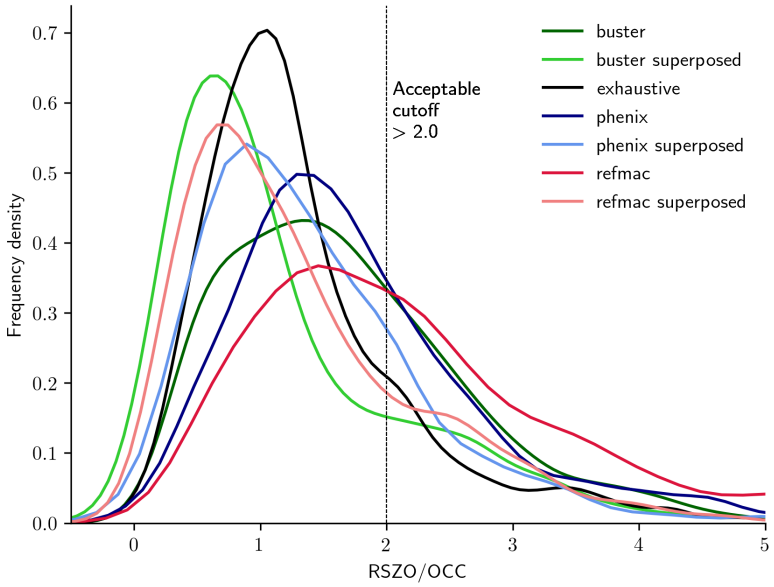
state of the ligand is being considered. The larger difference between RSCC values for superposed and non-superposed models when Buster is used suggests that there is more difference between these sets of models. This may be partly due to the much lower number of superposed models which were successfully processed by Buster (1179 compared to 1811). It may also be further evidence that the adjustments to refine superposed models in Buster, are not as well established as those used when `giant.quick_refine` is run using `phenix.refine` or `REFMAC5`.

Most ligands also have an RSZD below 3.0 (figure 4.11), suggesting that there is not an excess of difference density over the model, and that atoms are generally not over or under represented. Note that there were compatibility issues between superposed models and Edstats [201] so only the RSZD scores for the non-superposed models were reported.

Figure 4.12 shows distribution of RSZO with and without normalisation by occupancy. Most ligands had a surprisingly low RSZO score indicating that datasets had some misplaced model atoms. Most datasets have a relatively low RSZO score as normalised by occupancy, with only 27$\pm$10% of datasets above the previously determined cut off




**(a)** RSZO




**(b)** RSZO

**Figure 4.12: (a)**Distribution of RSZO, and **(b)** distribution of RSZO/occupancy for ligand atoms in XChem datasets. RSZO measures density precision. RSZO varies with refinement programs used, and this variation increases when RSZO/Occupancy is considered, as the occupancy distributions vary considerably between refinement program. Only a small proportion of datasets were above the acceptability threshold (>2.0) outlined in [202].

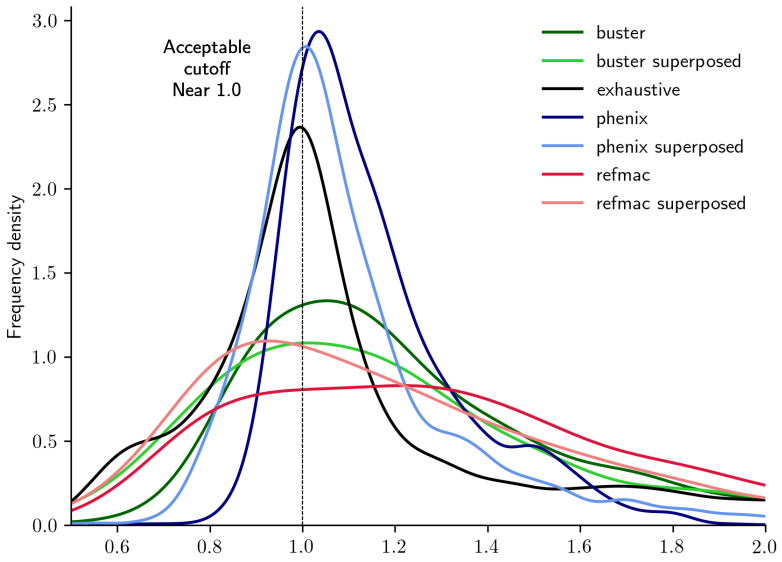



**Figure 4.13:** Distribution of the ratio of ligand B factor to the protein atoms surrounding the ligand. A ratio of 1.0 would indicate that the B factor of the ligand was the same as the surrounding residues. By default, phenix refinement methods restrains the ligand B factor to that of its surroundings in the protein, thereby having a ratio closer to 1.0.

of 2.0 [202]. This indicates if a higher volume of data were available, further filtering should be applied, and the applicability of this metric threshold should be explored for weakly bound ligands.

Figure 4.13 shows the ratio of ligand B factors to the protein atoms surrounding the ligand within 4 angstroms. A ratio of 1.0 would indicate that the B factor of the ligand was the same as the surrounding residues. As `phenix.refine` by default introduces restraints to keep the ligand B factors close to those of its surroundings, it is not surprising that these methods have a ratio much closer to 1.0. It is somewhat surprising how far away from 1.0 other methods take the occupancy, suggesting that such restraints of the B factor may produce more meaningful B factor distributions, and corresponding occupancies. It would be surprising to see a ligand that is much more mobile than the surrounding protein. The model should instead reduce the occupancy (i.e. the ligand either being present in the site, or not for a proportion of sites in the crystal) rather improving the fit to density by changing the ligand B factor.

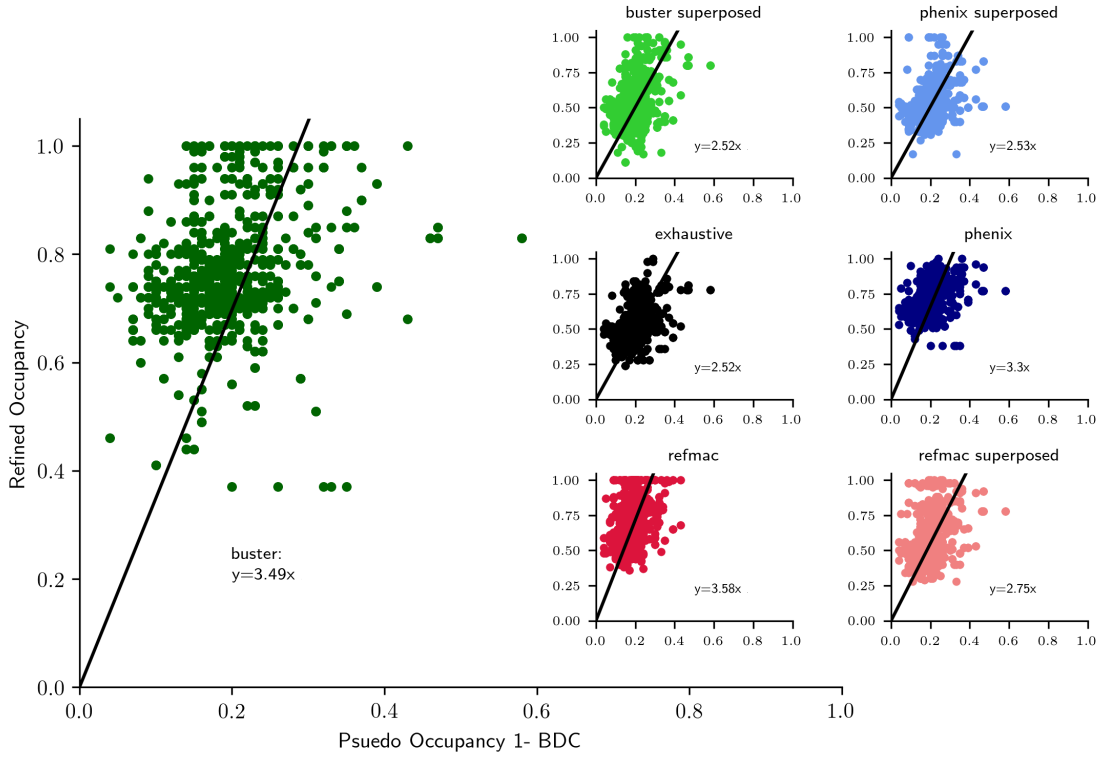



**Figure 4.14:** Occupancy of ligands is higher than estimated by background density correction. Scatter plots of occupancy against pseudo-occupancy (1- Background Density Correction). All refinement methods show a larger gradient (2.95±0.45) than established previously (2.26, which has been applied in PanDDA as 2.0) [203].

Exhaustive search provides a B factor ratio distribution closely centered to 1.0. As the B factor of the surrounding atoms in the superposed model were varied at the same time as the ligand, it is reasonable that difference density was minimised when these B factor values are close to one another.

### 4.3.7 PanDDA background density correction calibrated by refined occupancy of XChem hits

The background density correction is the proportion of ground state to take from each crystal dataset to show clear evidence in the changed state. This produces the event map. The previously determined benchmark slightly overestimates the proportion of ground state observed in the crystal, and the default proportionality factor in PanDDA should be adjusted.

The relationship between occupancy and background density correction (BDC) factor was originally established using four targets [203]:

$$\text{Occupancy} = 2.262\,(1 - BDC) \tag{4.2}$$

Figure 4.14 shows that on the larger test set of XChem targets presented, there is an increase in the proportionality constant to:

$$\text{Occupancy} = 2.95(\pm 0.45)\,[1 - BDC] \tag{4.3}$$

Removing data with RSCC below 0.7, does not effect the proportionality constant.

The slightly low default of proportionality used for generating event maps of 2.0, may lead to false negatives, due to uninterpretable density. There is still a large variation between refinement methods likely due to the uncertainty in ligand occupancy being represented in both occupancy and B factor variations.

## 4.4 Discussion and future work

Exploring alternate refinement routines for weakly binding ligands from XChem fragment screening campaigns it has been shown that low occupancies are not commonly sampled. This chapter cannot give definitive evidence why this is the case, only describe the trends observed in these refinements. As low occupancies reduce the scattering factors of an atom proportionally, it is possible such low occupancies are below the noise level achievable in the electron density supporting protein models. However, I believe that there is not currently a good way to parameterise a model with low occupancies,which limits their sampling.

The ligand B factor distributions across all methods are more similar to one another than occupancy, showing that this parameter is more stable under refinement across all refinement implementations.

The superposed methods consistently have a lower ligand occupancy which indicates each dataset has a consistently occupied ground state.

I acknowledge that although RSCC shows that the majority of the ligands are well supported in the density, the low occupancy normalised RSZO values suggest that there are systematic issues with evidence for the models. During chapter preparation, the XChem database only contained approximately 10% of the hits screened at the XChem facility. Expanding the number of datasets, by improving data cleaning steps in the pipeline would allow filtering by local data quality metrics such as RSCC, RSZO and RSZD. This could allow for further conclusions to be drawn on the applicability of the refinement of superposed models.

In the future re-refinement could be run automatically for all datasets. This would act as a benchmarking set to test model building algorithms (such as automated building into event density) for weakly bound ligands. It would also allow users to inspect their datasets after re-refinement in alternate programs.

Further work should also include development of information criterion metrics into refinement methods. Given a collection of models for the data, information criterion metrics estimate the quality of each model, relative to each of the other models. Thus, they provide means for model selection. The most common information criterion the Akaike information criterion is not directly applicable in crystallographic refinement, as models are restrained and not constrained [286].

Changes to the structural model, such as the introduction of additional parameters associated with any changed states or ligand in a superposed model, cannot be easily assessed to whether they have a sufficient improvement in explaining the electron density given the increase in model complexity. Some newer statistical techniques are being applied to crystallography [286] which may be able to surpass these limitations, but will need integration with both refinement and model building program developers and extensive benchmarking against existing models.

Given that in chapter 3 it was shown that repeated soaks of protein ligand complexes crystallised in the same manner vary in occupancy; it was expected that fragment ligands would be challenging to model, and thus refine with different occupancy values in current refinement programs. Future work to improve B factor parametrisation (such as the

bond based parametrisation being developed in Vagabond [287]) will likely lead to better local error models, and better models for crystallographic occupancy.

As there is variability in occupancy refinement, a crystallographer should not be concerned following up medium to high occupancies (0.3-1.0), but care should be taken to routinely trust occupancies below 0.3 of weakly bound ligands. In such cases it is recommended to try re-refinement to determine whether that occupancy value is stable, and assess the fit to density metric carefully before proceeding, as such a low value would be considered surprising given the occupancy distributions observed.

# 5

# Best practice for determining crystallographic occupancy using mass spectrometry to calibrate refinement methods

## Contents


**5.1 Introduction . . . . . 144**
5.1.1 Covalent binders as chemical probes . . . . . 145
5.1.2 Intact mass spectrometry for detection of covalent binding . . 145
**5.2 Methods . . . . . 145**
5.2.1 Compound selection . . . . . 145
5.2.2 Protein preparation . . . . . 146
5.2.3 Mass spectrometry methods for protein in solution and protein crystals . . . . . 149
5.2.4 Pre-crystallisation mass spectrometry detected ratio of covalent compound . . . . . 150
5.2.5 Post crystallisation mass spectrometry detected ratio of covalent compound . . . . . 152
5.2.6 Crystallographic data collection and refinement methods . . . 165
**5.3 Results . . . . . 166**
5.3.1 Crystallographic occupancy was correlated with ratio of covalent compound detected with intact mass spectrometry . . . 166
5.3.2 Restraining B factor allowed post crystallisation mass spectrometry to calibrate refinement methods . . . . . 170
**5.4 Discussion and future work . . . . . 170**

## 5.1 Introduction

Chapters 3 and 4 have shown that low occupancies are not sampled in refinement of weakly binding fragment ligands. To understand whether these low occupancies are not detectable, or whether the model parameterisations and refinement method are poorly designed to sample low occupancies that could be present, a calibration method for crystallographic occupancy was sought.

This chapter discusses the development of a MS based method to correlate the crystallographic occupancy of a protein-ligand complex with an orthogonal measurement of the amount of ligand bound. Using a covalent binder allows use of intact MS to measure the abundance of the compound.

By mixing protein labelled with the covalent compound with compound free protein it was hypothesised that the proportion of compound measurable in intact MS of crystals would correlate with the crystallographic occupancy of the ligand. If low occupancies were detectable using intact MS, but not by refinement, alternate model parameterisations could be tested to determine whether low occupancies could be detected and modelled using the X-ray diffraction data. Section 5.3.1 demonstrates the correlation between the refined occupancy of the ligand and the ratio of compound detected by intact MS of crystallised protein.

Ideally the same crystal which had undergone diffraction would be used for comparison by processing with intact MS. This would allow measurement of the crystallographic occupancy and ratio of compound in the same crystal. However, sections 5.2.5.2 and 5.2.5.3 show this was not possible due to signal to noise limitations.

By fixing the B factor of the ligand to the mean of its surroundings, a better fit to the intact MS data was possible. This suggests that B factor restraints, and an improved model for occupancy, are needed to improve the refinement of weakly binding protein-ligand complexes.

### 5.1.1 Covalent binders as chemical probes

Covalent probes are molecules which bind covalently and selectively to a protein of interest. Covalent probes typically have a prolonged duration of action [288], as they bind irreversibly, and therefore the protein of interest will remain inhibited until it is degraded and regenerated. Such probes are usually developed to selectivity bind protein nucleophiles that are not conserved across target families [289]. Covalent binders are detectable in intact MS, as they bind irreversibly with the protein.

Covalent inhibitors have historically been developed by incorporating an electrophile into an already optimised lead [290–293]. HTS of covalent binders has previously been avoided due to concerns about promiscuity and non-specific reactive compounds [59]. Recent development of covalent electrophile [94] libraries focuses on ruling out promiscuous compounds by evaluating reactivity, and confirming hits by X-ray crystallography.

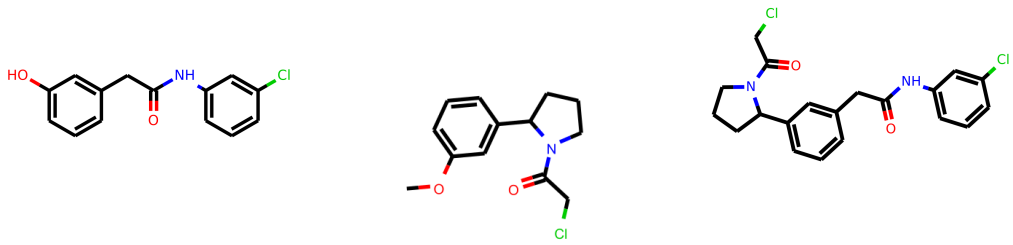
### 5.1.2 Intact mass spectrometry for detection of covalent binding

Intact MS detects a protein's total molecular weight without fragmentation. Electrospray ionization allows transfer of proteins and complexes intact into the gas-phase [294]. Although the protein is intact, it is typically linearised, therefore not in its native folded state. Covalent electrophiles which bind irreversibly to cysteine or other nucleophilic residues, are therefore detectable using intact MS. Further details are discussed in section 1.7.

## 5.2 Methods

### 5.2.1 Compound selection

A previous crystallographic fragment screen with non-covalent fragments against Peroxisomal coenzyme A diphosphatase (NUDT7) (residues 14-235) identified 18 fragments bound to the putative substrate binding region (`10.5281/zenodo.1244111`). A fragment screen of covalent electrophiles using intact MS identified 24 selective hits. Compound (**6**), shown in figure 5.1(c), was synthesised to obtain overlap of a phenyl ring between a non-covalent fragment (**4**) and a covalent electrophile (**5**). Compound




**(a)** NUDT7-REV-1 (**4**) **(b)** PCM-0102716 (**5**) **(c)** NUOOA000308a (**6**)

**Figure 5.1:** Covalent binder for NUDT7 was a result of fragment merging. Compound (**4**) was shown to bind to NUDT7 by fragment screening. Its binding pose had a phenyl ring overlap with the covalent electrophile hit compound (**5**) found by MS screening. Therefore both compounds were merged into compound (**6**) which was shown to bind in the same pose, and is used in this chapter.

(**6**) showed 100% labelling after 15 minutes, at 5 µM compound to 2 µM of protein. The compound binds to the catalytic cysteine 73 of NUDT7 [94]. Due to the high labelling efficiency and detectability of the ligand in the co-crystal complex, this compound was used to explore whether crystallographic occupancy is correlated with the amount of compound detectable using intact MS.

## 5.2.2 Protein preparation

### 5.2.2.1 Purification

NUDT7 purification is detailed in appendix section B.4.

### 5.2.2.2 Labelling and Gel Filtration

To produce a stock of labelled NUDT7: 600 µl of NUDT7 protein which was concentrated to 27 mg $ml^{-1}$ was initially incubated with 1 mM of compound (**6**). After 30 and 70 minutes, intact MS (method in section 5.2.3.1) still showed partial labelling, with peaks for both the unlabelled protein (25 124 Da) and labelled protein (25 479 Da) as seen in figures 5.2(a) and 5.2(b) respectively. The compound concentration was increased to 3 mM (**6**) to achieve 100% labelling, as in figure 5.2(c). Note that labelling of the protein in two distinct sites was possible, this gave a peak at 25 832 Da. This was minimised by reducing

the free compound by running gel filtration, after labelling the protein with a minimum amount of compound to achieve 100% single labelling.

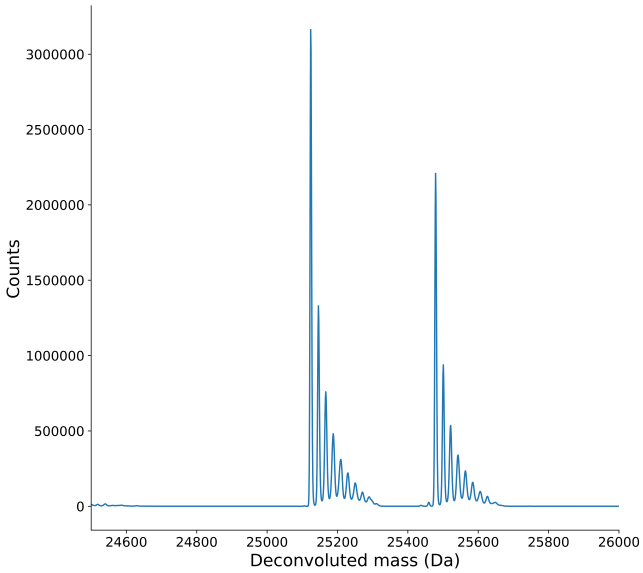
Removal of free compound was also important as labelled and unlabelled NUDT7 protein were mixed to form crystals. Further exposure to free compound would have altered the proportion of labelled protein.

For both labelled and un-labelled protein the flow-through was concentrated and purified further by size exclusion chromatography using a YARRA SEC-2000 PREP column (Phenomenex) equilibrated with binding buffer. Fractions containing protein were pooled, concentrated to 25 mg $ml^{-1}$ for unlabelled protein and to 24 mg $ml^{-1}$ for labelled protein and were stored at −80 °C.

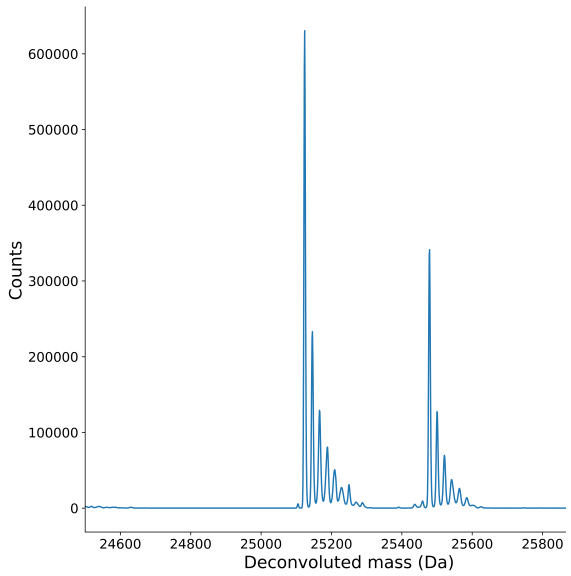
#### 5.2.2.3 Crystallisation

NUDT7 crystallisation was optimised around initial conditions used for the previous fragment screening, accessible on the protein databank group deposition id: G_1002045, also reported in [94]. Crystallisation trials were undertaken in sitting drop vapour diffusion plates (96-well 3-drop Swissci plates) with 150 -300 nl drops consisting of 2:1, 1:1 and 1:2 ratios of 25 mg $ml^{-1}$ protein solution and reservoir solution respectively. Crystals were obtained using reservoir solutions containing 0.1 M ammonium acetate and 0.1 M bis-tris pH 5.5 and 5-7% (w/v) PEG 10,000. Compact hexagon shaped crystals with a typical diameter of 50-150 µm usually nucleated within 24 hours, growing to a maximal size within 4 days when stored at 20 °C

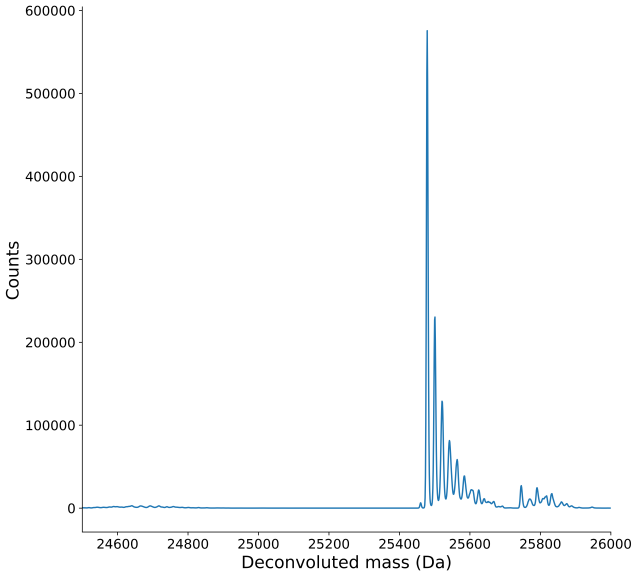
To increase reproducibility of crystallisation, micro-seeding stocks were used throughout. These micro-seeding stocks were produced from crystalline protein (crystals or crystalline precipitate) with added base buffer (10 mM HEPES, 500 mM NaCl, 5% Glycerol, 0.5 mM TCEP, pH 7.5) or reservoir solution. The crystalline material was broken into small pieces by vortexing the solution in an Eppendorf tube with a PTFE bead. Serial dilutions (typically 1:4 - 1:4096) of the seed stock were prepared. To evaluate each dilution a row of a crystallisation plate was setup to test whether seeds improve reproducibility of the crystals, or aid nucleation of larger crystals. For preparation of vapour diffusion sitting




**(a)** 30min after labelling with 6 µl




**(b)** 70min after labelling with 6 µl




**(c)** 30min after labelling with total of 16 µl

**Figure 5.2:** NUDT7 was 100% efficiently labelled by covalent binder. Successive increases of concentration of the compound (**6**). Unlabelled protein at 25 124 Da and labelled at 25 479 Da. The minor peaks seen are sodium adducts, repeating units of +22 Da.

drops typically 20 nl microseeding stock is added to a 150 -300 nl drop. Microseeding stocks diluted to 1:1, 1:50 and 1:1024 were used for the co-crystals with compound (**6**).

Co-crystals with compound (**6**) were produced using the same crystallisation conditions as for apo-protein crystals. The two species of labelled and unlabelled protein were mixed just prior to the setup of crystal plates, either by pipette or Mosquito™ as described in section 5.2.4. Increasing ratios of labelled to unlabelled protein were produced in 10% steps until fully labelled (100%), which were then adjusted for differences in concentration of the protein samples used to set-up the plates.

### 5.2.3 Mass spectrometry methods for protein in solution and protein crystals

Two methods have been used; a three-minute method is used primarily for analysis of pre-crystallisation samples. A twelve-minute method which was originally developed for separation of detergent from integral membrane proteins [295], is used to reduce the co-elution of low concentrations of PEG with the protein species.

#### 5.2.3.1 Three-minute liquid chromatography coupled with mass spectrometry

In this work the MS method outlined in [246] is used, described here briefly: Reversed-phase chromatography was performed in-line prior to MS using an Agilent 1290 ultra high performance liquid chromatography (uHPLC) system (Agilent Technologies inc. – Palo Alto, CA, USA). Concentrated protein samples were diluted to 0.02 mg $ml^{-1}$ in 0.1% formic acid and 50 µl was injected on to a 2.1 mm x 12.5 mm Zorbax 5 µm 300SB-C3 guard column (Agilent Technologies) housed in a column oven set at 40 °C. The solvent system used consisted of 0.1% formic acid in ultra-high purity water (Millipore) (solvent A) and 0.1% formic acid in methanol (LC-MS grade, Chromasolve) (solvent B). Chromatography was performed as follows: initial conditions were 90% A and 10% B and a flow rate of 1.0 mL $min^{-1}$. A linear gradient from 10 % B to 80 % B was applied over 35 seconds. Elution then proceeded isocratically at 95% B for 40 seconds followed by equilibration at initial conditions for a further 15 seconds. Protein intact mass was determined using a 6530 electrospray ionisation quadrupole time-of-flight mass spectrometer (Agilent

Technologies Inc. – Palo Alto, CA, USA). The instrument was configured with the standard ESI source and operated in positive ion mode. The ion source was operated with the capillary voltage at 4000 V, nebulizer pressure at 60 psig, drying gas at 350 °C and drying gas flow rate at 12 L $min^{-1}$. The instrument ion optic voltages were as follows: fragmentor 250 V, skimmer 60 V and octopole RF 250 V. Data analysis was performed using Masshunter Qualitative Analysis B7.0 proprietary software and deconvolution performed using the maximum entropy algorithm.

#### 5.2.3.2 Twelve-minute liquid chromatography coupled with mass spectrometry

This method was adapted from the method to separate detergent from integral membrane proteins [295]. Samples were diluted to a final volume of 50 µl in 30 % methanol, 0.1 % formic acid. This was injected on to an Agilent 1290 uHPLC system coupled to a 6530 QTOF mass spectrometer fitted with a Zorbax StableBond 300 C3, 2.1 x 150 mm, 5 µm HPLC column (Agilent Technologies, Santa Clara, USA). The solvent system used consisted of 0.1% formic acid in ultra-high purity water (Millipore) (solvent A) and 0.1% formic acid in methanol (LC-MS grade, Chromasolve) (solvent B). Initial conditions were 30 % B at 0.5 mL $min^{-1}$. After 1 min a methanol gradient was applied from 30 % to 95 % over 7 min. Elution was then isocratic at 95 % B for 2 min, followed by a further 2 min equilibration at 30 % B. The mass spectrometer was operated in positive ion, 2 GHz detector mode. The ion source was operated with the capillary voltage at 4000 V, nebulizer pressure at 60 psig, drying gas at 350 °C and drying gas flow rate at 12 L $min^{-1}$. Fragmentor was 250 V, collision energy 0 V and data acquired from 100-3200 m/z. Data analysis was performed using Masshunter Qualitative Analysis B7.0 proprietary software and deconvolution performed using the maximum entropy algorithm.

### 5.2.4 Pre-crystallisation mass spectrometry detected ratio of covalent compound

To quantitatively determine differing amounts of protein labelled with compound (**6**), intact mass spectra of different ratios of unlabelled NUDT7 to labelled NUDT7 were measured with both MS methods, outlined in sections 5.2.3.1 and 5.2.3.2. This involved

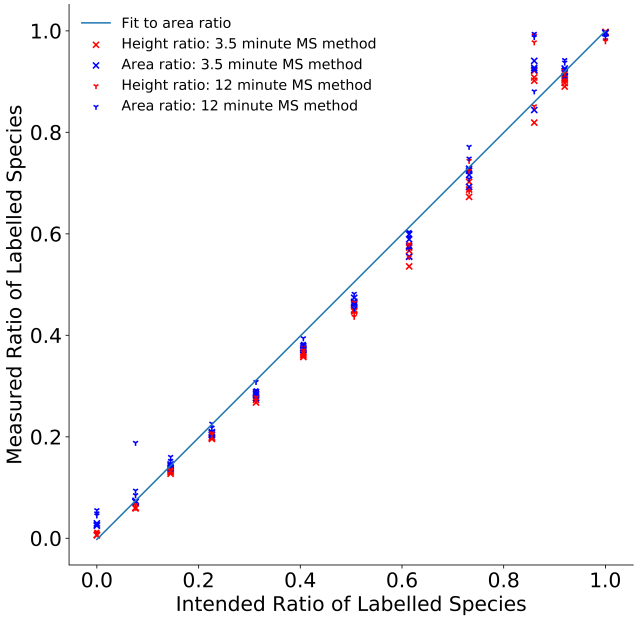



**Figure 5.3:** Intact MS measured the proportion of covalently labelled protein without systematic bias prior to crystallisation. The intended ratio was the ratio between protein labelled with compound (**6**), and unlabelled protein. This was adjusted by the concentration of the labelled and unlabelled species. The y axis shows the ratio between peaks associated with labelled protein (i.e. 25 479 Da and respective adducts) and unlabelled (i.e. 25 124.1 Da and respective adducts). The area ratio considers the area under the curve of the deconvoluted mass spectrum (-50, +100 from the largest peak). The height ratio takes the height of the largest peak in the appropriate series of adducts.

preparation of 96 samples using a Mosquito®HTS (TTP Labtech) to dispense a total of 2 µl of diluted protein in measured ratios. Frozen stock of NUDT7 was diluted in purification buffer. The intended ratio varied from 0 to 1.0 in 0.1 steps, with an additional ratio of 0.85. Measured concentrations were for unlabelled protein: 1.52 mg $ml^{-1}$±0.10 (n=5) and for labelled protein: 2.23 mg $ml^{-1}$±0.03 (n=5). Concentrations were measured using 2 µl samples on a NanoDrop™ 2000 Spectrophotometer, using protein A280 method using the Beer-Lambert law with an estimated extinction coefficient of 13,410 and a mass of 25 124 Da. To 48 samples 50 µl of 0.1% formic acid were added and processed according to the method outlined in section 5.2.3.1. To the other 48 samples 50 µl of 0.1% formic acid and 30% methanol were added and processed according to the method outlined in section 5.2.3.2. As the longer method was going to be used with all post

crystallisation examples, it was important to check the agreement of both methods to mitigate against any systematic bias.

Figure 5.3 shows that the dispensed ratios when corrected by protein concentration were well reproduced in the ratio measured by MS. Both the longer and shorter MS methods showed consistent results, and therefore were not detectably affected by systematic bias. Measuring the amount of covalent compound present was not affected by how the ratio between unlabelled and labelled NUDT7 is measured either by maximal peak height, or by total area associated with a species and it's adducts.

### 5.2.5 Post crystallisation mass spectrometry detected ratio of covalent compound

The aim was to explore whether crystallographic occupancy of compound (**6**) obtained from refinement of structural model against crystallographic data was correlated with the amount of covalent labelled NUDT7 detectable in MS. A consistent protocol for measuring the ratio of compound after crystallisation using MS was required. Firstly, the crystal preparation method was developed, as outlined in section 5.2.5.1, which consistently led to sufficient signal to identify the NUDT7 protein and to determine the extent of labelling by the covalent ligand.

Ideally crystals screened by X-ray crystallography would be used to compare crystallographic occupancy to the ratio of covalent labelling by intact MS. When attempting to work on smaller crystals, and crystals which had already been exposed to X-ray synchrotron radiation, ion suppression was identified, as discussed in section 5.2.5.3, which led to insufficient signal. In some cases there was no ion suppression detected, but there was insufficient signal to identify NUDT7. This is likely due to a lower amount of crystalline material after the crystal has been mounted and processed as outlined in section 5.2.5.1. Some samples used to assess the ratio of covalent compound post diffraction had no signal in intact MS due to loss of the crystal in transferring between the cryo-cooled harvesting loop and each of the wells used to wash the crystals.

To increase the signal to noise of these samples alternate sample preparation techniques were tested. These included new crystal conditions, discussed in section 5.2.5.6,

alternate selectivity of chromatographic columns, discussed in sections 5.2.5.8 and 5.2.5.9, and changing uHPLC methods, discussed in section 5.2.5.10. Unfortunately, these methods were at best partially successful. Thus this work focuses on the correlation of crystallographic occupancy with crystals prior to diffraction from the same crystallisation conditions in section 5.3.1.

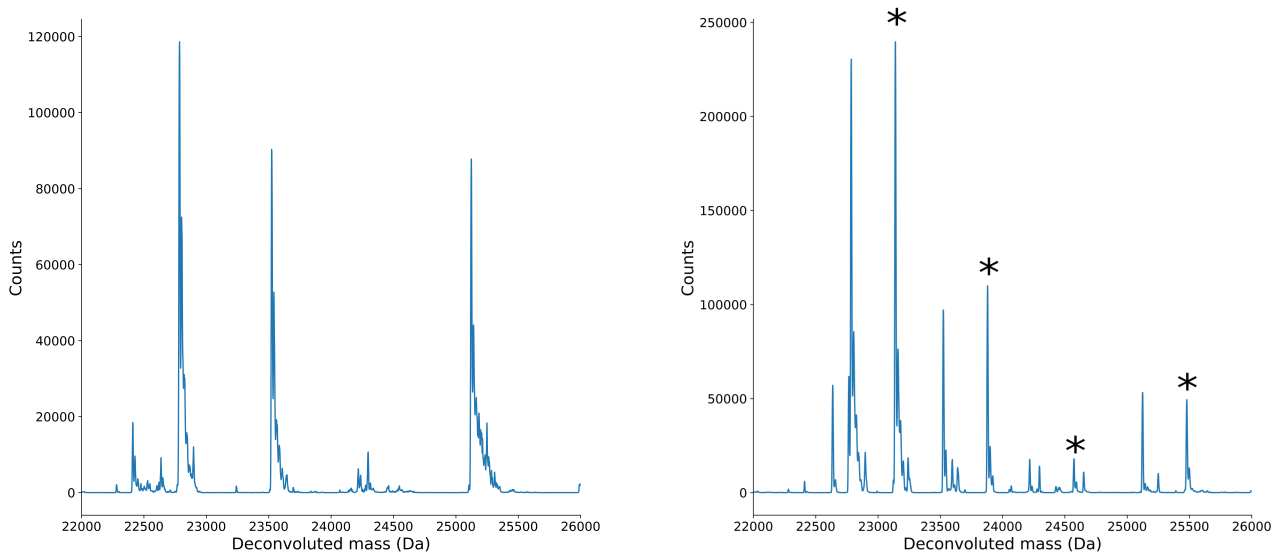
A major caveat was that the NUDT7 post crystallisation samples showed protein degradation. The lower mass degradation products all contained the cysteine which was labelled. This was identified by seeing two peaks separated by 354 Da, the mass of the compound addition, associated to each degradation product. The ratio between the covalently labelled and unlabelled degradation product was the same as in the undegraded peak. Therefore, the signal from all pairs of degraded peaks was used to show the ratio of covalent compound. This is described further in section 5.2.5.2.

#### 5.2.5.1 Sample preparation

Sample preparation for intact MS of crystals is based on the method outlined in [255].

After crystallisation, if a sub-well of a crystallisation plate contained at least one crystal, with a diameter greater than 50 µm, it was considered suitable for use in diffraction studies or post crystallisation intact MS. In some cases where multiple crystals were present, the crystal may first be transferred by means of a crystal harvesting loop into a alternate drop without a crystal, or a freshly prepared drop of either reservoir solution, or Milli-Q water. A single crystal preparation is preferred as any ratio determined from this will not be averaged over possibly varying crystals.

If the sub-well had dried, 2 µl of corresponding reservoir solution was added to the sub-well, before further processing. Qualitative filter paper, grade 1 11 µm (Whatman™), was cut into approximately 10 mm by 5 mm pieces and manipulated by tweezers were used to blot the reservoir solution from the crystal in the sub-well. Ideally blotting could be detected by a visible change in crystal opacity, as viewed on a microscope during blotting. After blotting a 2 µl solution of HPLC grade acetonitrile was applied to the crystal. This was allowed to evaporate until dry, aided by blotting with another piece of filter paper. Application and removal of acetonitrile was repeated three times. 2 µl




**(a)** No covalent labelling.

**(b)** Covalent labelling at 50%

**Figure 5.4:** Degradation of NUDT7 to lower mass species seen in mass spectra of crystals did not effect the proportion of covalent labelling. Peaks are present in 5.4(a) at 25124, 24218, 23525 and 22786 Da are present alongside labelled version, with a corresponding mass shift of +354 Da in 5.4(b), which are indicated with asterisks. The ratio between the degraded peaks was consistent with the ratio between the un-degraded labelled and unlabelled peaks.

solution of 0.1% or 1.0% formic acid was applied to the crystal in the sub-well. Normally this solution was sufficient to dissolve the crystalline matter.

The solution was transferred to an Eppendorf tube, or a 96 well MS storage plate. The formic acid and crystal solution was made up to 5 µl using 0.1% or 1.0% formic acid. Each sample was supplemented to 60 µl total volume with 30% HPLC grade methanol and 0.1% formic acid solution. The intact MS method discussed in section 5.2.3.2 was used for post crystallisation samples unless otherwise indicated.

#### 5.2.5.2 Understanding degradation in NUDT7 post crystallisation

There were multiple confounding factors reducing the signal to noise ratio in MS post crystallisation. One factor was the degradation of NUDT7 into lower mass species as shown in figure 5.4(a).

The covalent labelling of the protein occurs at each degradation peak as shown in figure 5.4(b). The degradation peaks were consistent in mass across samples. Sometimes

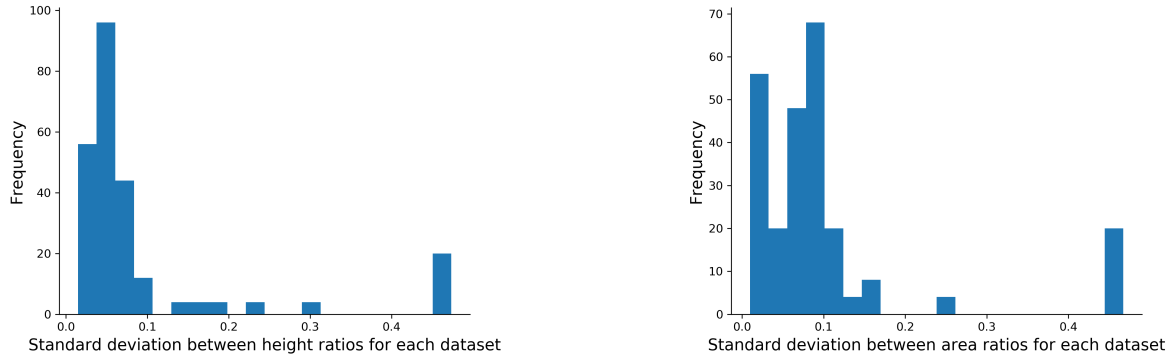
**Table 5.1:** Degradation peaks of NUDT7 present in intact MS after crystallisation with covalent labelling by compound (**6**).

| Unlabelled peak [Da] | Labelled peak [Da] |
|---|---|
| 25124 | 25478 |
| 24218 | 24574 |
| 23525 | 23879 |
| 22786 | 23140 |

adducts (such as oxidation) were detected at greater magnitude than the unmodified protein. Note that the oxidation was likely due to the length of time the protein has been in the crystallisation condition, which was typically 4-7 days, as larger crystals formed after 4 days. However, as the unmodified degradation peak remained, it was measured alongside its adducts. The peak locations are shown in table 5.1.

The reason for degradation is unclear. It may have been due to residual proteolytic activity present in the sample. However, as the degradation only occurred in samples prepared for crystallisation, it is either dependent on the change in conditions which the protein experiences due to the crystallisation buffer, or due to the extended time which the crystalline sample will have to degrade. Attempts to identify cleavage sites using ExPASy peptide cutter [296] and Protein Analysis Worksheet, show that cleavage by formic acid corresponds to all but one of the observed mass shifts. As formic acid was present at a low concentration of 0.1-1% to provide a denaturing environment for the protein, it is unlikely to account for the degradation, as this typically needs 70% formic acid [297]. Proteolytic enzymes such as proteinase K could explain the cleavage at the observed sites, but these would need to be carried along with purification. However, the pre-crystallisation MS show no degradation, this cannot satisfactorily explain the degradation.

However, as the degradation peaks were present for both unlabelled and labelled protein, see figure 5.4, it was hypothesised that these peaks might also be informative on the ratios between the labelled and unlabelled protein. To determine whether the ratio between different sets of peaks was consistent, both the peak height and area including any adducts were extracted from each deconvolution. Peaks were found by searching within $\pm 10$ Da of expected peak locations identified in table 5.1, for a peak using a peak detection algorithm; `scipy.signal.find_peaks` [298]. The peak height




**(a)** Height ratios **(b)** Area ratios

**Figure 5.5:** Histogram of the standard deviation of the ratio between the peak height and area ratios respectively. Mean of 0.064 (height ratios) and 0.072 (area ratios). The relatively low values show that the ratio between sets of peaks are similar for every degradation peak.

was taken from the maxima in the searched region. Due the presence of adducts which could alter in distribution between different samples (i.e. a higher rate of oxidation in one sample) the peak area was taken as the sum of points within -50/+100 Da of the specified peak location. Note that the asymmetry here is to reflect the fact that adducts present were either multiple sodium adducts (+22 Da) or multiple oxidations (+18 Da), both of which appear above the expected mass.

As the degradation peaks had different magnitudes the ratio between covalent labelled species and unlabelled species was calculated:

$$\text{Ratio} = \sum_{\substack{\text{Paired} \\ \text{degradation} \\ \text{peaks}}} w_i r_i \tag{5.1}$$

where $i$ indexes the pairs of degradation peaks, and:

$$w_i = \frac{\text{Labelled peak height} + \text{Unlabelled peak height}}{\sum_{\substack{\text{Paired} \\ \text{degradation} \\ \text{peaks}}} \text{Labelled peak height} + \text{Unlabelled peak height}} \tag{5.2}$$

and:

$$r_i = \frac{\text{Labelled peak height}}{\text{Labelled peak height} + \text{Unlabelled peak height}} \tag{5.3}$$

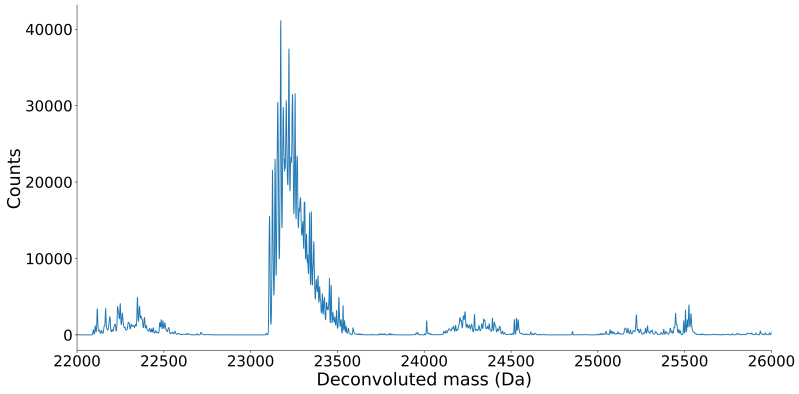



**Figure 5.6:** Ion suppression reduced MS signal to below background noise level. Ion suppression was due to presence of PEG after diffraction experiment. The 23 140 Da peak corresponds to a degraded covalently labelled peak. The mass deconvolution has two peaks shouldering a larger peak. This is a known artefact when a MS signal is suffering from ion suppression.

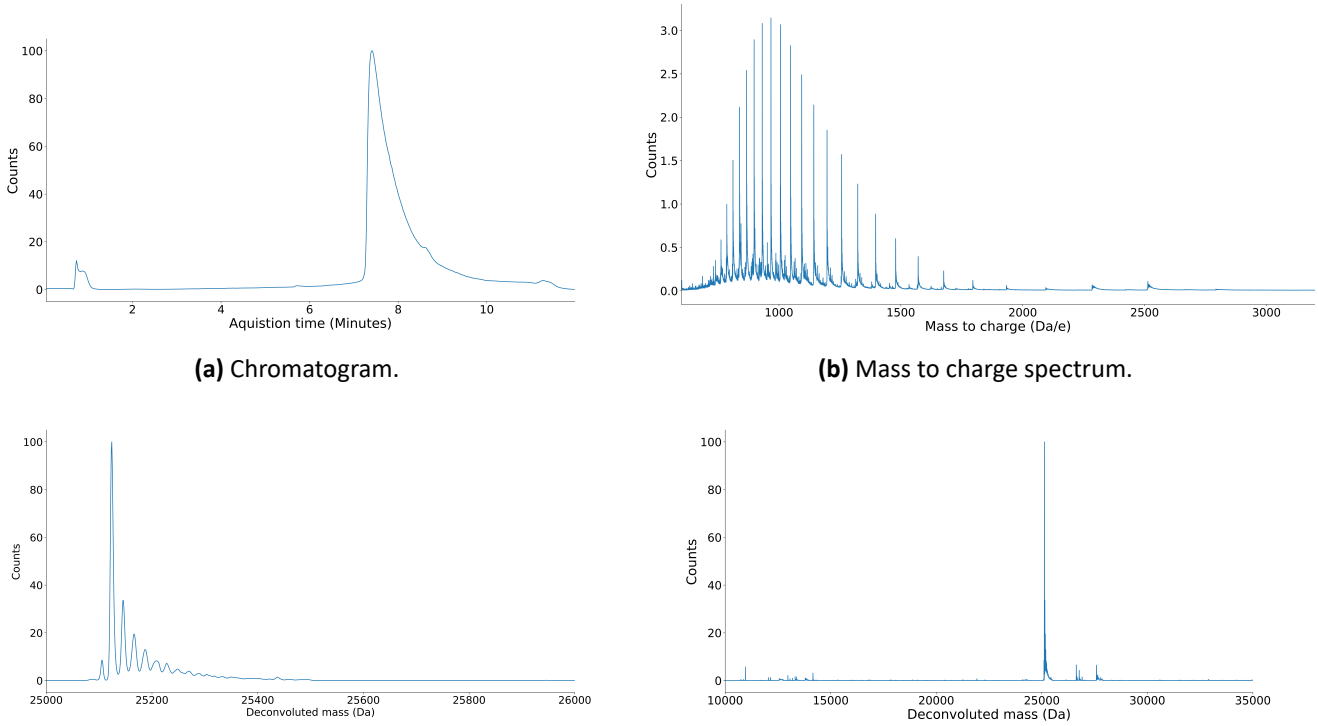
Respectively to calculate the weighted area ratio, the peak area was used instead of the peak height.

Figure 5.5 shows that the ratio of labelled to unlabelled protein was consistent across degradation peaks. As the degradation did not systematically affect the ratio between labelled and unlabelled species, the weighted area ratio was used in the remainder of analysis.

#### 5.2.5.3 Presence of PEG in the crystallisation condition causes ion suppression in intact mass spectrometry

Ion suppression in MS results from the presence of less volatile compounds which change the efficiency of droplet formation and evaporation, which then affects the amount of charged ions in the gas phase which can reach the detector [299]. The presence of PEG, introduced in the crystallisation buffer is known to cause ion suppression in intact MS [300].

Ion suppression of NUDT7, as shown in figure 5.6, was present on samples post X-ray exposure, and for smaller crystals which were not exposed to X-rays. The mass deconvolutions had two peaks shouldering a larger peak. This ion suppression was hypothesised to be due to the presence of an increased relative concentration of PEG due to the smaller amount of protein present. This meant that it was not possible to use




**(a)** Chromatogram.

**(b)** Mass to charge spectrum.

**(c)** Deconvoluted mass, near 25 124 Da peak.

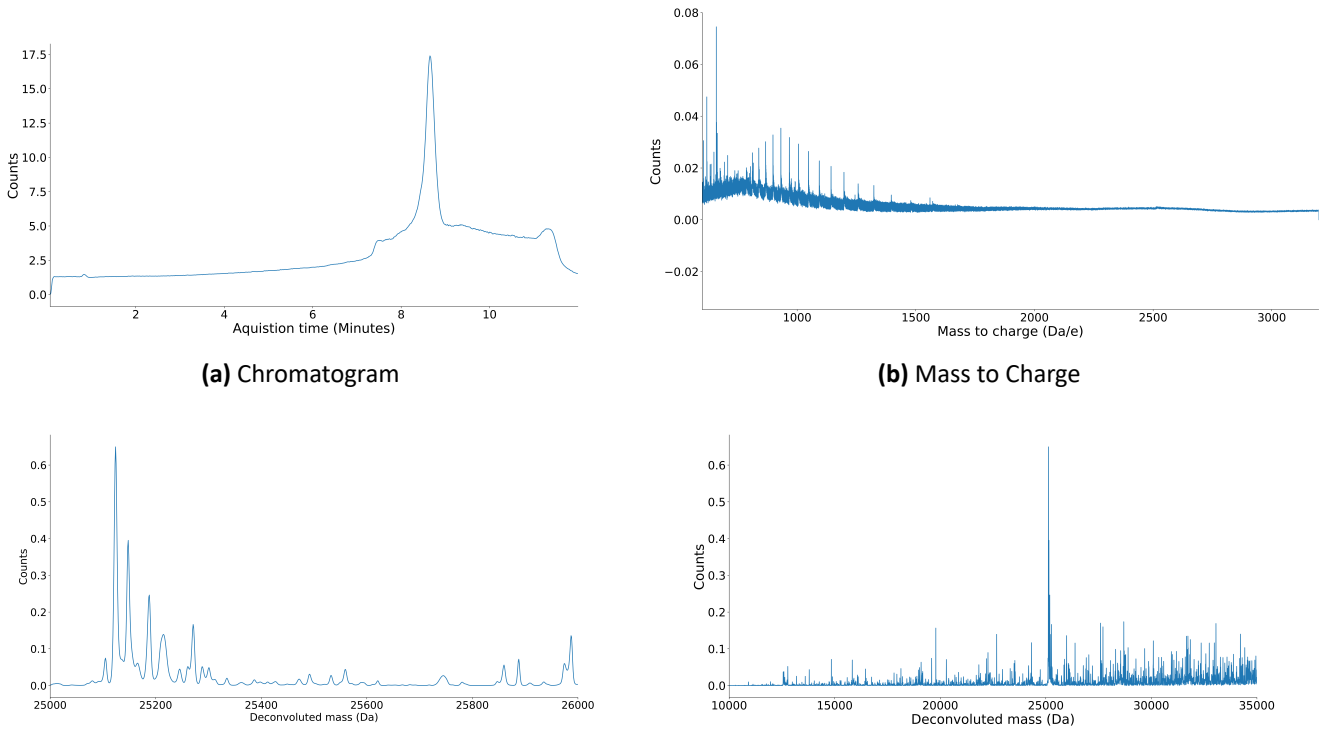
**(d)** Deconvolution of 1 mg ml$^{-1}$ NUDT7A.

**Figure 5.7:** Mass spectra of unlabelled NUDT7 showed high signal to noise. Unlabelled protein measured at 1 mg ml$^{-1}$.

the peak height or area to estimate the amount of covalently labelled species present. Therefore, to measure post diffraction samples either a PEG removal technique, or a chromatographic technique were sought to increase the separation of the elution of PEG and protein.

#### 5.2.5.4 Control experiments: Mass spectrometry of NUDT7 with increasing concentrations of PEG

To test alternate sample preparation methods to remove PEG, a baseline was required. MS samples were prepared by dilution of unlabelled protein to 1 mg ml$^{-1}$ and adding differing concentrations of PEG up to 5%. Figure 5.7(a) shows at this protein dilution there was still slight overloading of the column without PEG, as represented by the shouldered peak. The addition of 0.05% PEG (figure 5.8(a)) showed the dominance of the PEG peak, with the protein eluting slightly later. Although this can be deconvoluted to obtain the protein mass as shown in figure 5.8(d), it had a high signal to noise ratio.




**(a)** Chromatogram

**(b)** Mass to Charge

**(c)** Deconvoluted mass, near 25 124 Da peak

**(d)** Deconvoluted mass

**Figure 5.8:** Adding 0.05% PEG 10 kDa to NUDT7 reduced the signal to noise ratio of the MS. Unlabelled 1 mg ml$^{-1}$ NUDT7 with 0.05% PEG 10 kDa.

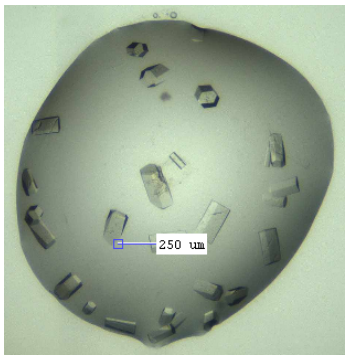
PEG has a distinctive signature in mass charge spectra, of evenly spaced peaks, with curved maxima. Peaks corresponding to a protein species increase in spacing as the mass to charge ratio increases. Figure 5.8(b) shows that PEG can dominate over the protein signal in the mass spectra. Higher concentrations of PEG were also added these are shown in appendix figures E.1 and E.2, in both these cases the PEG strongly altered the signal to noise ratio for the protein.

#### 5.2.5.5 Controls for radiation damage

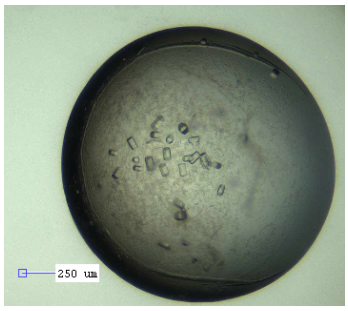
Radiation damage may irreversibly alter the protein structure [301], thereby reducing the likelihood a single species is identifiable by intact MS. An appropriate control experiment to quantify whether radiation damage or sample handling is responsible for observed losses in signal, is to mount and cryo-cool crystals as if they were to be exposed to synchrotron radiation. Then instead of exposure to X-rays, the cryo-cooling is immediately followed by warming up the crystal harvesting loops, and transferring crystals to a 2 µl




**(a)** 14 day image of NUDT7 crystals in 6% PEG 10 kDa,0.1 M bis-tris pH 5.5, 0.1 M ammonium acetate




**(b)** 14 day image of NUDT7 crystals in 25% PEG 3350 0.1 M bis-tris pH 5.5




**(c)** 24 hour image of NUDT7A crystals in 1 M imidiazole condition

**Figure 5.9:** Alternate crystallisation conditions include conditions with lower molecular weight PEG, and no PEG.

drop of Milli-Q water. The blotting and washing process described in section 5.2.5.1 was then applied, followed by the longer intact MS method. These control datasets show similar rates of failure due to crystal transfer and low signal to noise as those after diffraction. Therefore, it is expected these transfer failures are larger factors than any effects of radiation damage, as to why there was no signal detected in the post diffraction crystals.

#### 5.2.5.6 Exploration of additional crystallisation conditions

One way to remove the PEG suppression would be to move to a non-PEG containing crystallisation condition. This was explored by undertaking matrix seeding into a series of coarse screens typically used at the SGC. Microseeding stock was prepared as described in section 5.2.2.3.

Crystal trials identified a single crystallisation condition not containing PEG: 1 M imidazole (figure 5.9(c)). Without cryoprotectant these crystals showed diffraction to 2.73 Å with a similar unit cell to the previous condition. Initial optimisations with a fine screen could not replicate crystals, so instead other PEG removal techniques were tried.

Another option would be to use condition with alternate PEG species which may be easier to remove. Multiple conditions containing lower molecular weight PEG species

were found to lead to crystals, figure 5.9(b), with a similar morphology to those crystallised with PEG 10 kDa, figure 5.9(a). Lower molecular weight PEG may be easier to separate using techniques such as dialysis. Initial dialysis tests, using low volume Swiss CI dialysis plates, were unsuccessful. Due to time constraints these crystals could not be optimised to repeat these experiments with an alternative crystal system.

#### 5.2.5.7 Alternate sample preparation: Washing crystal with isopropanol alcohol

Replacing the acetonitrile washing step with isopropanol alcohol was attempted. As PEG is a hydrophilic molecule, an improvement in extraction from around the crystal by the hygroscopic isopropanol alcohol was expected. Initial attempts showed no improvement in failure rate or signal to noise over use of acetonitrile as the crystal washing solvent.

#### 5.2.5.8 Alternate sample preparation: Titanium dioxide spin columns

Another approach for separating the PEG and protein is to use a column material that has a different selectivity profile. Development can be done on the sample offline from the uHPLC, later using a similar selectivity material in an on-line uHPLC column.

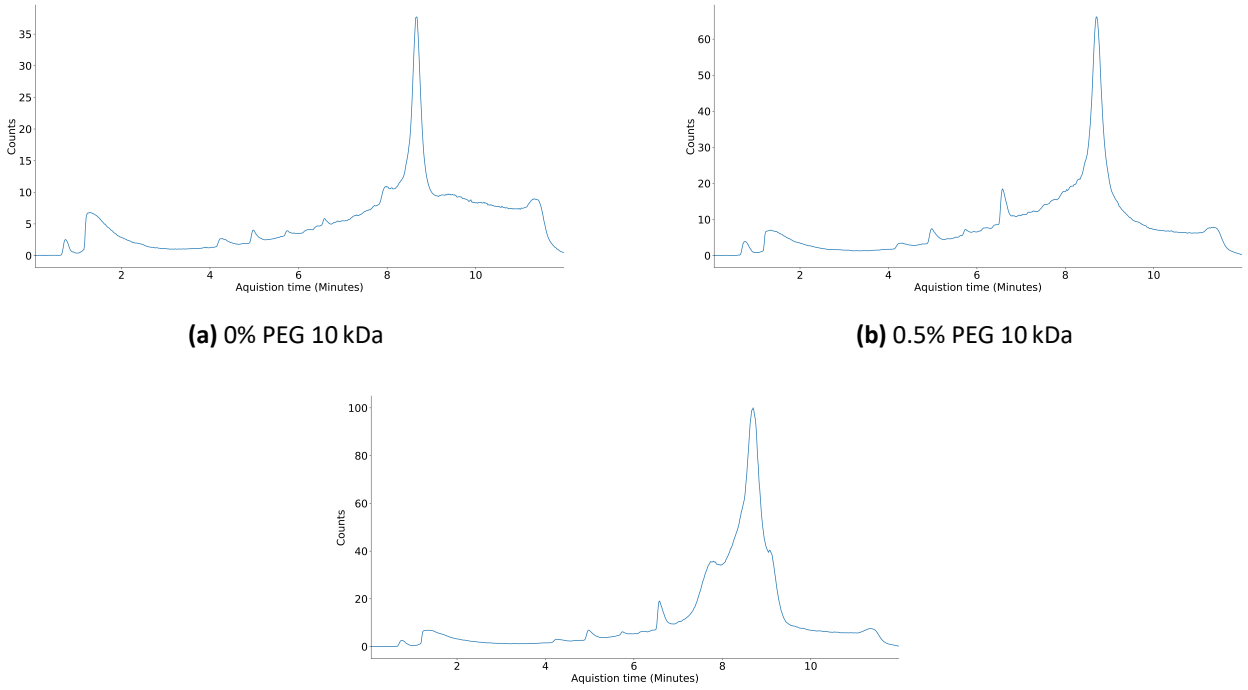
Titanium dioxide has been shown to separate PEG 4000 Da [252], detailed in appendix section E.1. This method was applied to NUDT7 protein samples as prepared in section 5.2.5.3. These samples showed no signal for the protein species in either mass spectra or deconvolution, as shown in figure 5.10.

#### 5.2.5.9 Alternate sample preparation: $C_4$ and $C_{18}$ ZipTips™

Another change in selectivity to seperate the PEG from NUDT7 may have been achieved by using $C_4$ and $C_{18}$ packed ZipTips™ rather than the $C_3$ columns used in the HPLC methods.

Samples of 0.05%, 0.5% and 5% PEG and 1 mg ml$^{-1}$ NUDT7 were prepared using the ZipTips™ .

To use ZipTips™ the sample was prepared to have a final concentration of 0.1% Trifluoroacetic acid (TFA). The ZipTip™ was affixed to 10 µl pipette, which was set to aspirate maximally. The tip is wetted by aspirating and dispensing 100% acetonitrile to waste repeatedly. The tip was washed with Milli-Q supplemented with 0.1% TFA (to waste), repeated twice. The sample was bound to the ZipTip™ by fully depressing the




**(a)** 0% PEG 10 kDa

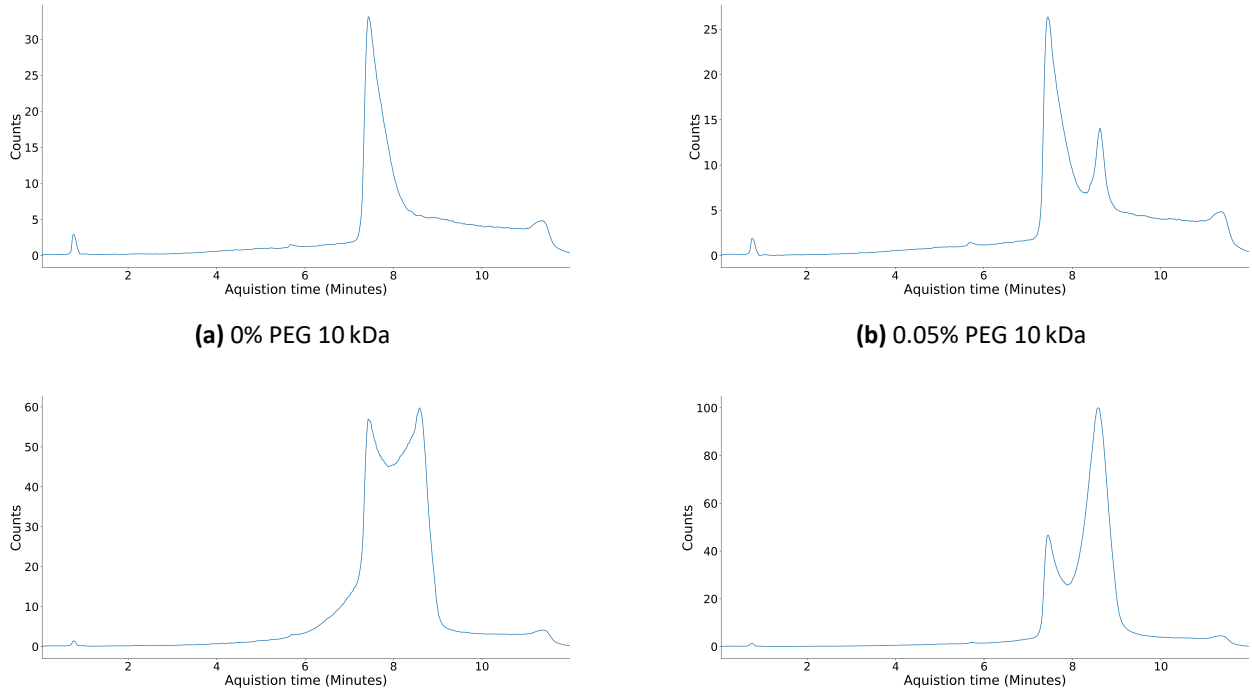
**(b)** 0.5% PEG 10 kDa

**(c)** 5% PEG 10 kDa

**Figure 5.10:** Titanium dioxide spin columns did not displace PEG 10 kDa. Chromatograms for 0.05%, 0.5% and 5% PEG and 1 mg ml$^{-1}$ NUDT7 after preparation by titanium dioxide spin column. All chromatograms only show the presence of PEG, and not the protein.

pipette plunger to a dead stop. The tip was then washed with Milli-Q supplemented with 0.1% TFA (to waste), repeated twice. Elution of sample was achieved with 5 µl of 50% methanol with 1% formic acid by aspiration and dispense, which was repeated at least 3 times.

Figure 5.11 shows chromatograms for PEG containing samples of 1 mg ml$^{-1}$ NUDT7 processed by $C_4$ ZipTips™ , and figure 5.12 shows the same samples processed by $C_{18}$ ZipTips™. Note that the ZipTips™ do not seem to introduce any loss of signal when processing protein without PEG present, as can be seen by comparing figure 5.11(a) to figure 5.7(a). Note also that a reasonable amount of separation of PEG is achieved, comparing figure 5.11(b) to figure 5.8(a), figure 5.11(c) to figure E.1(a) and figure 5.11(d) to figure E.2(a) respectively we see an improved separation of the PEG and protein peaks. The separation of PEG with the $C_{18}$ ZipTips™ is not particularly improved. The improved performance of $C_4$ over $C_{18}$ ZipTips™ is potentially due to the similar selectivity of the in-line $C_3$ column used in the uHPLC methods.




**(a)** 0% PEG 10 kDa

**(b)** 0.05% PEG 10 kDa

**(c)** 0.5% PEG 10 kDa

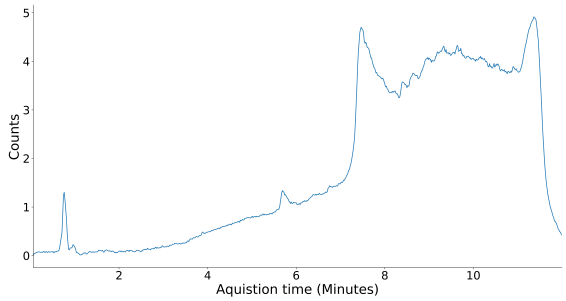
**(d)** 5% PEG 10 kDa

**Figure 5.11:** $C_4$ ZipTip™ reduced the amount of PEG present with NUDT7 crystals which were dissolved using formic acid, after the crystallisation drop contents were wicked away.. Chromatograms after $C_4$ ZipTip™method applied to NUDT7 with varying percentages of PEG introduced.

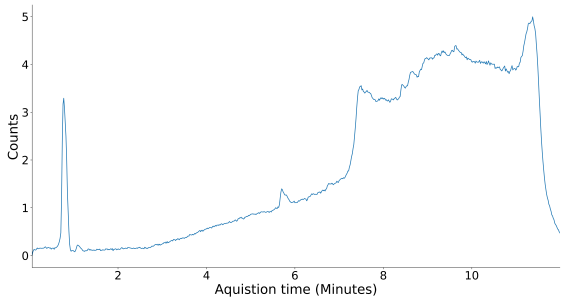
#### 5.2.5.10 Changing uHPLC methods

Although the twelve-minute method outlined in section 5.2.3.2 has been shown in other experiments to be robust to additions of small amounts of PEG, it was designed for the separation of detergents from membrane proteins purified using detergent mixes. Therefore, the separation of PEG and NUDT7 may be increased by varying the uHPLC method. The time taken of gradient step from 30% methanol to 95% methanol was increased from 7 to 11 minutes. Figure 5.13 shows the alerted chromatogram, there is not a significant increase in the separation of the PEG and NUDT7.

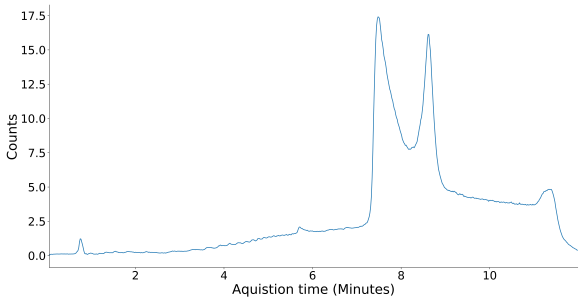
As there was little improvement in separating power with gradient length, a systematic search of uHPLC methods would need to be carried out. This would like include a search of orthogonal columns [302] to look for a differing selectivity profile, which was out of scope for this project. Development of a uHPLC method, could be made faster by



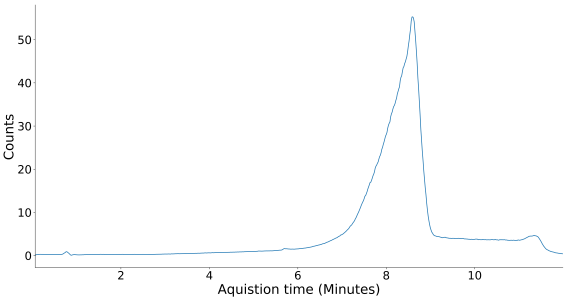

**(a)** 0% PEG 10 kDa




**(b)** 0.05% PEG 10 kDa

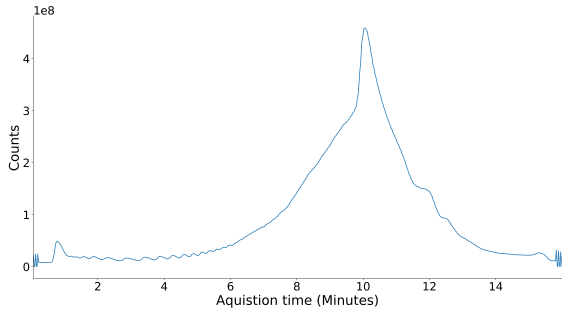



**(c)** 0.5% PEG 10 kDa

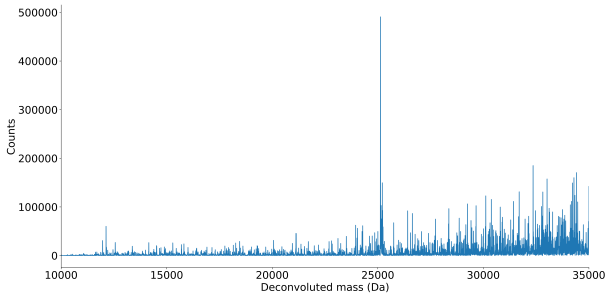



**(d)** 5% PEG 10 kDa

**Figure 5.12:** $C_{18}$ ZipTip™ did not reduce presence of PEG. Chromatograms after $C_{18}$ ZipTip™ method applied to NUDT7 with varying percentage of PEG.




**(a)** Chromatogram




**(b)** Deconvolution

**Figure 5.13:** Increasing length of methanol gradient did not improve PEG separation. 5% PEG added to 1 mg ml$^{-1}$ NUDT7 with gradient increased from 7 minutes to 11 minutes. No additional separation of PEG, although the deconvolution does show the protein.

utilising a shorter run by increasing the starting concentration of the methanol in the gradient, as both PEG and NUDT7 elute quite late.

### 5.2.6 Crystallographic data collection and refinement methods

#### 5.2.6.1 Data collection and processing

Initial data collection of the co-crystals was done without a cryoprotectant, as previous crystals of NUDT7 had performed well without. Noticing a lower than typical diffraction resolution, a study of potential cryoprotectants for the co-crystals was undertaken: 2 µl of either DMSO, ethylene glycol or glycerol at increasing concentration of 10%, 20% and 30% were added to the crystal well. Through screening at the SGC and sending a small number of crystals to be screened on I03 (DLS), DMSO at 30% and glycerol at 20% were deemed to be appropriate cryoprotectants and used for the remainder of the co-crystal datasets. Both DMSO and glycerol were used, as if either caused reduction of intact MS signal, the other subset could still be used for further analysis.

Screening diffraction data was collected in-house at the SGC on a Bruker Microstar generator equipped with an APEX II detector. Diffraction datasets were collected on beamline I03 and beamline I04-1 at the DLS. Datasets collected on beamline I03 were collected with a smaller beam size to mitigate the dose being spread across the whole crystal, as the dose may affect the ability to collect mass spectrometry data post diffraction. This could not be studied due to the overall low signal of post diffraction intact MS.

Diffraction datasets were automatically processed by software pipelines at the DLS [109]. The initial refinements and map calculations were carried out with DIMPLE [6]. Refinement and model building were carried out using REFMAC5 [8], Phenix [217], Buster [222] and COOT [3] respectively.

#### 5.2.6.2 Crystallographic refinement

A model of NUDT7 without the covalent compound was supplied to DIMPLE [6], for all datasets collected. Then the atom positions of the bound compound were extracted from the originally modelled complex structure and copied to each model. This allowed processing by the `giant.quick_refine` script. The ligand description CIF defining the

covalent bond to the ligand was generated using JLigand [303]. The corresponding link record was copied into each PDB file such that the model could be refined taking the covalent bond into account.

#### 5.2.6.2.1 Superposed refinement

As described in section 1.6 this allows refinement with REFMAC5, phenix.refine and Buster considering the crystal state as a superposition of the bound and unbound states of the protein. Refinement was carried out as outlined in section 3.2.1.5.

#### 5.2.6.2.2 Non-superposed refinement methods

One aim was to determine whether modelling the protein-ligand complex as a superposed model affected the correlation of occupancy to the ratio of labelled protein determined by intact MS. Therefore, refinements with and without the unbound state represented should be compared. Refining without the bound state utilises the non-superposed refinement protocols; specifically those using REFMAC5, phenix.refine and Buster, which are outlined in section 3.2.1.6. The model used for refinement in these non-superposed methods is the single conformer model that can be produced by running the `giant.split_conformations` script on the superposed model.

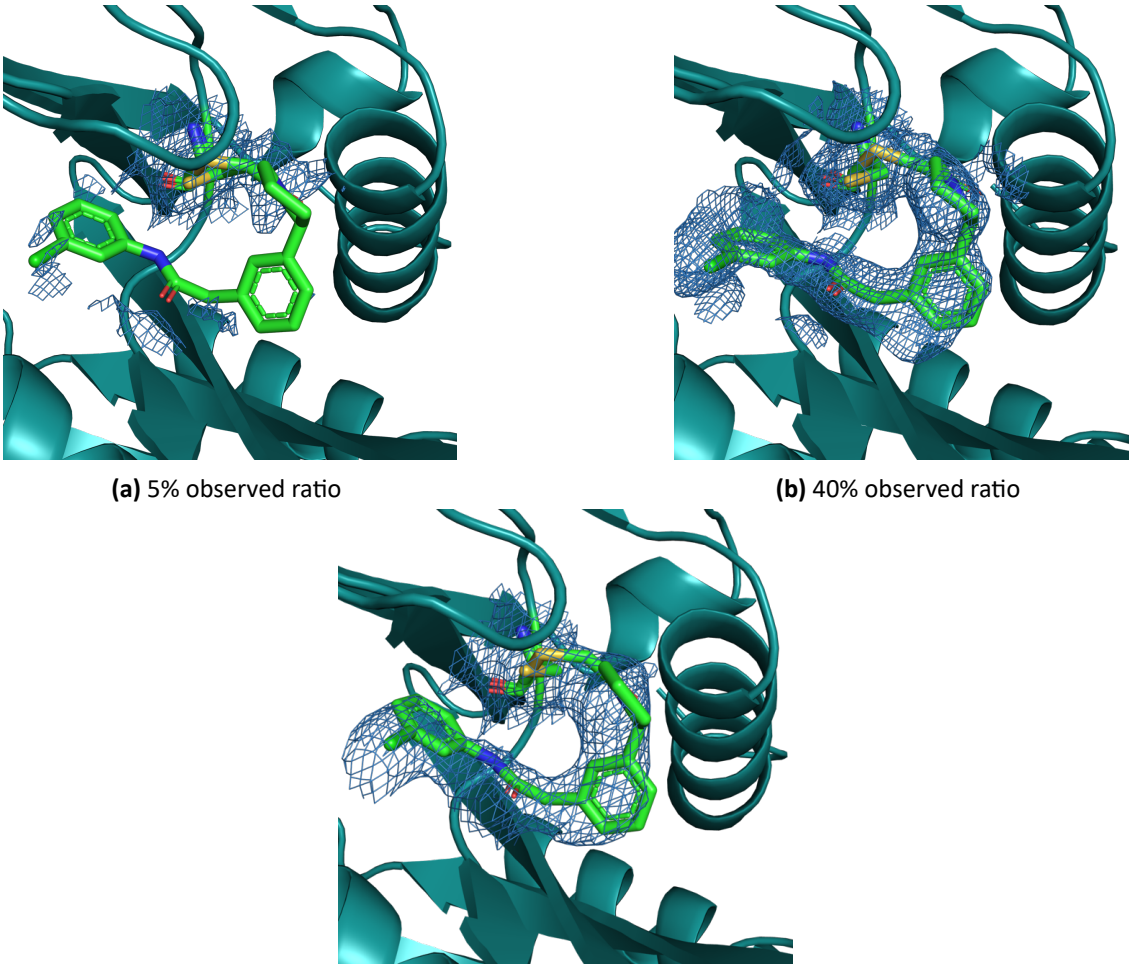
#### 5.2.6.2.3 Exhaustive search

Alongside superposed and non-superposed refinement, the exhaustive search as defined in 3.2.2 was applied. The input model is the superposed model used for refinement in `REFMAC5`.

## 5.3 Results

### 5.3.1 Crystallographic occupancy was correlated with ratio of covalent compound detected with intact mass spectrometry

Initial refinement of datasets from crystals with increasing ratios of protein covalently labelled with compound (**6**) showed increasing electron density support for the ligand,



**(a)** 5% observed ratio

**(b)** 40% observed ratio

**(c)** 100% observed ratio

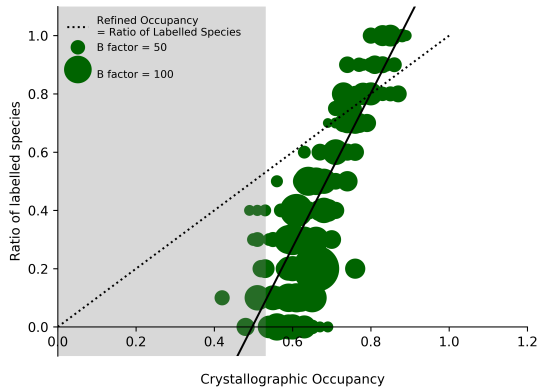
**Figure 5.14:** Covalent binder (compound (**6**)) was identified in the electron density of NUDT7 crystals at 40% and 100% ratio of compound present, but not at 5%. REFMAC5 with a superposed model was used for refinement. Electron density ($2mF_{obs} - DF_{calc}$) map contoured at 1$\sigma$, shown up to 2.2 Å from ligand atoms.

as shown in figure 5.14. This initial work was followed by collecting diffraction data on a larger number of crystals across the range of ratios of covalently labelled protein.

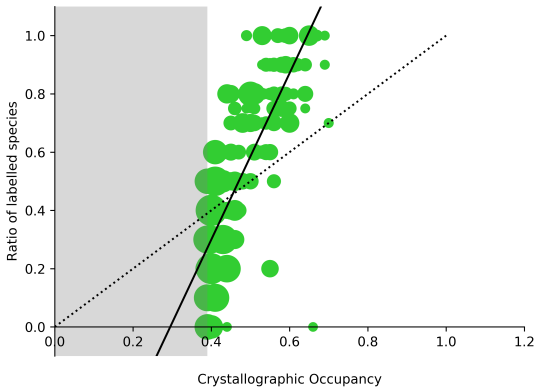
Using this larger set of crystals, a comparison of crystallographic occupancy of the ligand to the ratio of labelled to unlabelled protein was produced. This was obtained by post crystallisation MS, as shown in figure 5.15.

For each refinement method it was assumed that the ligand was present, and the occupancy was allowed to vary. If the ligand were not present, it was expected that it's occupancy should refine to a low value. The same model was used for all refinements.

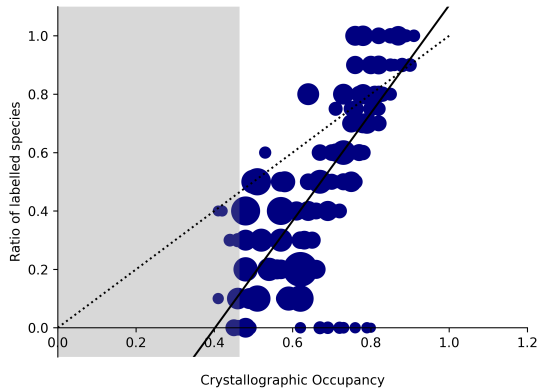
The root mean squared deviation (RMSD) reported in figures 5.15 and 5.17 measures



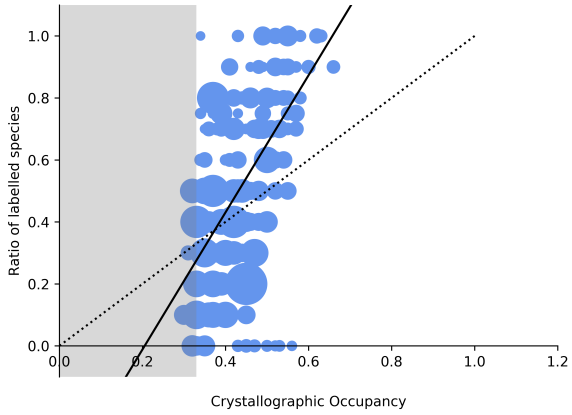

**(a)** Buster refine. RMSD 0.22



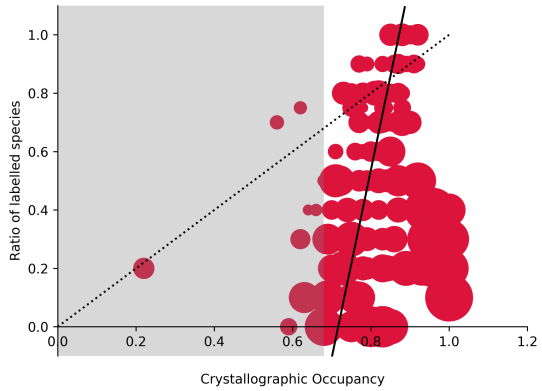

**(b)** Buster refine superposed RMSD 0.21



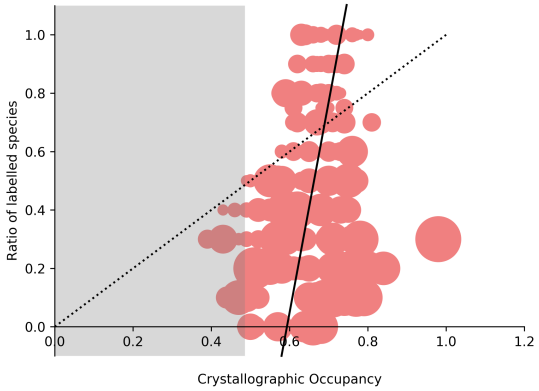

**(c)** `phenix.refine` RMSD 0.19




**(d)** `phenix.refine` superposed RMSD 0.22




**(e)** `REFMAC5` RMSD 0.33




**(f)** `REFMAC5` superposed RMSD 0.24

**Figure 5.15:** Crystallographic occupancy increased with the proportion of compound measured by MS(continued on next page).

**Figure 5.15:** (continued) Crystallographic occupancy was fixed to have the same value across all atoms in the ligand/ bound state residues in each individual refinement. B factor is shown by the size of the plotted points. The dashed line represents where the refined occupancy is equal to ratio that is detected from intact MS of crystals grown with the same ratio of covalently labelled protein present. The solid line represents a linear fit to the occupancies of each refinement. The grey shaded area corresponds to the 5$^{th}$ percentile of the data. The reported root mean squared deviation (RMSD) measured the difference between the crystallographic occupancy and the expected ratio given by the proportion of covalent compound.

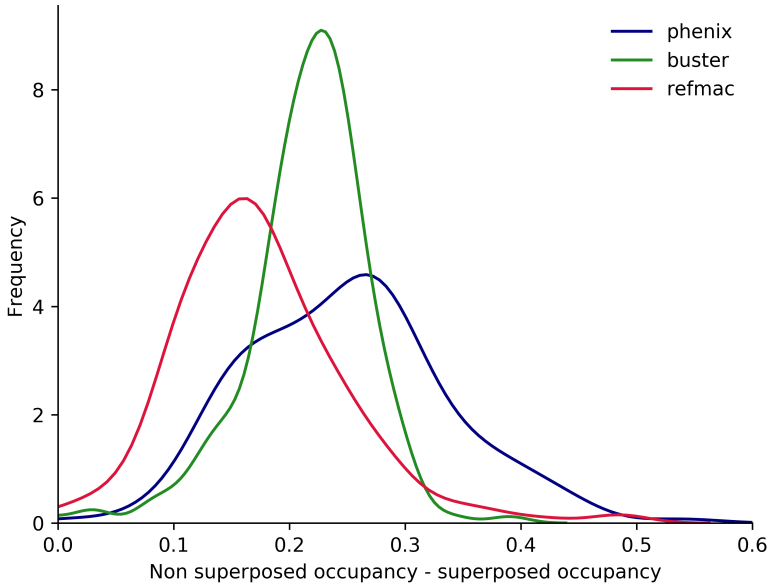



**Figure 5.16:** Ligand occupancies refined with superposed methods were lower. The change between superposed and non-superposed was near normally distributed for Buster 0.22$\pm$0.052 (standard deviation). However artefacts in refinement meant that `phenix.refine` 0.25 $\pm$0.085 and `REFMAC5` 0.17 $\pm$0.08 varied more between the superposed and non-superposed methods.

the difference between crystallographic occupancy and the expected ratio given by the proportion of covalent compound. A low value corresponds to a good fit between the data and the expectation (as shown by the dashed line). `phenix.refine` shows the lowest RMSD, but still has a high minimal occupancy. The shaded regions show the 5$^{th}$ percentile of occupancies observed, this highlighted that all refinement programs did not refine to low occupancy and tend to put uncertainty into the B factor instead.

Comparing the refinement of superposed and non-superposed models, it was shown that the superposed refinement was lower in occupancy in all programs. A consistent gap between the superposed and non-superposed refinement shown in figure 5.16

shows that ground state was occupied consistently across different crystals compared to the bound state.

The exhaustive search algorithm was also tested, as seen in figure 5.17(a). This also showed a minimal occupancy of 0.3, and a reasonably consistent B factor across all refinements. The range of occupancies (0.3 to 0.8) and correlation is quite similar to the other refinement methods.

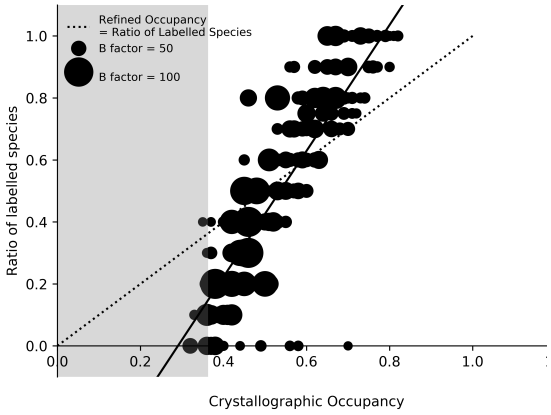
### 5.3.2 Restraining B factor allowed post crystallisation mass spectrometry to calibrate refinement methods

As the refinement methods had not refined low occupancies as would have been expected by the proportion of covalent compound measured by intact MS, restraining the ligand B factors was explored.

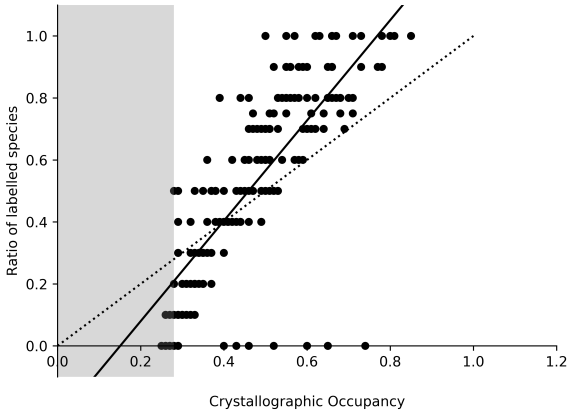
Applying the exhaustive search algorithm, (exploring the flatness of difference density map over occupancy and B factor values), whilst keeping the isotropic B factor of the ligand constant is shown in figure 5.17(b). The average B factor of the local residues (from a single example structure) was applied to all refinements. The reported root mean squared deviation (RMSD) is lower for the exhaustive search method and remains low when the occupancy is fixed. The B factor in `phenix.refine` can be fixed by selecting all other atoms to refined with individual isotropic B factors, this was applied alongside the superposed refinement restraints. This showed a lower RMSD than its non-restrained counterpart. This suggests that improvements to the model are needed to improve refinement, namely the application of B factor continuity restraints for proximal atoms and a better model of ligand occupancy.

## 5.4 Discussion and future work

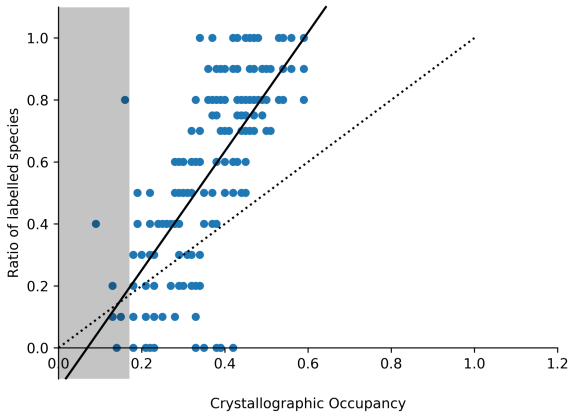
The central hypothesis of this chapter was that crystallographic occupancy of a ligand within a protein-ligand complex would be correlated with a non-crystallographic quantification of the amount of compound present. This correlation exists, and has been shown using intact MS of crystals. The exploration of different refinement methods to obtain crystallographic occupancy show that the parametrisation of the protein-ligand




**(a)** Exhaustive search. RMSD 0.16




**(b)** Exhaustive search fixed B factor. RMSD 0.17




**(c)** `phenix.refine` superposed fixed B factor. RMSD 0.18

**Figure 5.17:** Fixing B factor of ligand improved fit to ratio of covalently labelled protein. For exhaustive search and phenix superposed, the B factor of the ligand was not refined, occupancy remained variable. In the case of `phenix.refine` it also reduced the occupancy which was sampled, as uncertainty was prescribed to the B factor, this fitted better than other refinement methods with an RMSD of 0.18, but not as well as the exhaustive methods. The grey shaded area corresponds to the $5^{th}$ percentile of the data. The reported root mean squared deviation (RMSD) is from the reported crystallographic occupancy to the expected ratio given by the proportion of covalent compound.

complex is a non-trivial task, and may account for some of the large biases in the protein databank for unitary occupancy ligands.

Note that Buster refine and `phenix.refine` typically have more consistent B factors, and reach lower occupancy values. However, all of the methods can fail in producing the most appropriate model.

Fixing the B factor to a low value, similar to that of the nearby residues, refining with `phenix.refine` using superposed models, and exhaustive search produced a better fit to the ratio of covalent compound, see figure 5.17. Therefore, it is expected that applying B factor continuity restraints to residues near the ligand is likely to improve occupancy refinement. An alternate refinement method to apply in `REFMAC5` would be to iteratively adjust the ligand atom occupancy until their atomic temperature factors refine to the value of the surrounding residues [256]. In future work it may be possible to produce and refine a better structural model by altering the weights on the interatomic B factor restraints, and utilising B factor continuity restraints from proximal atoms, rather than fixing the B factor. Understanding the best model parameterisation that fits this calibration data should inform how to model ligand occupancy and B factor without such detailed calibration data.

The exhaustive search method produces smoothly varying function of B factor and occupancy near each minima of difference density. Using the exhaustive search method the correlation between B factor and occupancy is similar to refinement programs. This suggests that the refinement is getting a good result given the model. Therefore, future work should focus on providing tools and metrics to choose the most appropriate model and corresponding restraints rather than adjusting underlying refinement algorithms. Altering refinement to a bond-based parametrisation [287] rather than atomic parametrisation, may improve the understanding of appropriate B factor models for ligand binding.

One limitation was that the crystallisation condition with PEG present had insufficient signal to noise ratio with post diffraction samples. Future work in this area would likely focus on repeating these results with a new covalent protein-ligand complex with crystallisation conditions that were more amenable to intact MS. Alternatively, the experiment could be repeated with NUDT7 crystals from newly found non-PEG containing

crystallisation conditions. Further work to separate PEG would involve exploring the selectivity profiles of different uHPLC columns, alongside developing uHPLC methods.

To address the question of whether there is a correlation between low occupancy ligands and low binding affinity, which initiated this work, non-covalent ligand binding would need to be studied. Specifically, a method where the crystallographic occupancy of non-covalent ligands can be compared to the amount of compound present as detected by an orthogonal method. Native MS, keeps the protein in a non-linearised state, and therefore the binding of non-covalent species can be studied. However, studying crystalline protein using native MS is not possible. Therefore, techniques such as micro-spectrophotometry [304, 305] using Raman [306, 307], infrared or UV-Visible [308], to monitor the amount of compound in a single crystal could be adapted to study the relationship between compound present and refined crystallographic occupancy.

# 6

# Detecting correlated motions by sampling rotamer distributions across many electron density maps

## Contents


**6.1 Introduction . . . . . 176**
**6.2 Methods . . . . . 177**
6.2.1 Datasets . . . . . 178
6.2.2 Scaling structure factor amplitudes allowed comparison of electron density across multiple datasets . . . . . 178
6.2.3 Ringer routine on fragment screening datasets . . . . . 179
6.2.4 Pairwise analysis of datasets . . . . . 180
6.2.5 Clustering Ringer plots over datasets for each residue . . . . . 181
6.2.6 Scoring to detect sub-populations of similar datasets . . . . . 183
**6.3 Results . . . . . 185**
6.3.1 Idealised rotamers were detected in multiple datasets . . . . 185
6.3.2 Conformational heterogeneity between datasets varied by residue . . . . . 185
6.3.3 Gaussian fits approximate the Ringer plot . . . . . 187
6.3.4 Ringer across multiple datasets detected correlated motion in *T. cruzi* FPPS case study . . . . . 187
6.3.5 Sub populations of datasets such as bound ligands were not detected by automated method . . . . . 188
**6.4 Discussion and future work . . . . . 190**
6.4.1 Weaknesses in the multiple dataset ringer routine and opportunities for improvement . . . . . 191
6.4.2 Future work . . . . . 193

## 6.1 Introduction

This chapter aimed to detect correlated motions between residues, by sampling electron density across multiple datasets with a structural model. Correlated motions are hypothesised to be linked to allosteric signal propagation.

An existing method, Ringer [245] forms the basis of this method. It samples the electron density around protein side-chains by rotating the side chain about its dihedral angle. This identifies whether there is electron density evidence for an alternate conformation. Extending the Ringer sampling method across multiple datasets was hypothesised to allow detection of residues which frequently have alternative conformations. Datasets which have multiple residues with alternate conformations in common were expected to belong to a distinct sub-population. Groups of residues with alternate conformations in common were hypothesised to represent correlated motion between those residues.

A case study of multiple dataset Ringer applied to fragment screening datasets of *T. cruzi* FPPS, is discussed in section 2.3 and in appendix section B.2. This target has existing allosteric binders [268, 269]. Multiple fragment binders detected in the XChem screening have conformational changes of the protein near the ligand binding site and in the distal active site. Ringer [245] applied across the *T. cruzi* FPPS fragment screening datasets was sufficiently sensitive to detect ligand bound datasets as outliers at both the ligand binding site and distally near the active site metal ion.

The major aim of the multiple dataset Ringer method was to detect correlated motion of residues by analysing which residues had outlying Ringer plots. The approach explored, whereby the Ringer plots were fitted by a model of three gaussian curves and then hierarchically clustered, did not select sub populations of datasets with surprising alternate conformations. However, the *T. cruzi* case study shows that there may be more appropriate methods to determine related residue conformations.

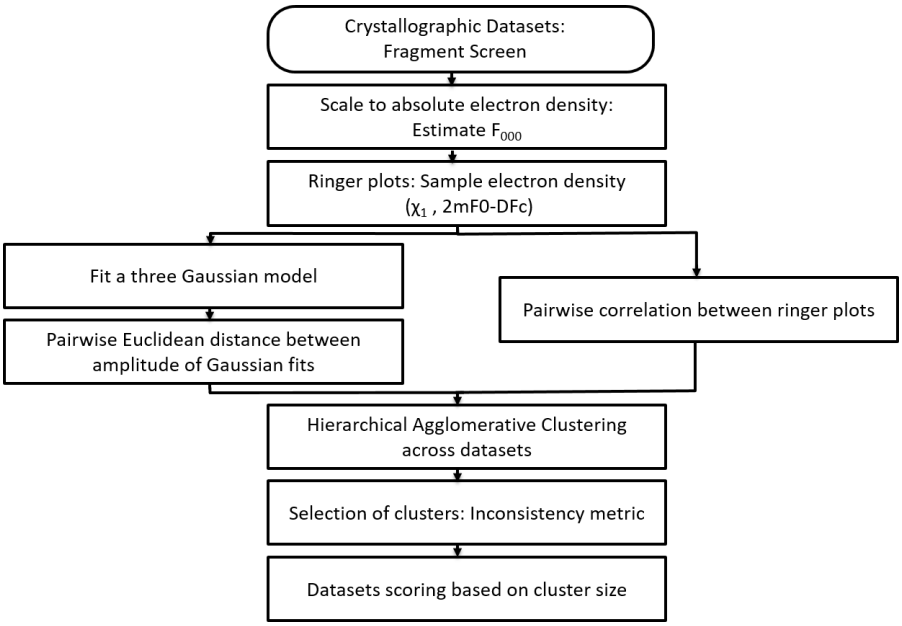



**Figure 6.1:** Multiple dataset Ringer method summary: Crystallographic fragment screening data which has been locally aligned is used as this is near isomorphous and each residue can be compared. This is scaled to a common scale, and then the Ringer routine samples each residue in each dataset. The results of the Ringer routine are then compared for each residue. This is done either by fitting a three Gaussian model and taking the Euclidean distance between pairs of datasets, for the amplitudes of the fits, or by Pearson correlation coefficient of pairs of datasets. Each pairwise comparison is then clustered using hierarchical agglomerative clustering. These clusters are selected out using the inconsistency metric, and finally the datasets are scored based on the cluster size.

## 6.2 Methods

In order to test the hypothesis that correlated motion can be detected in sub populations of datasets, a method for comparing Ringer plots and extracting datasets which were outliers across multiple residues was developed. Figure 6.1 summaries the multiple dataset Ringer routine.

To extract similarities between datasets pairwise comparisons were generated followed by clustering the comparisons for each residue. Two pairwise metrics, Pearson correlation coefficient of the Ringer plots, and the Euclidean distance between fit parameters for a three Gaussian fit to the Ringer plots were compared. The clustering was hypothesised to group similar datasets for each residue. The per residue clustering of datasets was combined using a score that assessed in how many residues was that dataset in a small cluster. It was envisaged that this score would be able to detect

datasets which were consistently outliers, and thus had a set of residues which indicated correlated motion.

### 6.2.1 Datasets

A crystallographic fragment screen of a lysine demethylase (JMJD2D) with 249 datasets, of which 24 datasets have bound ligands identifiable using PanDDA[1] was used. This is accessible on Zenodo at `10.5281/zenodo.290220`. We also analysed our fragment screening results for *T. cruzi* FPPS outlined in appendix B.2.

The fragment screening datasets used were aligned in the same reference frame by PanDDA, so models did not need further alignment to be able to compare Ringer plots. The electron density maps required scaling to be on a comparable scale. Ringer was run on these rescaled electron densities.

### 6.2.2 Scaling structure factor amplitudes allowed comparison of electron density across multiple datasets

In X-ray crystallography the forward scattered reflection, with structure factor $F_{000}$, is obscured by the direct beam, and not measurable. The structure factor of this reflection is equal to the total number of electrons in the unit cell, including the contribution from the disordered solvent molecules. Therefore, it is not possible to represent electron density maps on an absolute scale directly from experimental measurements alone.

To estimate $F_{000}$, the total number of electrons from the model and bulk solvent model in the unit cell were summed, as outlined in [197]. The experimental structure factors $F_{\text{obs}}$ were scaled to the structure factors from the model ($F_{\text{calc}}$ which are on an absolute scale) and the estimate of $F_{000}$ was added to each map voxel. In sigma-scaled maps, zero corresponds to the average electron density, whereas in these maps (END maps in [197]) zero corresponds to vacuum.

The scaling of the structure factor amplitudes leads to scaling of the corresponding electron density maps. Therefore, the electron density maps are on the same (comparable) absolute scale. This scaling was required such that the amplitudes of the Ringer plots can be compared between different datasets.

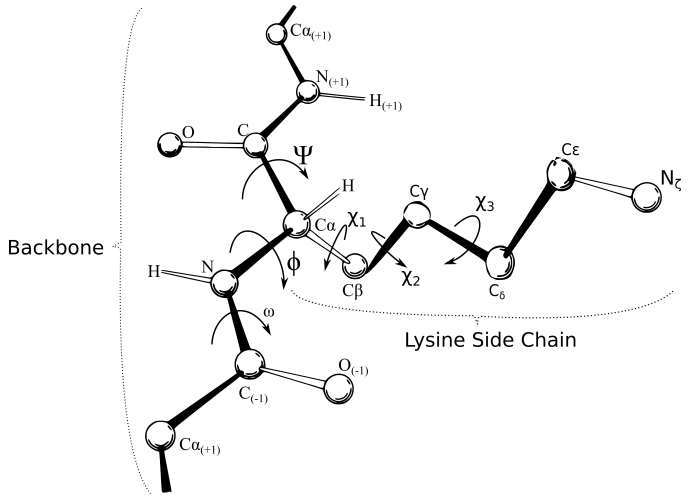



**Figure 6.2:** Ringer samples electron density by rotating side chain models around their side chain dihedral angles. Schematic diagram depicting heavy atoms of a lysine side chain. Shows the backbone dihedral angles $\psi$ and $\phi$ as well as the positional labelling ($C_\alpha$,$C_\beta$…) of atoms in the side chain. The arrows represent the bond which form the axis of the dihedral angle, so that in the case of $\chi_1$ the angle is between the nitrogen atom in the backbone and the heavy tom at the $C_\gamma$ position. Adapted from `https://commons.wikimedia.org/wiki/File:Protein_backbone_PhiPsiOmega_drawing.svg`

### 6.2.3 Ringer routine on fragment screening datasets

As an electron density map averages the many copies of the asymmetric unit of the protein, a single electron density map can show evidence of multiple rotamer positions. The Ringer routine [245] can be used to detect alternate side chain conformers within an electron density map.

In the Ringer routine the protein backbone atoms are fixed in their originally modelled positions. The electron density is sampled around the side chain dihedral angles ($\chi_1$,$\chi_2$,$\chi_3$, see figure 6.2 for a schematic diagram) by moving the model of the side chain within the electron density map around the side chain dihedral axis. At each sampled point, the Ringer routine interpolates the value of the electron density, producing a two dimensional plot of dihedral angle versus electron density, as in figure 6.6.

Alternate conformers detected by the Ringer routine can be below typical electron density noise thresholds (1$\sigma$ in relative contoured maps), but have been shown to represent signal rather than Fourier errors or model bias [245].

For non-branchingside chains each peak within such a plot is evidence for a rotamer conformation. Note that isoleucine (ILE), valine (VAL) and threonine (THR) residues are

branched at the $\beta$ position, therefore these residues show two intense $\chi_1$ Ringer peaks.

For simplicity this work focuses on sampling around $\chi_1$ dihedral axis, as this should detect the largest changes between electron density support for a conformation. Glycine and alanine residues were not included in the analysis as they have no heavy chain atoms beyond the $\gamma$ position. If the analysis were undertaken at higher dihedral angles, a further reduced selection of side chain residues could be studied, as some side chains are too short to have atoms beyond these higher dihedral angles.

When side chain atoms are freely rotated about the dihedral axis, atoms at the $\gamma$ position sterically clash with the backbone, restricting the side chain to three conformations separated by 120 ° [309], which are known as gauche (+), gauche (-) and trans rotamers.

Prior to determining comparison metrics, the ringer plots for each dataset are interpolated onto the same angle scale so that they can be compared.

### 6.2.4 Pairwise analysis of datasets

Two different pairwise comparisons are considered, Pearson correlation coefficient and a fit to a model comprised of three Gaussian curves representing idealised rotamers.

The Pearson correlation coefficient between Ringer plots was employed as it was a simple method of pairwise comparison that could represent the similarity between two datasets. However, as it is a linear measure of correlation, it was limited in comparing the non-linear Ringer plots.

Therefore, the expected electron density distribution for the rotamers was approximated by three Gaussians. The pairwise Euclidean distance between the amplitudes of the Gaussians allows comparison of the relative occupancy of the different conformations of that residue.

The idealised model, in which dihedral angle $\chi_1$ is restricted to three possible conformations separated by 120° is considered as a basis for a curve fitting routine. Therefore, it was expected that the Ringer routine will find at most three peaks, these were fitted with three Gaussian curves, of the form:

$$f(x) = A_1 \exp(-\frac{x - x_{01}}{\sigma_1}) + A_2 \exp(-\frac{x - x_{02}}{\sigma_2}) + A_3 \exp(-\frac{x - x_{03}}{\sigma_3}) + B \quad (6.1)$$

The curve fitting routine uses a non-linear least squares routine using the Levenberg-Marquardt algorithm with 20° bounds on the mean value of the peaks, figure 6.7 shows a sample of these fits. For ease of visualisation the map values are interpolated such that most likely peak locations are at 60°, 180° and 300°, see figure 6.5. The other fit parameters; amplitudes of each Gaussian, $\sigma$ for each Gaussian, and a linear offset were allowed to vary to any value. As the width of the Gaussian can increase arbitrarily (in the case of fitting a single peak for example), the $\sigma$ values for each Gaussian were not further analysed. Section 6.4.1 discusses future work to compare the $\sigma$ values of the fitted Gaussians.

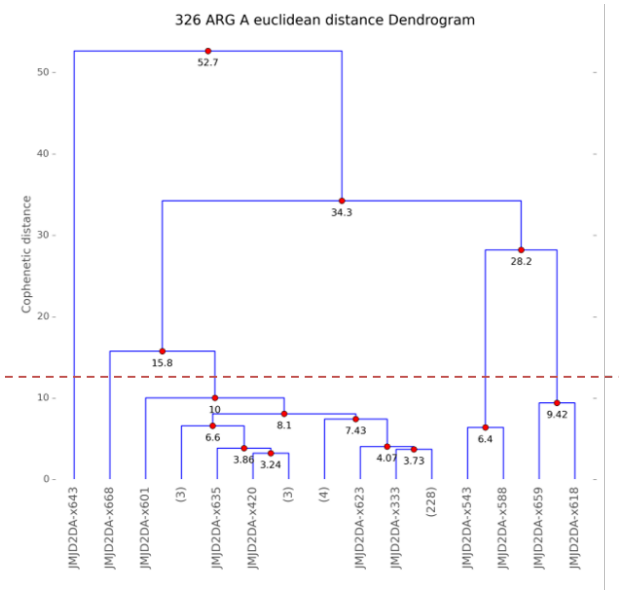
The linear offset used in the Gaussian model can negate the need for scaling the structure factor, and thus the electron density, as the change in scale be accounted for in the offset. The scaling however was necessary for the comparison using the Pearson correlation coefficient, and thus we use the scaled data in both methods.

The amplitudes of the fitted Gaussians should represent the relative occupancy of alternate conformers, and therefore are compared. The Euclidean distance between the amplitudes of two datasets was used as the pairwise comparison between datasets. Variation in the mean of the Gaussian curve, measuring the movement away from the ideal 120° separation can also be analysed.

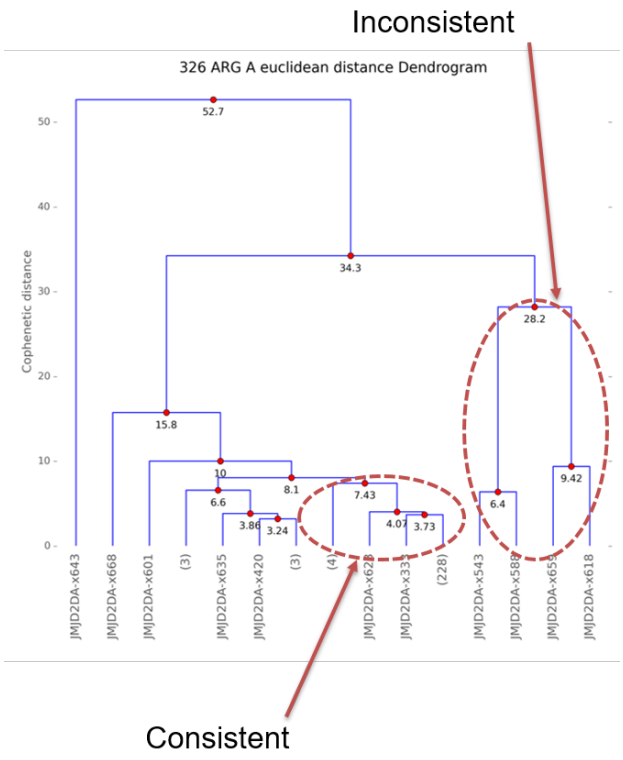
### 6.2.5 Clustering Ringer plots over datasets for each residue

Clustering was used to select sub-groups of datasets which were similar for every residue. Hierarchical agglomerative clustering was used to group the pairwise comparisons. Each dataset starts as a separate cluster, and the two that are most similar, i.e. closest values of the pairwise metric are linked. This process was repeated, with the link being made between either singleton datasets or clusters using the nearest point algorithm (within the `scipy.cluster.hierarchy.linkage` function). The output can be represented using a dendrogram, as in figure 6.3.

The output of hierarchical agglomerative clustering has every dataset linked, with a cophenetic distance (vertical height in figure 6.3) representing how similar the datasets were. Next the data needs to be split into clusters, ideally so that the most similar




**(a)**




**(b)**

**Figure 6.3:** Clustering ringer plots into groups of similar datasets for each residue. Hierarchical clustering joins similar datasets, due to variation in cophenetic distances of differing residues we use the inconsistency threshold to clusters. This can detect abrupt changes, which implies a group of datasets is dissimilar from another. However, it requires two free parameters depth and threshold. Dendrograms represent the output of hierarchical agglomerative clustering.

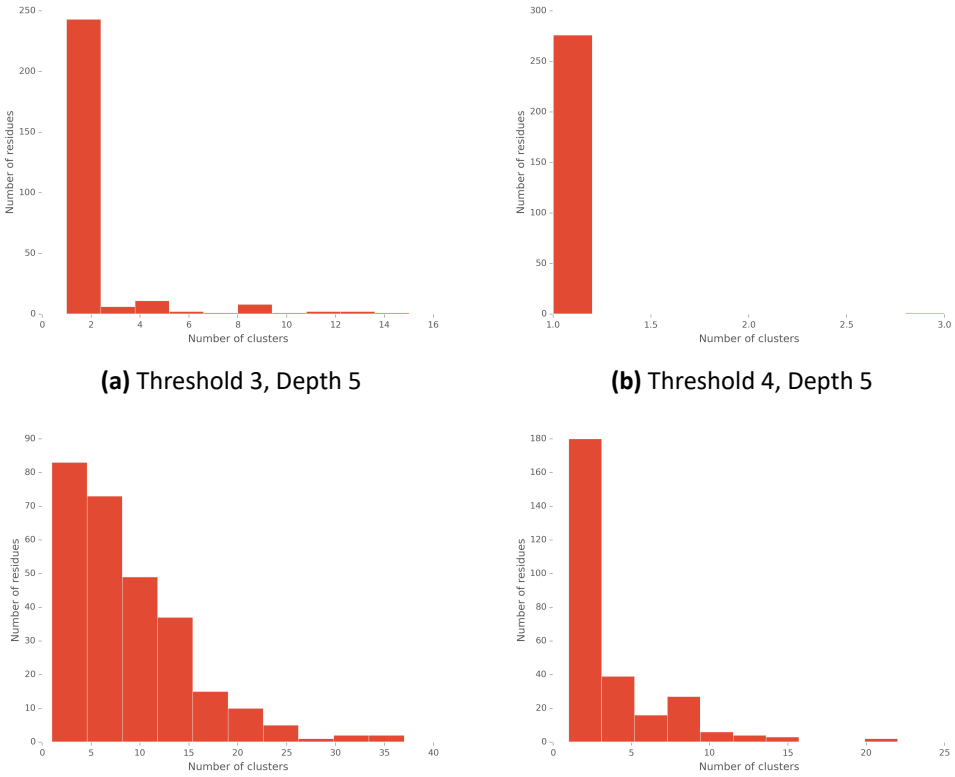
datasets are grouped together, and any natural divisions between the datasets are followed. One of simplest ways to split the dendrogram is to pick a value of cophenetic distance, and truncate the clustering at that value. For example, in figure 6.3(a), the line shown would divide the data into five clusters. However as there was a different clustering of datasets for each residue, the appropriate cut off for cophenetic distance varied greatly between datasets.

Therefore, the inconsistency coefficient was used to choose clusters. This measures abrupt changes in the cophenetic distance, so that only when there were relatively large changes between clusters, a division was made. The cophenetic distance of a link between clusters was compared to that of the average of those below it. This averaging was done over a specified number of cluster links, given by the depth parameter. The ratio between the link, and the average of those below was compared to an inconsistency threshold, which determined how the data was clustered. It was hypothesised that datasets which had evidence for an additional alternative conformation would cluster together.

In typical crystallographic fragment screens 75-95% of datasets do not have bound ligands [104]. It was hypothesised that many of the datasets will be broadly similar, such that the distribution of the clusters would be well represented by a single large cluster and a few smaller clusters. Figure 6.4 shows the depth and threshold free parameters which need to be set to cluster using the inconsistency threshold. These parameters determine number of clusters and the distribution of cluster sizes. In this analysis the cluster distribution in figure 6.4(c) was used, as this most consistently gave a large cluster, with a few small clusters, which we hypothesised would be able to select out sub populations of datasets such as those with bound ligands.

### 6.2.6 Scoring to detect sub-populations of similar datasets

Bound ligands are expected to alter the protein conformational state. Different ligands may cause the same conformational changes in the protein. For example, to bind the active site, all ligands might have to displace a gate-keeping residue, which blocks the site from being fully solvent accessible. In that case all bound ligands would have a significantly different Ringer plot for the gatekeeper residue.




**(a)** Threshold 3, Depth 5

**(b)** Threshold 4, Depth 5

**(c)** Threshold 3, Depth 10

**(d)** Threshold 4, Depth 10

**Figure 6.4:** Free parameters in clustering that were chosen to give a few residues per residue, however, these were not sufficient to discriminate sub-populations of datasets. Distribution of clusters for different values of inconsistency threshold and depth. **(a)** Selection criteria that lead to a few small clusters, and one large cluster. Having so few clusters leads to a lack of separation of the scoring metric used. **(b)** Selection criteria that led to a single cluster for almost every residue. **(c)** Shows the selection criteria used throughout further analysis. **(d)** Shows selection criteria that provided a lower variation in dataset score than **(c)**. This did not lead to discrimination of bound ligands.

A score from the clustering of the datasets, aimed to identify datasets that are consistently different across the residues analysed:

$$\text{Dataset Score} = \sum_{\text{All Residues}} \frac{1}{\text{Number of Clusters} \times \text{Size of Cluster}} \tag{6.2}$$

Datasets that consistently appeared in small clusters, for residues with a small number of clusters were hypothesised to be outlying populations which would score highly. Datasets which are consistently not in the main cluster across residues were hypothesised to have a perturbation, such as a binding ligand, or alternate protein conformation.

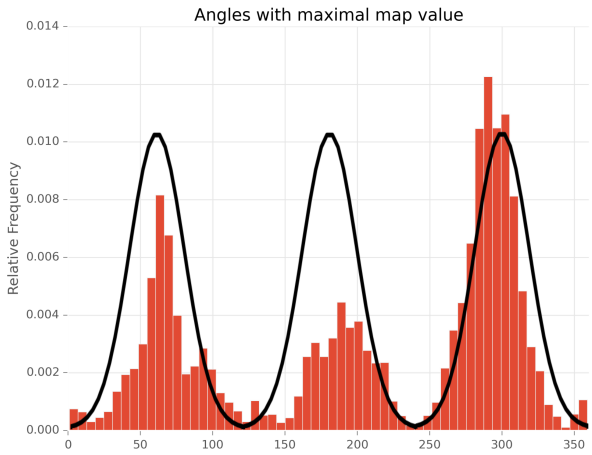



**Figure 6.5:** Dominant peaks found by Ringer matched the theoretical distribution of side chain rotamers, with the most common rotamer locations being separated by 120°. Black line represents the theoretical distribution, not adjusted for the residues seen in JMJD2D. Histogram of angular location of the dominant Ringer peak: All residues of JMJD2D.

## 6.3 Results

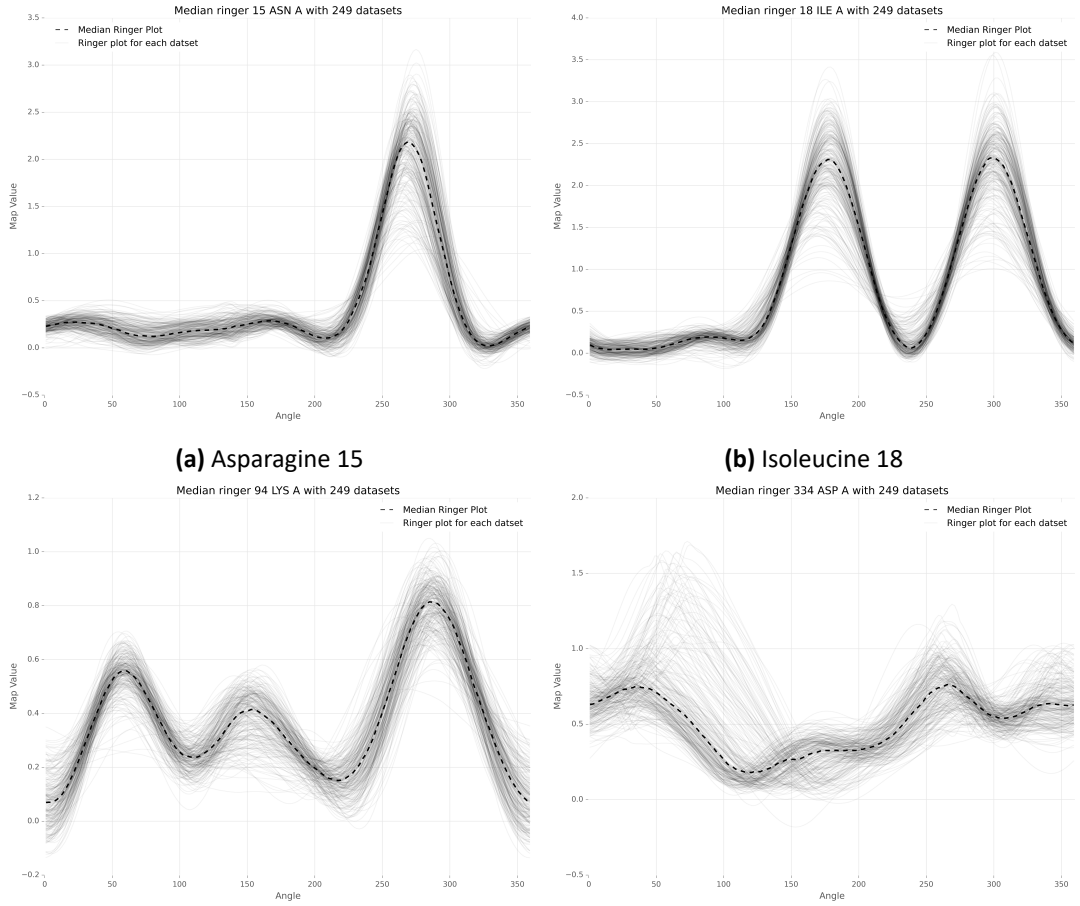
### 6.3.1 Idealised rotamers were detected in multiple datasets

Over multiple datasets the main Ringer peak follows the expected distribution of rotamers, as shown in figure 6.5. Note that the theoretical distribution shown in 6.5 is not adjusted for the residue distribution in the residues of JMJD2D. This showed that we have sufficient structural alignment between the datasets and that the interpolation to make the rotamer angles consistent between datasets worked well.

### 6.3.2 Conformational heterogeneity between datasets varied by residue

Ringer applied across multiple datasets allowed quick comparison of the number of different conformations sampled by each residue. Figure 6.6 shows Ringer plots for residues of the JMJ2D datasets. The asparagine 15 residue (figure 6.6(a)) showed a single peak without much variance, therefore the residue had a single conformation present in all datasets. Low conformational sampling also implies that this residue position was likely relatively inflexible, and unlikely to change given ligand perturbation.

The isoleucine 18 residue (figure 6.6(b)) had a small variation in peak height, this implies in some datasets there was weaker electron density evidence for the residue, or




**(a)** Asparagine 15 **(b)** Isoleucine 18

**(c)** Lysine 94 **(d)** Aspartic acid 334

**Figure 6.6:** Different residues had different amounts of conformational heterogeneity. Plots from the Ringer routine, showing rotamer occupancy of protein side chains identified by sampling electron density around $\chi_1$ dihedral angle, for 249 datasets of the lysine demethylase fragment screen. The median of all datasets is shown by a dashed line, highlighting the differing degrees of variability between datasets across residues, with **(d)** Aspartate 334 showing the highest variability of those shown.

that the scaling between datasets was not perfect. Also note there were two dominant peaks indicative of a branched chain at the $\beta$ position.

The lysine 94 residue (figure 6.6(c)) showed three peaks, as lysine is not a branched chain, this represents sampling of multiple conformations. There was some variability in this residue suggesting different occupation of each rotamer.

The aspartic acid 334 residue (figure 6.6(d)) showed a large variance in the amplitude and position of its major peak, suggesting a large amount of structural heterogeneity, and thus flexibility. This was likely as it was near the end of the protein chain.

The benefit of comparing Ringer plots for many datasets is that we have a quick visual

analysis tool for the available conformations, in which outliers and overall conformational heterogeneity can be quickly assessed.

### 6.3.3 Gaussian fits approximate the Ringer plot

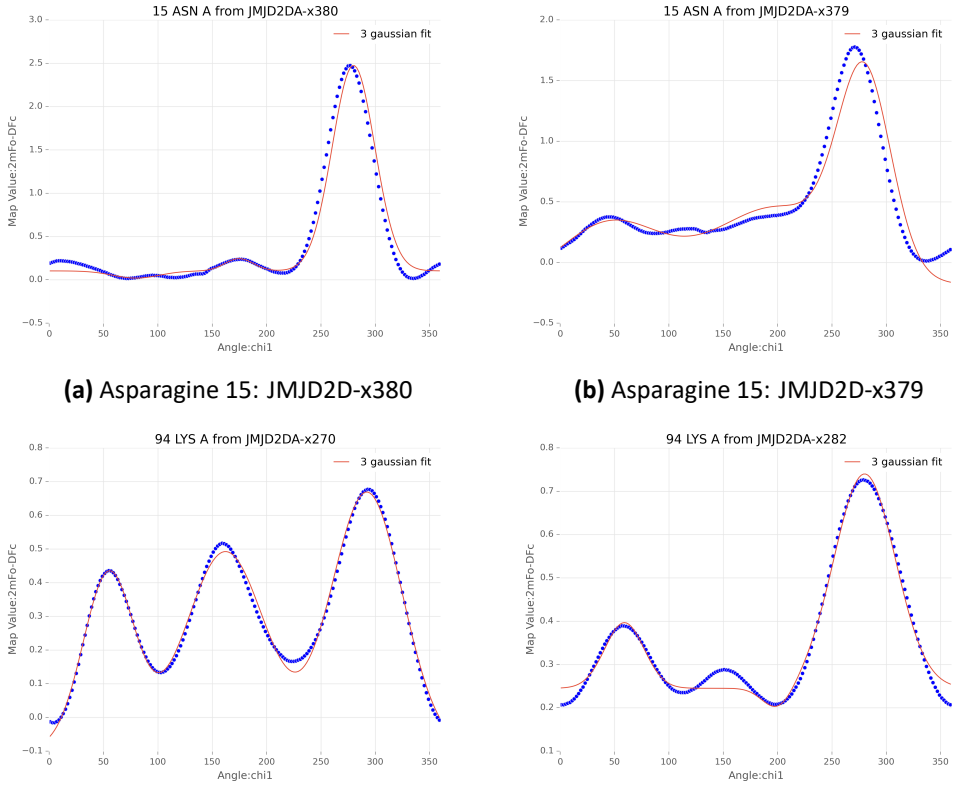
Gaussian fitting to the Ringer plots captures most peak features. However, some fits did miss the peak maxima, as in figure 6.7(b). The ability to detect a subtle change in the Ringer plot once fit with Gaussian curves may be sensitive to such relatively small errors, as some of the differences between Ringer plots can be relatively small. Comparing figures 6.7(c) and 6.7(d) it was identified that the fits can capture the significant differences between the Ringer plots. Other fitting errors can be seen such as in figure 6.7(d) does not well fit the smallest peak.

### 6.3.4 Ringer across multiple datasets detected correlated motion in *T. cruzi* FPPS case study

Fragment screening undertaken at the XChem facility identified two compounds (**7**) and (**42**), as shown in figure 6.8 , which bind to *T. cruzi* FPPS which induced the opening of an allosteric pocket involving the rotation of phenylalanine 50. This movement shown in figure 6.8(c) can be identified from sampling electron density without modelling the explicit conformers. This was shown in the Ringer plot in figure 6.8(d). Both datasets have a reduced main peak, and a secondary rotamer peak, not present in many datasets separated by 108 ° and 125 ° when binding with compounds (**42**) and (**7**) respectively.

Datasets which differ from the median Ringer plot at aspartic acid 98, figure 6.8(f). This residue was involved in the first DDXXD motif which coordinates the metal ions in the active site. The TCRUFPPS-x0051 structure was dominated by a single rotamer, whereas TCRUFPPS-x0106 had its density spread over two commonly occupied rotamers. This site is 18 Å away from the phenylalanine 50 ligand binding, as shown in figure 6.8(e). As theses changes in phenylalanine 50 and aspartic acid 98 occur in the same datasets, this is correlated motion.

The movement of the phenyl residue near the ligand binding, and the distal movement of the aspartic acid coordinating the metal ions was not detected by the clustering of




**(a)** Asparagine 15: JMJD2D-x380

**(b)** Asparagine 15: JMJD2D-x379

**(c)** Lysine 94: JMJD2D-x270

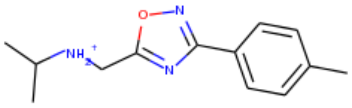
**(d)** Lysine 94: JMJD2D-x282

**Figure 6.7:** Gaussian fitting captured most peak features, but was not ideal, missing peak maxima. Three Gaussian fitting routine applied to four Ringer plots. **(a)** and **(b)** show two different datasets for asparagine 15, a major peak showing the dominance of a single rotamer. Note in (b) the peak was shifted away from 300°, by more than the 20°bound, forcing a poorer fit. **(c)** and **(d)** show two different datasets for lysine 94, identifying rotamer occupancy at three positions in both datasets presented, showing a large difference in the relative occupancy of the rotamers between these datasets.

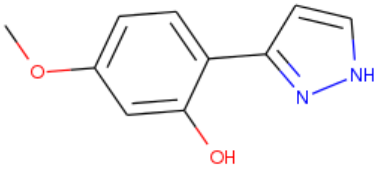
Gaussian fits to the ringer plots. However, as these residues are outliers when compared to other rotamers, there is scope to identify appropriate methods to analyse these relationships, as discussed in section 6.4.1.

### 6.3.5 Sub populations of datasets such as bound ligands were not detected by automated method

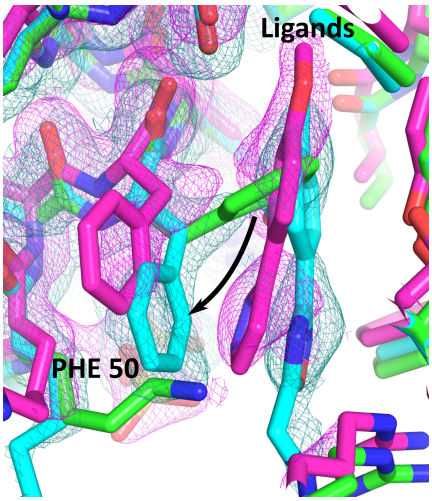
The correlation and clustering methods were applied to the JMJ2D datasets and then scored based on size of cluster each residue belonged to, did not discriminate bound ligands from unbound ligands, as can be seen in figure 6.9. This could be due to the score being a poor way to summarise cluster information. The score involves summation



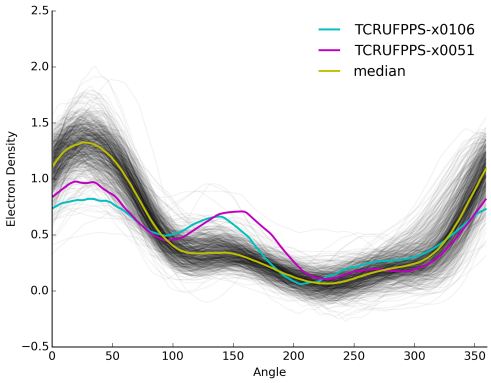

**(a)** FMOPL000293a (**42**)



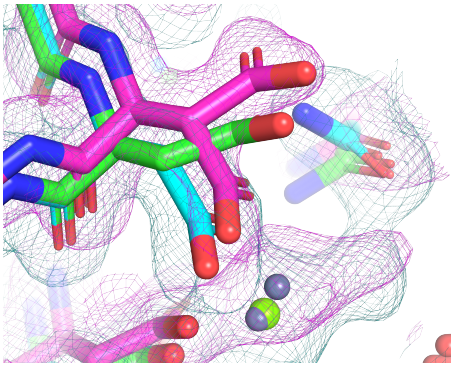

**(b)** FMOPL000554a (**7**)



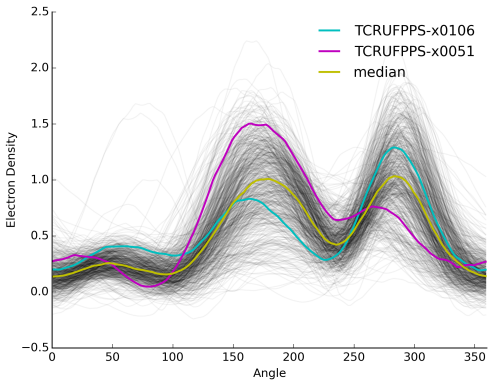

**(c)** Ligand binding near phenylalanine 50




**(d)** Ringer plot for phenylalanine 50



**(e)** Motion in aspartic acid 98




**(f)** Ringer plot for aspartic acid 98

**Figure 6.8:** Correlated distal motion between Phe 50 and Asp 98 on ligand binding. Figure 6.8(c) shows the motion of phenylalanine 50 on binding of two ligands, with these alternate conformations identifiable in the Ringer plot Figure 6.8(d). 18 Å distal to the ligand binding there is a associated movement in a aspartic acid residue which coordinates the metal ion in the active site Figure 6.8(e). This shows as a selection of two possible conformers in Figure 6.8(f).

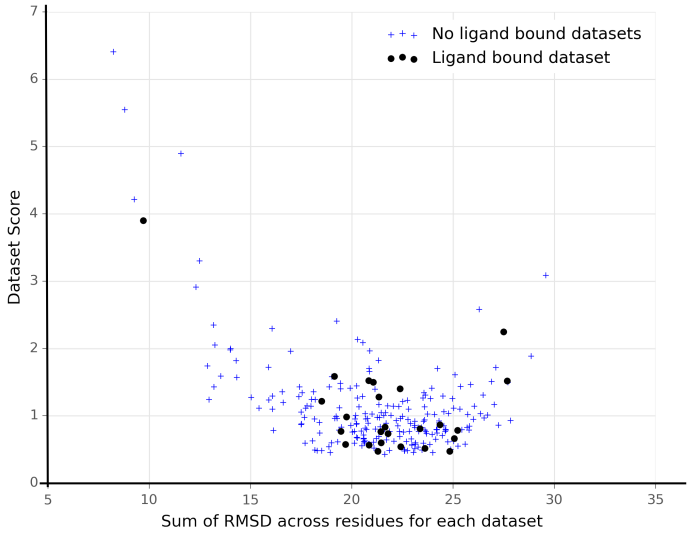



**Figure 6.9:** Scoring datasets via the clustering of Ringer plots does not discriminate bound ligands. Dataset score for all 249 datasets from the JMJ2D fragment screen versus the quality of the three Gaussian fit. Fit quality was measured by the root mean squared deviation from the fit to the data, summed over all residues analysed by the Ringer routine at $\chi_1$. Datasets with bound ligands are not selected out by the dataset score.

over all residues for a dataset, but the differences between datasets with bound ligands and without may only be present in a few residues. The assumptions about cluster size and consistently belonging to small clusters could also be incorrect, or the clustering method or parameters were chosen poorly.

## 6.4 Discussion and future work

The aim was to detect correlated motions between residues in fragment screening datasets. To achieve this the electron density near each residue of each dataset was sampled with Ringer, a tool which identifies alternate side-chain rotamer conformations. Clustering pairwise comparisons of Ringer plots between datasets for each residue was expected to detect groups of residues which belonged to sub-population of datasets. Scoring of these clusters was explored to detect correlated motion between datasets, however the method was unable to automatically identify any significant realtionships between residues.

In the case study of *T.cruzi* FPPS, Ringer over multiple datasets can detect protein motion associated with distal ligand binding. In this case study the ligand binding was associated with the movement of a phenylalanine residue, which could be seen in the ringer plot as outliers in this residue. The multiple dataset ringer showed that there were a few other datasets which also sampled the alternative conformation of the of the phenylalanine residue. The outlying datasets also had large difference from the median in aspartic acid 98 which is in the active site, which coordinates a metal ion.

Multiple dataset Ringer is a good tool for visualising conformation heterogeneity and thus flexibility. It can be used to assess the flexibility of residues in the binding pocket, and identify if there are existing datasets which explore similar conformations to those seen when ligands bind.

### 6.4.1 Weaknesses in the multiple dataset ringer routine and opportunities for improvement

The clustering techniques applied to multiple Ringer plots have been unsuccessful at automatically identifying correlated motion, or identifying bound ligands. However there are consistently new conformations identifiable across multiple datasets. Ringer plots of many datasets can be a quick way to identify residues with high conformational heterogeneity. Further work would involve testing alternate ways to determine which datasets have outliers and thus alternative conformations. Furthermore, work on how to detect correlations between residues with outlying conformations is needed.

The three-dimensional real space of an electron density map is reduced to a projection along the dihedral axis that the protein side chains sample. This means that we can more meaningfully compare the distribution of electron density, as there is a model with which to interpret it. Gaussian curves were fit to the distribution and compare those, as we expect the distribution to be well explained by one to three Gaussians. However, one could instead utilise the Kullback-Leibler divergence, a measure of how one probability distribution is different from another. One would look for outliers in the KL divergence across datasets and residues. Outliers would suggest rotamers which

are rarely seen. Alternatively, one could use the Earthmover's distance to measure the difference between the distributions, looking for outliers.

Another potential method would be to use Fourier analysis to identify the major peaks in the rotamer distributions. The peaks identified could then be clustered. This has the benefit that the Ringer routine samples 360°, and thus are periodic and thus could be well suited to Fourier analysis. The current analysis which fits three Gaussians has discontinuities at 0/360 ° which is problematic as these are the same point in space.

Analysing the width ($\sigma$) and mean of the Gaussians rather than just their amplitude would be beneficial, especially for cases such as the *T. cruzi* in figure 6.8, where the main difference is the mean angle of the rotamer in the outlier. In this implementation this was not done, as the width of the Gaussian could vary to allow a single peak to dominate. An alternative method could detect the number of peaks before fitting the Gaussians. This was unsuccessfully attempted in this project, likely unsuccessful due to the choice of a poorly suited peak finding routine.

Furthermore, the clustering step should be analysed; firstly, the shape of the underlying rotamer data should be analysed using dimensionality reduction techniques such as principal component analysis. This can help determine the correct number of clusters, and help determine which clustering method may be more effective.

Sampling common rotamers positions rather than sampling around the dihedral angle could allow easier identification of outliers, and sampling of rotamers which are relevant for the residue in question. However, it may reduce the likelihood that a weakly occupied rotamer is detected, as not all angles would be sampled.

As the Ringer routine fixes the position of the backbone when sampling the electron density, this limits the potential detection of conformations that are associated with backbone movements. A routine that also samples variation in the backbone atoms would allow for larger conformational changes to be detected, however constraints on possible backbone torsion angles ($\psi, \phi$) would be needed to reduce the search space of potential conformations. qFit [244, 310], is a tool that samples many side chain conformers and backbone fragments to generate a single multi-conformer model that explains local

features in electron density maps. Multi-conformer models generated by qFit could be compared between different datasets, to show the differences in available conformers.

### 6.4.2 Future work

New strategies are being developed in protein crystallography group at the Structural Genomic Consortium (private correspondence Conor Wild) for clustering outlying datasets which rely on electron density, and not the structural model. Separating out conformational states prior to refinement, would be beneficial as detecting after refinement requires manual model building to identify regions of change. This work clusters the electron density of datasets, using principal component analysis, t-distributed stochastic neighbour embedding and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to determine groups of datasets which are most similar. These groups of datasets have shown to identify the same clusters of datasets as in the Kalirin-Rac1 complex discussed in chapter 2, as well as highlighting residue movement in a recent fragment screen of MurE Ligase. Currently this allows crystallographers to separate datasets into alternate PanDDA runs to better characterise ground states. Combining this analysis with multiple dataset Ringer, could allow for hypothesis generation on which local perturbations can be associated to the outlying sub-population of datasets. Thereby detect what structural difference there are between sub-populations of crystals.

# 7

# Conclusions and future work

## Contents


**7.1 Superposed modelling improves the description of the crystal state 196**
7.1.1 Superposed modelling can better describe crystal state . . . . 196
7.1.2 Metal ions require independent restraints in superposed models 197
7.1.3 Future work: model comparison metrics are needed to improve models of weakly binding ligands . . . . . 197
**7.2 Refinement across repeated datasets of protein-ligand complexes and many weakly binding fragment ligands show occupancy refinement is inconsistent . . . . . 198**
7.2.1 Future work: developing a benchmarking set for refinement of weakly binding ligands . . . . . 198
**7.3 Intact mass spectrometry of crystals calibrated occupancy refinement methods . . . . . 199**
7.3.1 Future work: detailed measurement of crystal parameters may be required to ensure that maximal ligand occupancy is achieved prior to cryo-cooling . . . . . 200
**7.4 Correlated motions were detected by sampling electron density across multiple datasets . . . . . 200**
7.4.1 Future work: relating sub-populations detected globally over multiple datasets to associated local residue movements . . . 200
**7.5 Summary . . . . . 201**


This thesis explored how to utilise electron density across multiple datasets to improve models of protein-ligand complexes and alternate protein conformations. As well as detect correlated motions of protein residues. As shown in chapter 3 models

parameterising ligand occupancy had a large variance in occupancy values for repeat datasets of protein ligand-complexes.

Collection of calibration data using intact MS of crystals which contained a proportion of covalent compound, showed that occupancy correlated with the proportion of compound. By fixing the B factor a better fit between occupancy and proportion of covalent compound was achieved. This suggests that restraints in the continuity of B factor between the ligand and protein would improve occupancy refinement.

In chapter 6, sampling electron density across multiple datasets allowed the detection of correlated motions of protein residues, associated with the binding of a ligand at an allosteric site in *T.cruzi* FPPS.

## 7.1 Superposed modelling improves the description of the crystal state

### 7.1.1 Superposed modelling can better describe crystal state

Modelling a protein as a superposition of multiple conformations can better explain the state of the crystal. The Kalirin/Rac 1 complex case study, discussed in chapter 2, showed that the superposed model was the most appropriate model for the loop rearrangement in some datasets.

A rearranged position of the variable switch 1 loop was identified to have an alternate conformation across multiple residues. This conformation was first found alongside the binding of covalent compounds, but was confirmed to occur without a covalent binder. The region is typically mobile in GEF/GTPase complexes, but understanding where the loop is positioned is key to developing chemical probes against the target. Refinement was carried out with models which contained: the original loop position, the alternate loop position, or a superposition of both models. This was followed by the calculation of the RSCC for loop residues in each of the models. In the datasets studied which are suspected to have loop rearrangement, around 20% had a high RSCC for the original loop position. In approximately 50% of cases the loop rearrangement had a better fit to density. In the remaining 30% of the crystals; some were well explained by a superposition of

both loop rearrangements, with both loop models having high RSCC and reasonable B factors. Therefore, in some cases a superposed model was necessary to describe the state of the crystal. However, some datasets had high $R_{\mathrm{work}}$ and $R_{\mathrm{free}}$ values suggesting a superposed model was still inappropriate.

### 7.1.2 Metal ions require independent restraints in superposed models

In the fragment screening data of *T. cruzi* FPPS, both magnesium and zinc ions were considered as potentially bound to the active site. Electron density evidence showed that partially occupied zinc ions should be modelled. To ensure correct coordination of the metal ion, restraints were needed. In the model used the metal ions' occupancies varied between 0 and 1 and were not related to the occupancy of the nearby ligand. This produced the most stable refinement, and reasonable model for the ligand binding site with low level of difference electron density over the ions.

Automated methods to appropriately determine restraints for metal co-ordination, such as `phenix.metal_coordination` [217] need to be tested and automated with the restraint generation for superposed models. This would likely involve updating `giant.make_restraints` to allow crystallographers to routinely refine metal ion containing structures with a superposed model.

### 7.1.3 Future work: model comparison metrics are needed to improve models of weakly binding ligands

The lack of model comparison metrics in crystallography, such as information criterion limits the ability to directly compare the applicability of two models. For example, when is adding a superposed model a more appropriate way of explaining the electron density, given the increase in free parameters. Application of such metrics on synthetic data [286] has shown the ability to distinguish between residues with one and two conformations more reliably than $R_{free}$. As crystallographic data is in reciprocal space, any model change would need an independent assessment, therefore each additional conformer added should be refined and assessed under an information criterion metric, rather than a bulk of changes.

Introducing information criterion metrics into crystallographic refinement would make choosing refinement strategies, such as superposed modelling an easier and more well evidenced choice.

## 7.2 Refinement across repeated datasets of protein-ligand complexes and many weakly binding fragment ligands show occupancy refinement is inconsistent

Application of `phenix.refine`, `REFMAC5` and Buster refine to repeated datasets of NUDIX hydrolase protein ligand complexes showed that the simultaneous refinement of B factor and occupancy was inconsistent, giving a wide distribution of ligand occupancy values. As B factor and occupancy are correlated, changes in one parameter can compensate for the other. Therefore, a systematic search of B factor and occupancy was developed. This method considers the ligand, and the surrounding protein residues which change upon ligand binding (referred to as exhaustive search, chapter 3). This showed that by varying B factor and occupancy, a smoothly varying function of difference density could be obtained. This implied that refinement methods were likely using appropriate algorithms, but the correlation of B factor and occupancy requires one of these parameter to be quite tightly restrained to make the other more interpretable.

The refinement methods were also applied across approximately 2000 protein ligand complexes of fragment sized weakly binding ligands from XChem targets, chapter 4. This showed similar inconsistencies in occupancy refinement between different programs as was identified in the repeated datasets of protein-ligand complexes.

Therefore it is not currently possible to use occupancy to rank confidence in ligand binding. Furthermore, the occupancy cannot be linked to the relative binding affinity of the ligands.

### 7.2.1 Future work: developing a benchmarking set for refinement of weakly binding ligands

In this work I developed a generalised task based re-refinement of many XChem fragment screening hits in different refinement programs. Extending this slightly to improve

reproducibility could allow it to be used as a benchmarking dataset for comparing crystallographic refinement of weakly binding ligands in different software.

Instead of the current file based management of tasks, a Django database model could be deployed. This would allow centralised records of different refinement program results. These alternate refinements could then be presented to the user as alternative interpretations of the data. The results could also be stored for benchmarking of the refinement programs. Given that many XChem targets are in the public domain, such as those from the SGC, these refinements could be studied to improve the refinement of weakly binding ligands.

## 7.3 Intact mass spectrometry of crystals calibrated occupancy refinement methods

Intact MS of crystals to detect the proportion of a covalently bound ligand enabled the calibration of refinement of crystallographic occupancy.

Two species of NUDT7 protein were produced, one labelled with a covalent compound, the other without. These were carefully combined and crystallised to produce crystals with different proportions of covalent compound. Some of the crystals at each ratio were processed by intact MS, some of the crystals were sent to the synchrotron for diffraction studies. The ratios measured by intact MS on the crystal were used to calibrate the different crystallographic occupancy refinement methods.

When occupancy and B factor were refined simultaneously, the refinement programs did not produce consistent results as was expected from the results obtained in chapters 3 and 4. In section 5.3.1 it was shown that by fixing the ligand B factors, the fit to the expected occupancy from the intact MS calibration improves.

The crystal system used in the intact MS experiments was problematic due to the presence of polyethylene glycol, so alternate crystal systems without this could be studied to further explore the relationship. Micro-spectrophotometry techniques to look at functional behaviour in the crystalline state [311], such as Raman difference method [306] or UV-Visible [308] could be adapted to look at occupancy in ligand binding examples.

### 7.3.1 Future work: detailed measurement of crystal parameters may be required to ensure that maximal ligand occupancy is achieved prior to cryo-cooling

Existing work showing that occupancy is dependent on soaking time and crystal size [256] employed a similar restriction of B factor to that employed in chapter 5 to calculate the occupancy of the ligand. In the study of crystal size versus occupancy the authors applied a method to decouple occupancy and B factor. The datasets were truncated to have a common resolution, whilst applying a reciprocal space temperature factor adjustment. Although such changes may be beneficial in these controlled cases, this work aimed to find a refinement approach that would be applied to many different target-ligand complexes and did not utilise these correction methods as they are difficult to apply to all crystal studies.

The effects of crystal size and soak time are not studied in detail in this work, as XChem fragment soaks are generally at high concentration of fragments and soaks are usually carried out over 1-3 hours suggesting that the compound would be at it's maximal occupancy [256]. The addition of cryoprotectants could also be looked at, to determine whether adding a protectant affected the detectable occupancy. As the majority of the crystal systems did not need additional cryoprotectant, this was not studied herein.

## 7.4 Correlated motions were detected by sampling electron density across multiple datasets

Chapter 6 focused on detecting sub-populations of datasets with correlated motions. In a case study of *T Cruzi* FPPS we showed that Ringer [245] plots of multiple datasets could be manually inspected to find correlated motion. However, initial attempts to automate the selection of residues with correlated motions were unsuccessful.

### 7.4.1 Future work: relating sub-populations detected globally over multiple datasets to associated local residue movements

A new technique of detecting sub-populations of datasets is under development in the protein crystallography group at the SGC (private correspondence Conor Wild). This

method clusters the electron density of datasets, using principal component analysis, t-distributed stochastic neighbour embedding and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [312] to determine groups of datasets which are most similar.

The method can identify the same clusters of datasets as in the Kalirin-Rac1 complex case study discussed in chapter 2, as well as highlighting residue movement in a recent fragment screen of MurE Ligase. The current role of this code allows crystallographers to separate datasets into alternate PanDDA runs to better characterise ground states.

However as this new method relies on dimensionality reduction of the electron density, sub-populations of datasets detected in this technique cannot be traced back to a local change in the protein.

By combining this technique with multiple dataset Ringer outlined in chapter 6 to detect recurring local side chain motions. It may be possible to localise the changes, and allow a crystallographer to quickly build alternative models.

## 7.5 Summary

This thesis outlines the current limitations in the refinement of crystallographic occupancy, and develops experimental and computational strategies to produce more informative models. The superposed model determined from multi-dataset analysis has been shown to be a useful tool in understanding protein behaviour, such as alternative loop conformations in crystals. It is expected that the ability to collect many crystal datasets and more easily process information from multiple datasets with tools such as superposed modelling will lead to new insights in protein flexibility and the behaviour of proteins under perturbation by ligands.

# Appendices

# A

# Code and data accessibility

- DCP2 fragment screening data: coordinates and structure factor are available in the protein data bank under group deposition ID: G_1002061. The complete PanDDA analysis, which was carried out in two stages due to updates to the platform, is available under DOI 10.5281/zenodo.1437589

- *T. cruzi* FPPS fragment screening data is available at DOI 10.5281/zenodo.2649077 and is in process of submission to the protein databank.

- The repeated protein datasets (chapter 3), and NUDT7 datasets (chapter 5) are not currently deposited and can be requested from the author. The datasets used from XChem (chapter 4) are not available, as some of the analysis is based on XChem user data. Future release of the open targets will occur with further development of the codebase.

- Code for analysis of multiple loop rearrangements discussed in chapter 2 is available at `https://github.com/nelse003/multiple_loop_refinement/releases/tag/v0.1`.

- Code for the exhaustive search algorithm (chapter 3) is available at `https://github.com/nelse003/exhaustive_search/releases/tag/v0.1`.

- Code for the analysis of XChem hits (chapter 4) is available at `https://github.com/nelse003/parse_xchemdb/releases/tag/v0.1`.
- Code for the analysis of MS data and associated crystals (chapter 5), alongside summary MS data is available at `https://github.com/nelse003/parse_xchemdb/releases/tag/v0.1`.
- Code for the analysis of multiple dataset Ringer (chapter 6) is available at `https://github.com/nelse003/Multiple_Dataset_Ringer/releases/tag/v0.1`.

# B

# Fragment screening data

## B.1 DCP2 (NUDT20)

### B.1.1 Biological relevance of DCP2

Regulation of mRNA decay, is a highly regulated process, that allows for rapid control of mRNA levels in the response to cellular events. Eukaryotic mRNAs are protected from degradation, by both the 5' 7-methylguanosine cap and the 3' poly-A tail. For decay to occur, either one of these structures must be compromised, alternatively the mRNA could be cleaved internally by endonucleolytic attack.

The first step of degradation in eukaryotes is usually initiated by the shortening of the poly-A tail. This is followed by either decapping, the removal of 5' cap, allowing the mRNA to be degraded in the 5' $\rightarrow$ 3' direction by the XRN1 exoribonuclease, or the 3' end can be attacked by the exosome, a large complex of 3' $\rightarrow$ 5' exonucleases.

The decapping process is undertaken by the an enzyme consisting of DCP1 and DCP2 domains. This complex is weakly active on its own, but requires accessory factors for full activity [313]. These include the Pat1-Lsm1-7 complex, and the enhancers of DCP2 activity (Edc1-Edc2 and Edc3). Different structures of DCP2 containing complexes have been solved: figure B.1 shows the structure of the *Kluyveromyces lactis* DCP1-DCP2-EdC3 complex. The catalytic NUDIX domain belongs to the superfamily of hydrolytic enzymes capable of cleaving nucleoside diphosphates lined to x (which represents any moiety)

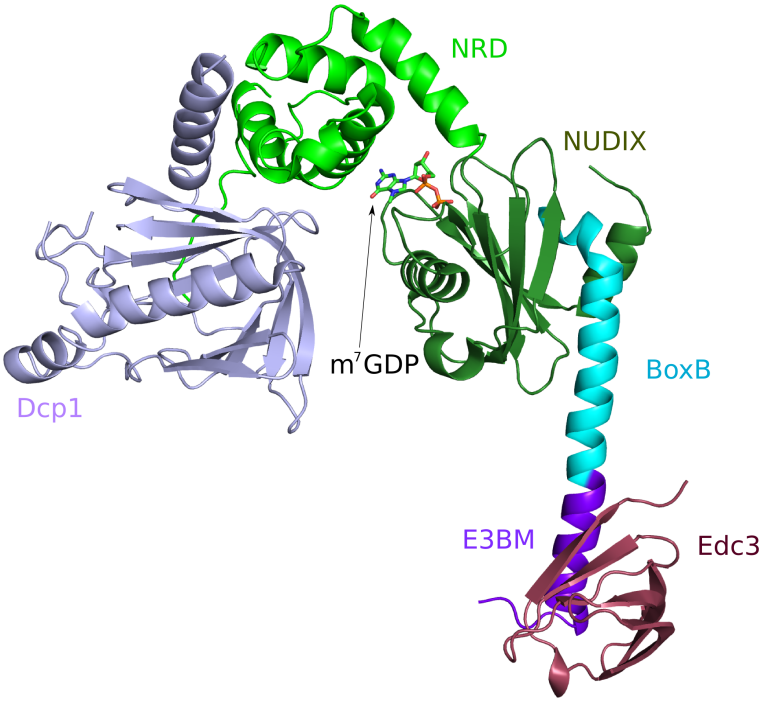



**Figure B.1:** Structure of the *K.lactis* Dcp1-Dcp2-Edc3-m$^7$GDP complex. PDB code: 5LOP [315].

[314]. This superfamily has been studied by the SGC, attempting to gain a structural understanding of multiple family members, many of which are represented throughout the crystallographic studies in this thesis. The NUDIX motif is GXXXXXEXXXXXXXREUX-EEXGU, where U is isoleucine, leucine or valine, and X is any amino acid. This forms a short helix, which usually contains the catalytic amino acids of the enzyme.

Human m7GpppN-mRNA Hydrolase (DCP2, NUDT20) catalytic domain, figure B.2,was crystallised at the SGC. To understand potential inhibitory routes, a crystallographic fragment screening campaign was undertaken.

### B.1.1.1 Protein preparation

Protein expression, and purification were carried out by Micheal Fairhead (SGC), following the PREPX method [manuscript in preparation].

Human DCP2 (residues 95-260) was cloned into pNIC28-Bsa4 with a TEV-cleavable N-terminal hexahistidine tag. After transformation into *E. coli* (BL21(DE3)-R3), expression

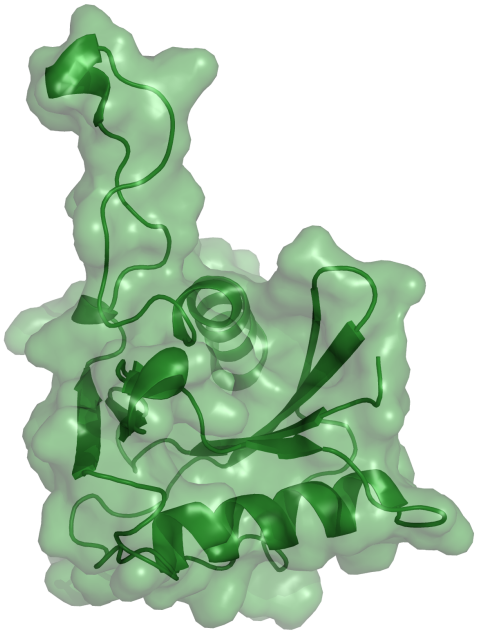


**Figure B.2:** Structure of the Human m7GpppN-mRNA Hydrolase (DCP2, NUDT20) catalytic domain. PDB code: 5MP0

was performed in TB auto induction media (FroMedium), supplemented with Kanamycin at 50 µg ml$^{-1}$ and Chloramphenicol at 34 µg ml$^{-1}$, 1 ml l$^{-1}$ 10% antifoam, 1 ml l$^{-1}$ *E.coli* metals, as defined in B.5.1, 10 ml l$^{-1}$ 1 M ammonium sulfate, 1 ml l$^{-1}$ 1 M magnesium sulfate and 20 ml l$^{-1}$ 25% glucose. Cultures were grown for 4 hours at 37 °C shaken at 300rpm, the temperature was decreased to 17 °C and the cultures were grown for another 44 hours. Cells were spun at 4,000g for 20 mintues, 4 °C, and stored at −80 °C. After thawing cells were re-suspended in base buffer (10 mM HEPES, 500 mM NaCl, 5% Glycerol, 0.5 mM TCEP, pH 7.5) with 0.5 mg µl$^{-1}$ Lysozyme and 1 µg µl$^{-1}$ Benzonase, vortexed and left at room temperature for 30 minutes. 1% Triton X-100 was added, and the cells frozen at −80 °C for a minimum of 1-2 hours. On thawing cells were centrifuged at 4,000g for and the supernatant applied to a His GraviTrap column (GE healthcare) equilibrated with base buffer. After washing with base buffer supplemented with 20 mM imidazole, DCP2 was eluted with buffer supplemented with 500 mM imidazole. The eluted protein

was applied to PD-10 desalting column (GE healthcare) and eluted with base buffer supplemented with 20 mM imidazole. The N-terminal affinity tag was removed by TEV cleavage overnight, and uncleaved protein was removed by applying it again to a His GraviTrap column. The flow-through was concentrated and purified further by size exclusion chromatography using a YARRA SEC2000 PREP column equilibrated with base buffer. Fractions containing protein were pool, concentrated and stored at −80 °C.

### B.1.2 Crystal preparation

Crystallisation experiments were carried out by Elliot Nelson, Srikannathasan Velupillai and Patrick Collins.

DCP2 crystals were initially obtained by crystallisation trials against coarse screens [145]: Hampton Index (HIN), Ligand Friendly Screen (LFS), JCSG+ [316], Crystal Screen HT HCS (All supplied by Molecular Dimensions). Protein at 40 mg ml$^{-1}$ in base buffer (10 mM HEPES, 500 mM NaCl, 5% Glycerol, 0.5 mM TCEP, pH 7.5) was mixed, in 3 drop sitting drop vapour diffusion plates (SWISSCI 3 Lens Crystallisation Microplate) at three ratios 50:100nl, 75:75nl and 100:50nl with the crystallisation components present in the coarse screens. Two crystallisation conditions that led to well diffracting crystals (PDB ID: 5MP0,1.63 Å) were identified: 25%(w/v) PEG3350, 0.1 M acetate pH 4.5 and 0.1 M acetate pH 4.6, 0.2 M ammonium acetate, 30%(w/v) PEG 4000.

Micoseeding seed stocks [317] were prepared by flooding the crystallisation well with mother liquor from the reservoir well, then pipetting the crystals into a eppendorf tube containing a PTFE bead. The solution was vortexed for 1 minute, then serially diluted from 1:1 to 1:4096 with reservior solution in four-fold stages. These serial dilutions were used to set up crystallisation trials. Microseeded crystallisation trials were set-up in 3 drop sitting drop vapour diffusion plates, with protein at 18 to 22 mg ml$^{-1}$ in base buffer, at three ratios 1:2, 1:1 and 2:1 of protein to crystallisation components from the reservoir well, and 20 nl seed stock, to give either 220 or 320 nL total volume. Crystals used for fragment screening were from fine screens, with 1:16 diluted microseeds, with crystallisation conditions: 11-17% PEG 3350 with 0.1 M Acetate pH 4.5. Rod shaped crystals, or

clusters of crystals typically appeared within 12 to 24 hours. Fragment screening X-ray data collection occurred on beamline i04-1 at the Diamond Light Source (Harwell, UK).

### B.1.3 Fragment screening libraries

DCP2 crystals were soaked against two libraries, Diamond and SGC Poised Fragment Library (DSPL) [123] and Edelris library. Soaking experiments were undertaken as described in [108] and section 1.2.5.1. No additonal cryoprotectant was needed.

### B.1.4 DCP2 structure determination

Diffraction data were automatically processed by software pipelines at Diamond Light Source[109]. Preparation of ligand geometries (CIF generation) was carried out with AceDRG [2], eLBOW [5] or GRADE [4]. Initial refinement and map calculation was carried out with DIMPLE [6]. PanDDA [1] was used for hit identification and further refinement and model building was performed with REFMAC [8] and COOT [3] respectively. Unless mentioned otherwise, all structure determinations steps were carried out within XChemExplorer [115].

Coordinates and structure factor are available in the Protein Data Bank under group deposition ID: G_1002061. The complete PanDDA analysis, which was carried out in two stages due to updates to the platform, is available under DOI 10.5281/zenodo.1437589.

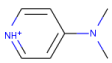
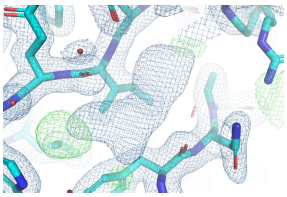
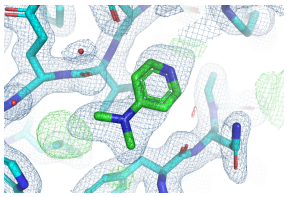
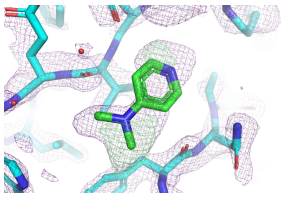
### B.1.5 Fragment hits from screening

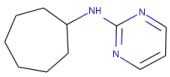
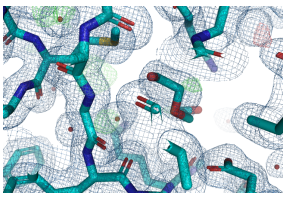
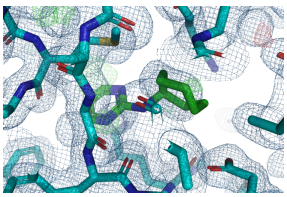
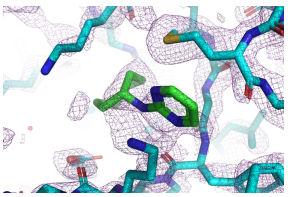
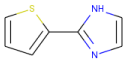
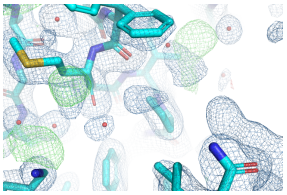
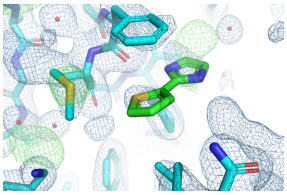
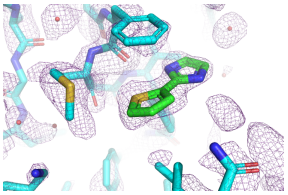
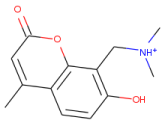
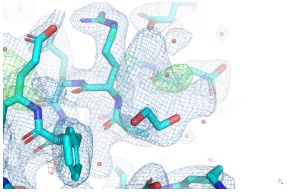
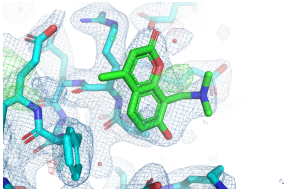
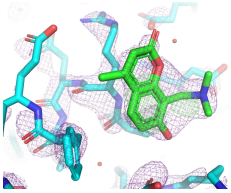
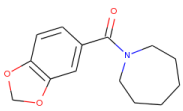
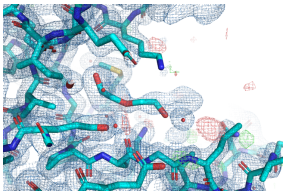
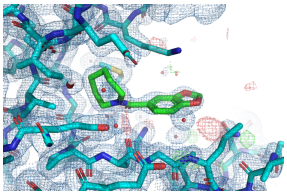
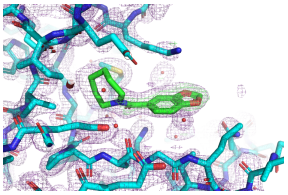
DCP2 was screened against 1035 compounds, 906 of which led to crystals. Of these 10 protein ligand complexes once refined, were of sufficiently high confidence for deposition to the protein databank.

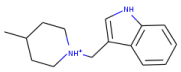
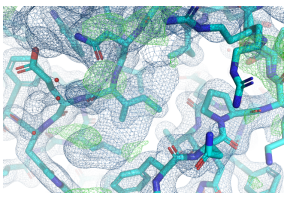
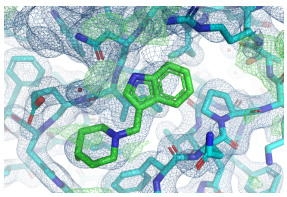
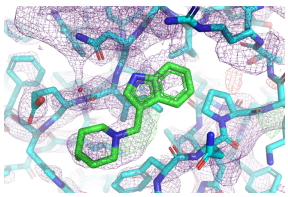
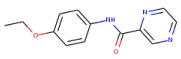
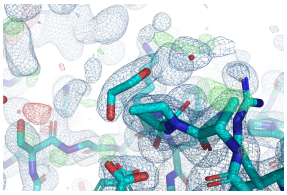
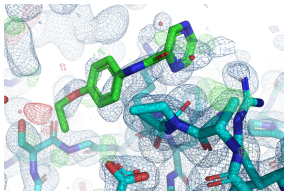
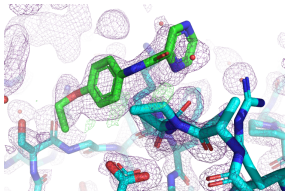
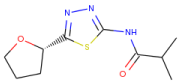
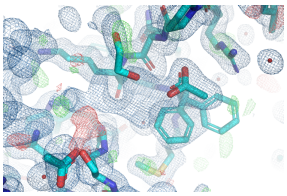
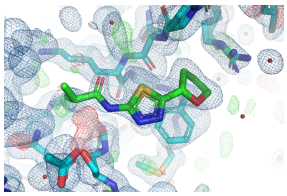
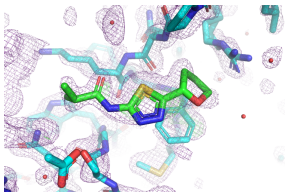
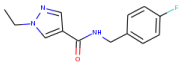
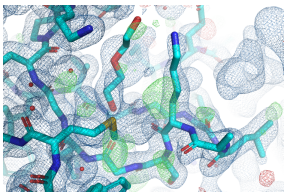
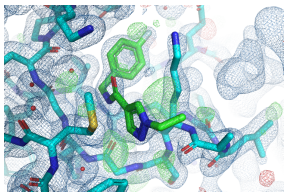
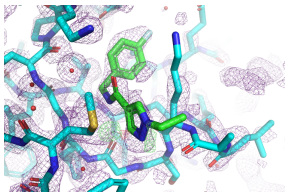
| Protein | DCP2 |
|---|---|
| Experiment Date | 2016-2017 |
| Crystallisation Conditions | 11-17% PEG 3350 with 0.1 M Acetate pH 4.5 |
| Cyro-Protectant (modal concentration) | None |
| Solvent (modal concentration) | DMSO (20%) |
| Mean Soaking Time (SD) | 3hr 9 mins (1 hr 31 mins) |
| Resolution Range [Å] | 1.25-3.51 |
| Mean Resolution (SD) [Å] | 2.03 (0.36) |
| Space Group | P 21 21 21 |
| Mean Unit Cell Axis | 47.67 60.64 64.85 |
| Unit Cell Axis SD [%] | 1.92 1.02 1.89 |
| Unit Cell Volume SD [%] | 5.00 |
| Median $R_{free}/R_{work}$ | 0.269/0.206 |
| Mean $R_{free} - R_{work}$ (SD) | 0.064 (0.018) |
| NCBI gene ID | 167227 |
| Domain Range | 95-260 |
| Number of unique Fragments Screened (DSPL) | 701 |
| Number of unique Fragments Screened (EDELRIS) | 206 |
| Number of unique compounds with crystals (DSPL) | 699 |
| Number of unique compounds with crystals (EDELRIS) | 205 |

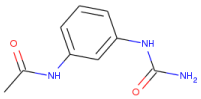
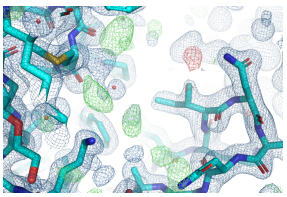
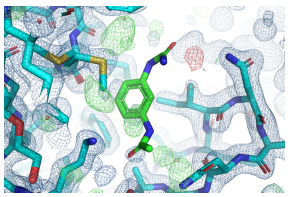
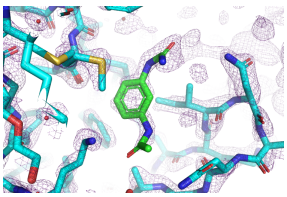
**Table B.1:** Summary of DCP2 fragment screening and follow up chemistry campaign.

**Table B.2:** Fragment screening hits for DCP2 (DCP2B refers to construct produced at SGC). Bound and ground structure shown with 1 $\sigma$ $2mF_{obs} - DF_{calc}$ and 3 $\sigma$ $mF_{obs} - DF_{calc}$. Event map contoured at 2 times the background density correction value, Z map contoured at 4 $\sigma$.

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x0016 | | | |
|  |  |  |  |
| XST00000847b (**8**) | | | 1-BDC: 0.19 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x0033 | | | |
|  |  |  |  |
| FMOPL000213a (**9**) | | | 1-BDC: 0.13 |
| DCP2B-x0042 | | | |
|  |  |  |  |
| FMOCR000171b (**10**) | | | 1-BDC: 0.17 |
| DCP2B-x0067 | | | |
|  |  |  |  |
| FMOPL000294a (**11**) | | | 1-BDC: 0.35 |
| DCP2B-x0146 | | | |
|  |  |  |  |
| FMOPL000435a (**2**) | | | 1-BDC: 0.27 |

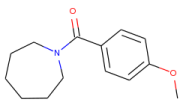
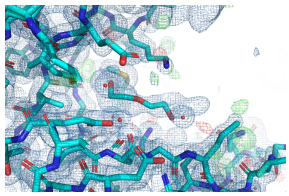
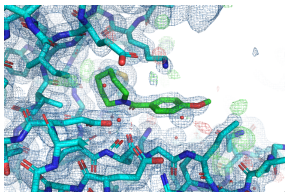
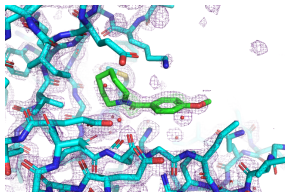
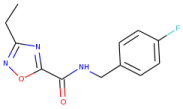
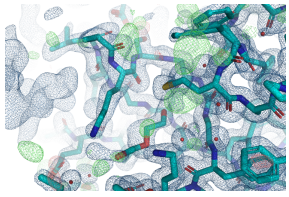
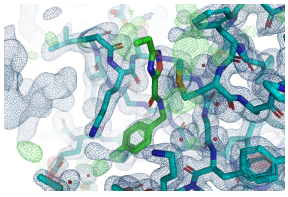
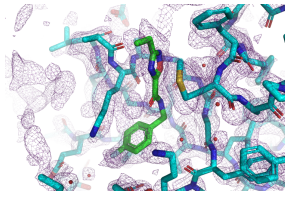
| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x0290 | | | |
|  FMOPL000576a (**12**) |  |  |  1-BDC: 0.21 |
| DCP2B-x0354 | | | |
|  FMOPL000446a (**13**) |  |  |  1-BDC: 0.18 |
| DCP2B-x0628 | | | |
|  FMOPL000144a (**14**) |  |  |  1-BDC: 0.23 |
| DCP2B-x1518 | | | |
|  FMOPL000449a (**15**) |  |  |  1-BDC: 0.27 |

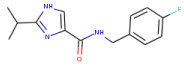
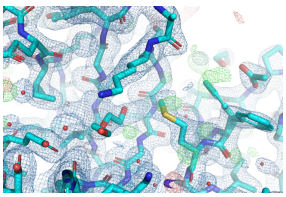
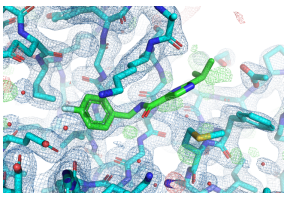
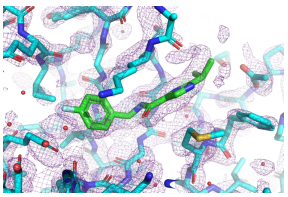
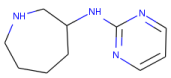
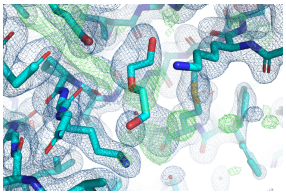
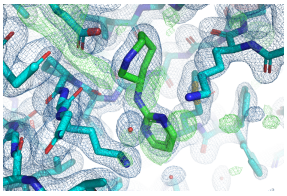
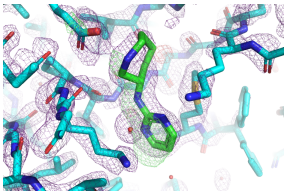
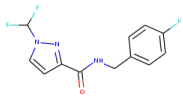
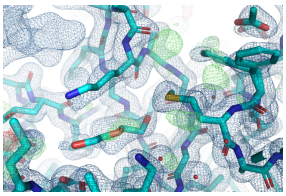
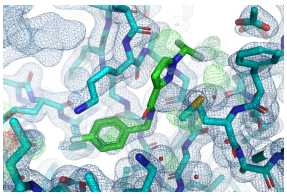
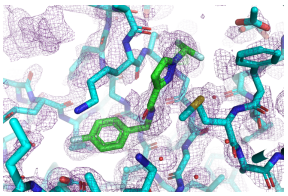
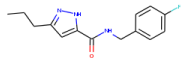
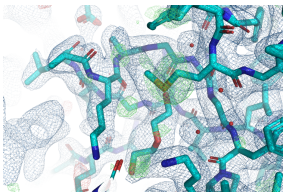
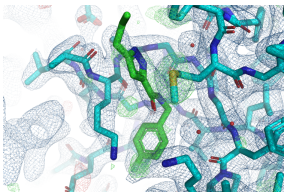
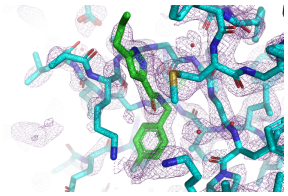
| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1561 | | | |
|  |  |  |  |
| FMOPL000701a (**16**) | | | 1-BDC: 0.14 |

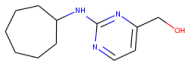
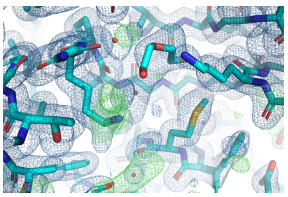
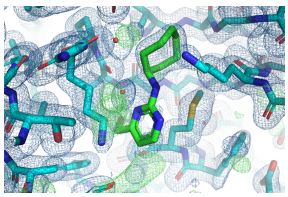
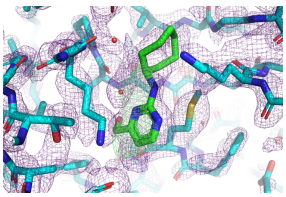
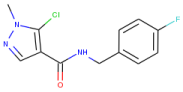
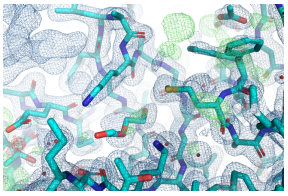
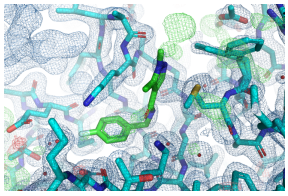
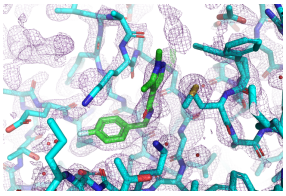
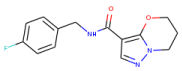
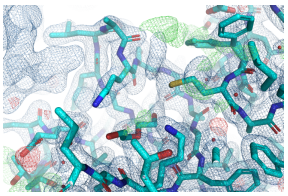
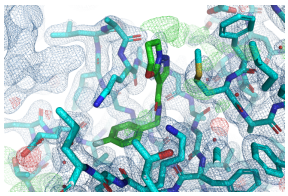
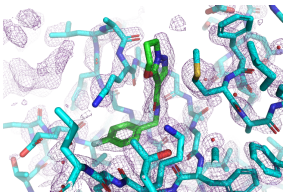
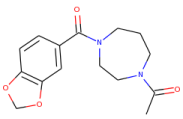
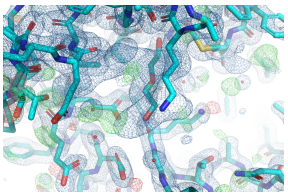
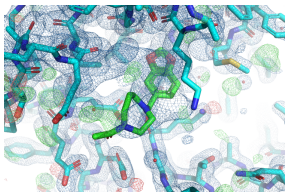
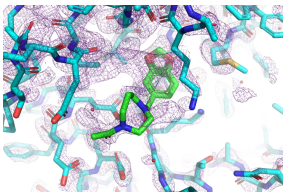
The binding pocket of DCP2 in complex with compound (**2**) is similar to the apo-form. The fragment compound (**9**) displaces two ordered water molecules. Compound (**15**) causes a conformational change in the position of Methionine 127 creating a distinct pocket.

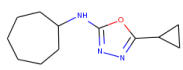
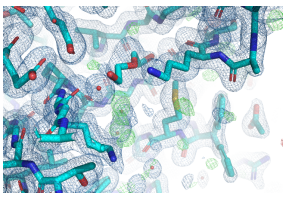
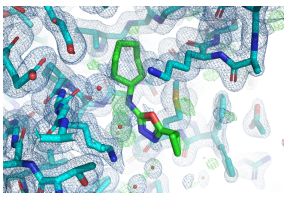
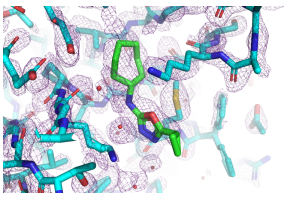
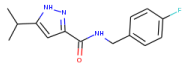
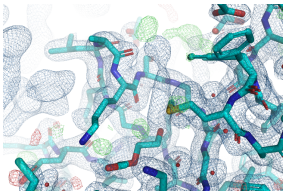
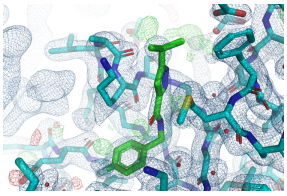
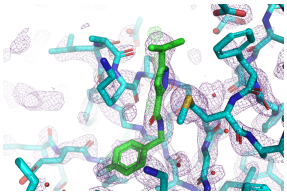
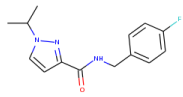
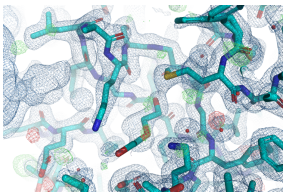
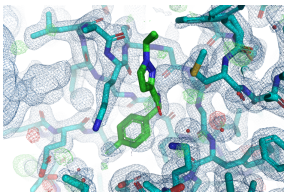
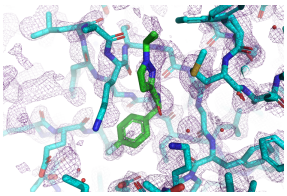
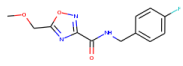
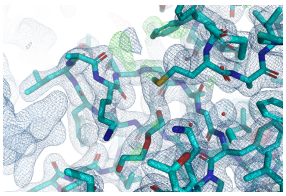
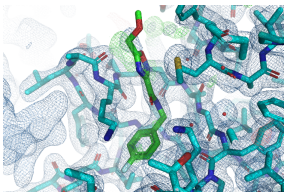
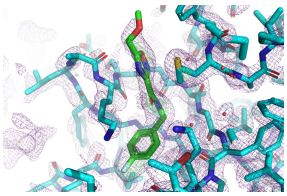
### B.1.5.1 Initial chemistry follow up

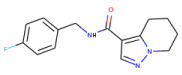
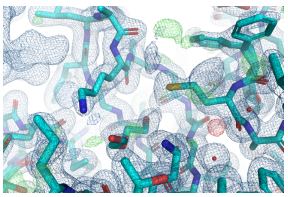
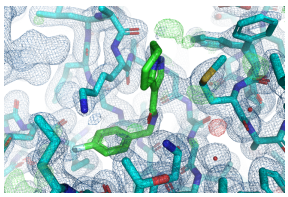
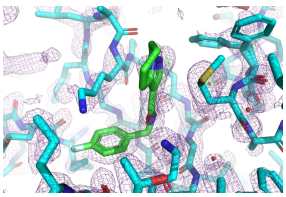
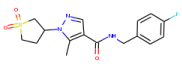
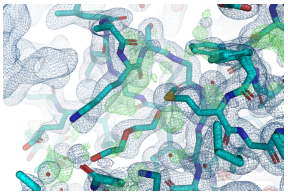
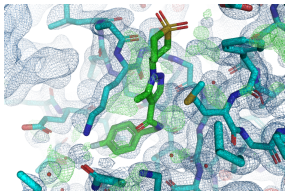
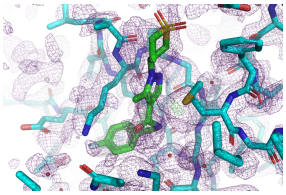
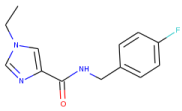
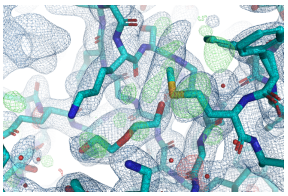
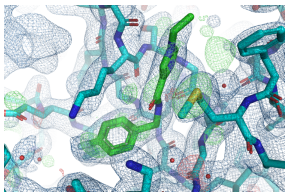
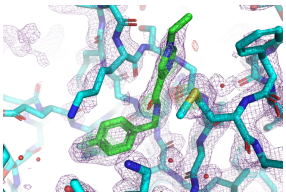
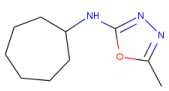
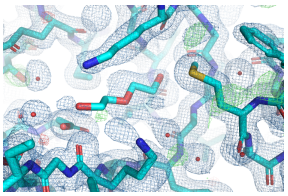
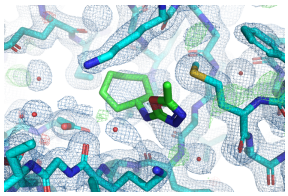
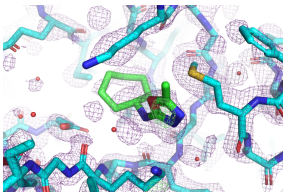
**Table B.3:** Follow up hits for DCP2. Bound and ground structure shown with 1 $\sigma$ $2mF_{obs} - DF_{calc}$ and 3 $\sigma$ $mF_{obs} - DF_{calc}$. Event map contoured at 2 times the background density correction value, Z map contoured at 4 $\sigma$.

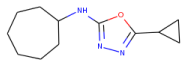
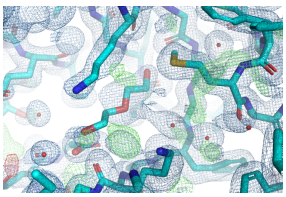
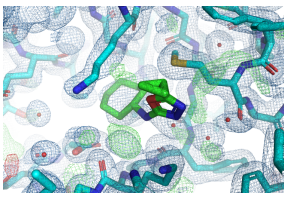
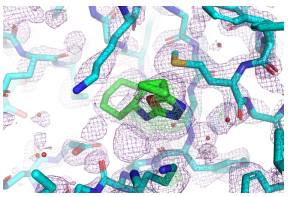
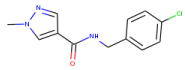
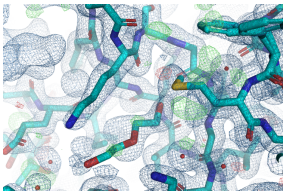
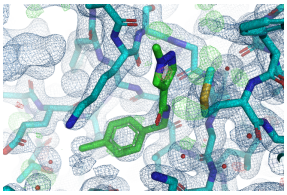
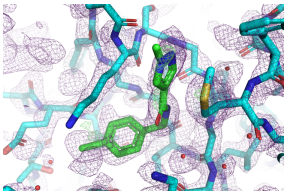
| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1181 | | | |
|  |  |  |  |
| NUOOA000023a (**17**) | | | 1-BDC: 0.36 |
| DCP2B-x1214 | | | |
|  |  |  |  |
| Z2895259680 (**18**) | | | 1-BDC: 0.09 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1217 | | | |
|  Z1203490773 (**19**) |  |  |  1-BDC: 0.17 |
| DCP2B-x1225 | | | |
|  Z1187701032 (**20**) |  |  |  1-BDC: 0.39 |
| DCP2B-x1235 | | | |
|  Z1592710382 (**21**) |  |  |  1-BDC: 0.11 |
| DCP2B-x1246 | | | |
|  Z296300542 (**22**) |  |  |  1-BDC: 0.30 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1247 | | | |
|  |  |  |  |
| Z1699011516 (**23**) | | | 1-BDC: 0.18 |
| DCP2B-x1256 | | | |
|  |  |  |  |
| Z2895259675 (**24**) | | | 1-BDC: 0.11 |
| DCP2B-x1259 | | | |
|  |  |  |  |
| Z2212124043 (**25**) | | | 1-BDC: 0.17 |
| DCP2B-x1276 | | | |
|  |  |  |  |
| YW-FY-378 (**26**) | | | 1-BDC: 0.10 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1344 | | | |
|  PB1787571279 (**27**) |  |  |  1-BDC: 0.16 |
| DCP2B-x1361 | | | |
|  Z454376544 (**28**) |  |  |  1-BDC: 0.27 |
| DCP2B-x1362 | | | |
|  Z1190363272 (**29**) |  |  |  1-BDC: 0.11 |
| DCP2B-x1367 | | | |
|  Z1170065264 (**30**) |  |  |  1-BDC: 0.19 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1372 | | | |
|  |  |  |  |
| Z1494850193 (**31**) | | | 1-BDC: 0.14 |
| DCP2B-x1382 | | | |
|  |  |  |  |
| Z450133538 (**32**) | | | 1-BDC: 0.10 |
| DCP2B-x1393 | | | |
|  |  |  |  |
| Z2895259681 (**33**) | | | 1-BDC: 0.48 |
| DCP2B-x1475 | | | |
|  |  |  |  |
| PB1230873739 (**34**) | | | 1-BDC: 0.14 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| DCP2B-x1479 | | | |
|  |  |  |  |
| PB1787571279 (**27**) | | | 1-BDC: 0.13 |
| DCP2B-x1502 | | | |
|  |  |  |  |
| Z100435060 (**35**) | | | 1-BDC: 0.13 |

I was also involved in the crystallographic refinement of the first round follow up compounds (Laura Diaz Saez, Paul Brennan, Patrick Collins) of approximately 50 compounds were initially designed as analogues of compound (**2**). The follow up compound (**17**) was shown to bind both crystallographically (shown in table B.3) and with surface plasmon resonance to bind at 307 µM. Another 138 follow up compounds were obtained following up from compound (**2**) and compound (**15**) . 48 compounds were synthesised by Enamine leading to 17 X-ray complexes, 14 of these were flourobenzyl compounds which bound similiarly to (**35**), the remaining 3 hits included a 7-membered ring; including complexes with compounds (**20**) and (**23**). 24 compounds synthesised by collaborators at Fudan university; 2 of which were found to bind using X-ray crystallography, The complex with compound (**35**) shows binding, the chlorobenzene overlapping with the binding of the 7 membered ring of compound (**26**). 67 compounds were purchased, of which 3 led to X-ray complexes, 8 were shown to bind weakly in surface plasmon resonance, but there were no hits found in both X-ray and SPR.

## B.2 *Trypanosoma. Cruzi* Farneysl pyrophosphate synthase

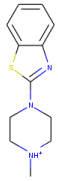
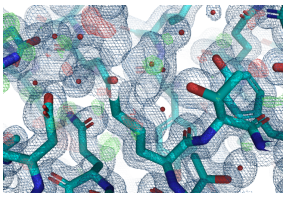
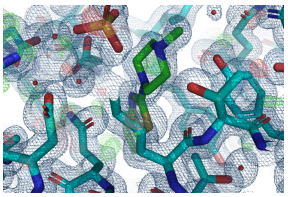
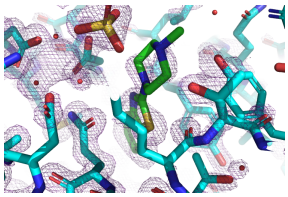
Here we briefly report the results of the XChem facility fragment screen of *Trypanosoma Cruzi* Farnesyl pyrophosphate synthase.

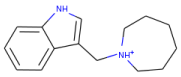
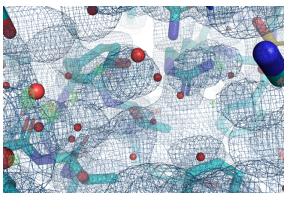
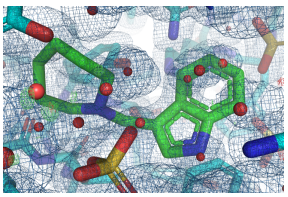
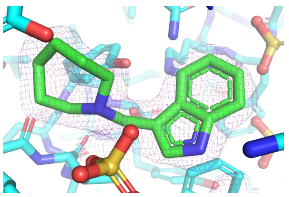
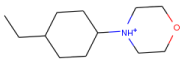
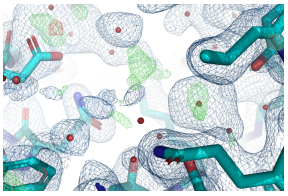
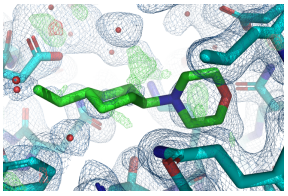
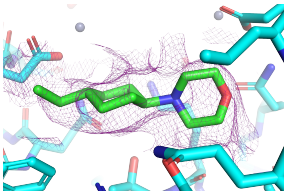
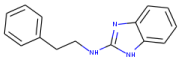
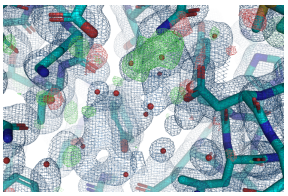
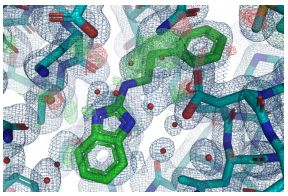
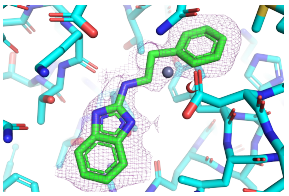
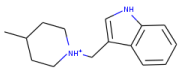
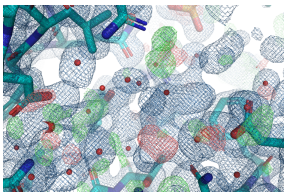
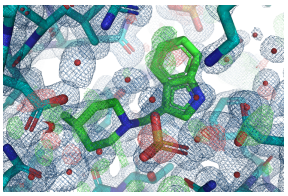
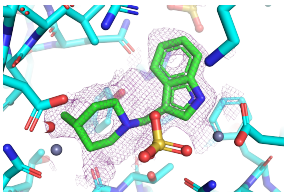
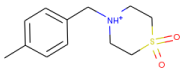
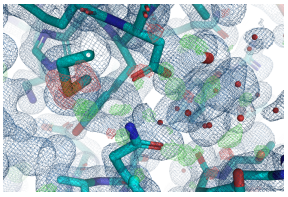
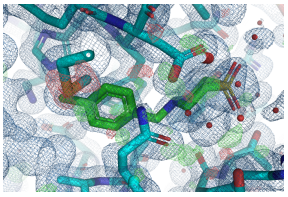
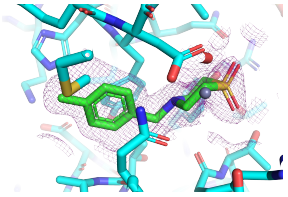
666 datasets of *T. cruzi* were collected on the XChem platform, and analysed using PanDDA [257] using the XChemExplorer platform [115]. 51 binding events with non-bisphosphonate scaffolds were obtained. As 12 structures showed binding to multiple copies of the same ligand, there were 35 protein ligand complexes, spread over 9 sites which were deposited in the protein databank. Binding sites were distributed all over the protein, including the active site, the allosteric site, the homodimer interface, sites on the surface and a new site in close proximity to the active site.

Note that a second round of screening was carried out at the Crystallisation lab at the European Molecular biology Laboratory's end-station at the European Synchrotron Research Facility, are not presented herein. Similar hit rates were obtained, and knowledge of metal ion binding from this work proved valuable for expedited structure solution.

Table B.4 summarises the XChem fragment screening campaign. Table B.5 shows the active site hits. Table B.6 shows the allosteric hits. Table B.7 shows a novel binding site hit. Note that the remaining six sites were not proximal to the metal ion binding or active site, and thus are not discussed herein. All structures, electron density information and PanDDA event maps can be found at (DOI 10.5281/zenodo.2649077).

**Table B.5:** Active site screening hits for T.cruzi FPPS. Bound and ground structure shown with 1 $\sigma$ $2mF_{obs} - DF_{calc}$ and 3 $\sigma$ $mF_{obs} - DF_{calc}$. Event map contoured at 2 times the background density correction value.

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| TCRUFPPS-x0064 | | | |
|  |  |  |  |
| FMOPL000478a (**36**) | | | 1-BDC: 0.35 |

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| TCRUFPPS-x0076 | | | |
|  |  |  |  |
| FMOPL000291a (**37**) | | | 1-BDC: 0.57 |
| TCRUFPPS-x0086 | | | |
|  |  |  |  |
| FMOPL000315a (**38**) | | | 1-BDC: 0.21 |
| TCRUFPPS-x0163 | | | |
|  |  |  |  |
| FMOPL000586a (**39**) | | | 1-BDC: 0.28 |
| TCRUFPPS-x0196 | | | |
|  |  |  |  |
| FMOPL000576a (**12**) | | | 1-BDC: 0.40 |
| TCRUFPPS-x0284 | | | |
|  |  |  |  |
| FMOPL000642a (**40**) | | | 1-BDC: 0.29 |

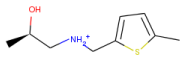
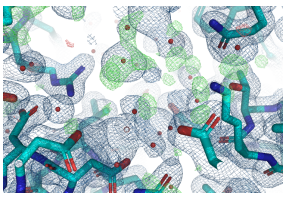
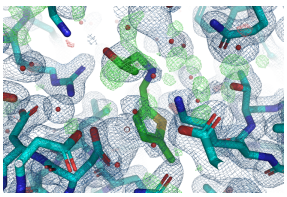
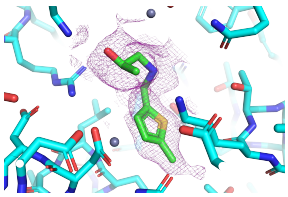
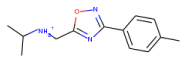
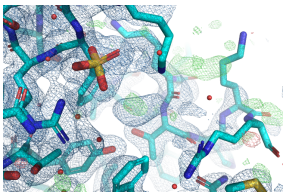
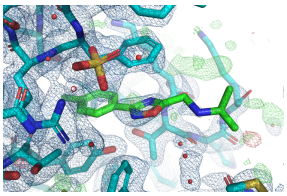
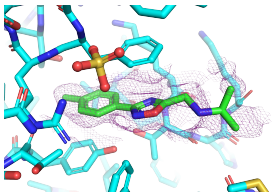
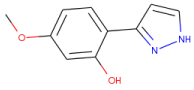
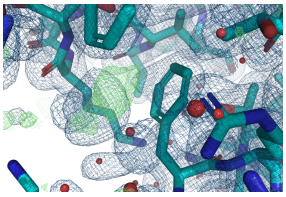
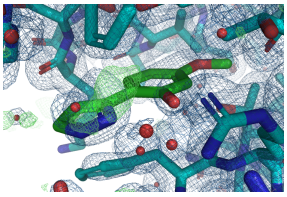
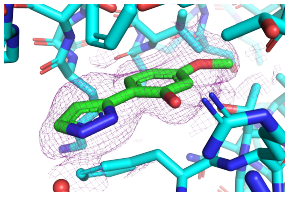
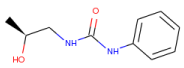
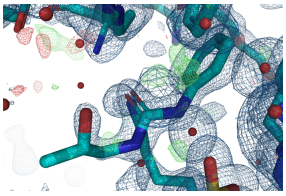
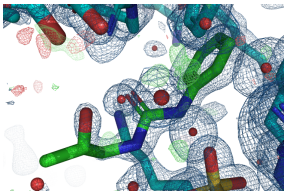
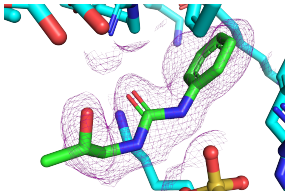
| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| TCRUFPPS-x0359 | | | |
|  |  |  |  |
| FMOPL000693a (**41**) | | | 1-BDC: 0.32 |

**Table B.6:** Allosteric site screening hits for T.cruzi FPPS. Bound and ground structure shown with 1 $\sigma$ $2mF_{obs} - DF_{calc}$ and 3 $\sigma$ $mF_{obs} - DF_{calc}$. Event map contoured at 2 times the background density correction value.

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| TCRUFPPS-x0051 | | | |
|  |  |  |  |
| FMOPL000293a (**42**) | | | 1-BDC: 0.25 |
| TCRUFPPS-x0106 | | | |
|  |  |  |  |
| FMOPL000554a (**7**) | | | 1-BDC: 0.25 |
| TCRUFPPS-x0165 | | | |
|  |  |  |  |
| FMOPL000464a (**43**) | | | 1-BDC:0.38 |

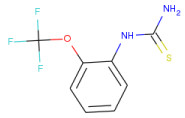
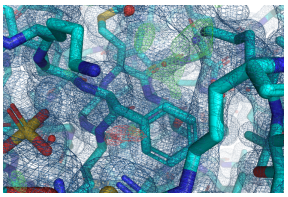
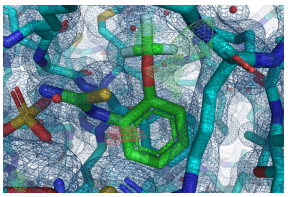
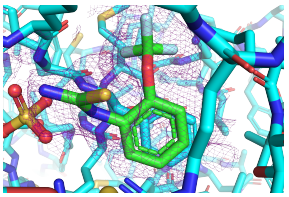
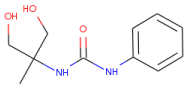
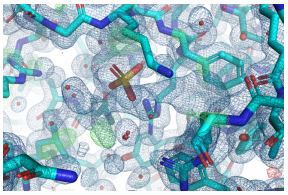
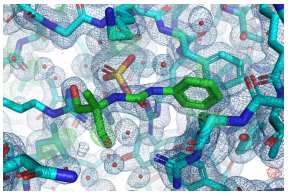
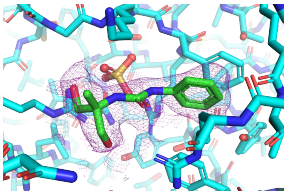
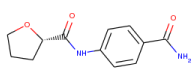
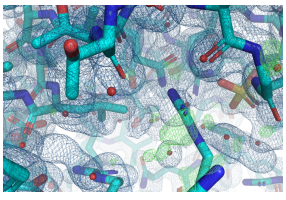
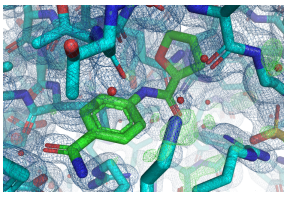
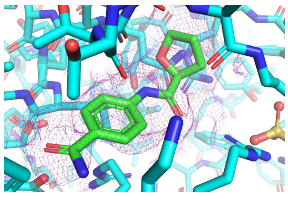
| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| TCRUFPPS-x0231 | | | |
|  |  |  |  |
| FMOPL000512a (**44**) | | | 1-BDC:0.27 |
| TCRUFPPS-x0325 | | | |
|  |  |  |  |
| FMOPL000524a (**45**) | | | 1-BDC:0.34 |

**Table B.4:** Summary of *T. cruzi* FPPS fragment screening.

| Protein | *T. cruzi* FPPS |
|---|---|
| Experiment date | October 2017 |
| Crystallisation Conditions | Reservoir: pH 6.5 80 mM MES, 4 mM $ZnSO_4$, 12.36% w/v PEG MME 550, 11.57% v/v glycerol |
| Cyroprotectant (modal concentration) | None |
| Solvent (modal concentration) | DMSO (15%) |
| Mean soaking time (SD) | 3hr 32 mins (39 mins) |
| Resolution range [Å] | 1.40-3.26 |
| Mean resolution (SD) [Å] | 1.75 (0.26) |
| Space group | P 61 2 2 |
| Mean unit cell axis | 57.045 57.045 395.47 |
| Unit cell axis SD [%] | 0.37 0.37 0.23 |
| Unit cell volume SD [%] | 0.80 |
| Median $R_{free}/R_{work}$ | 0.261/0.2162 |
| Mean $R_{free} - R_{work}$ (SD) | 0.041 (0.012) |
| Uniprot ID | Q8WS26 |
| Domain range | Met 64 - Lys 425. No mutations |
| Number of unique fragments screened (DSPL) | 406 |
| Number of unique fragments screened (EDELRIS) | 279 |
| Number of unique compounds with crystals (DSPL) | 395 |
| Number of unique compounds with crystals (EDELRIS) | 271 |
| Fragment Screening Hits | 35 |

**Table B.7:** Novel site screening hit for T.cruzi FPPS. Bound and ground structure shown with 1 $\sigma$ $2mF_{obs} - DF_{calc}$ and 3 $\sigma$ $mF_{obs} - DF_{calc}$. Event map contoured at 2 times the background density correction value.

| Compound | Ground Structure | Bound Structure | Event Map |
|---|---|---|---|
| TCRUFPPS-x0304 | | | |
|  |  |  |  |
| FMOPL000416a (**46**) | | | 1-BDC:0.21 |

# B.3 NUDT22

## B.3.1 Protein preparation

NUDT22 purification was carried out by Laura Diaz Saez, Paulina Siejka and Ioanna Panagakou.

4 l of TB media were prepared: 24 g $l^{-1}$ yeast extract granular, 12 g $l^{-1}$ peptone and 4 ml $l^{-1}$ 100% glycerol. A 20× stock of salts were prepared in 700 ml with 32.34 g potassium phosphate monobasic,$KH_2PO_4$ and 175.56 g potassium phosphate dibasic, $K_2HPO_4$. Media and salts were autoclaved.

1 l solutions were prepared in flasks containing 950 ml of TB media,50 ml salts, 1 ml 50 µg $ml^{-1}$ kanamycin and 1 ml 34 µg $ml^{-1}$ chloramphenicol and 10 ml preculture were incubated at 37 °C and 200rpm. After 6 hours the flasks were cooled at 6 °C for 20 mintues. Protein expression was induced using 200 µM IPTG. The flasks were incubated overnight at 18 °C.

80 g of bacteria were collected via centrifugation at 6000rpm for 25 minutes. The cell mass was re-suspended in lysis buffer (20 mM HEPES pH7.5, 500 mMNaCl, 20 mM Imidazole, 5% Glycerol, 0.5 mM TCEP with protease inhibitors) at 4 °C until homogenous. The re-suspended bacteria were then passed through a Homogeniser (EmulsiFlex-c5), once without pressure applied, and three times with a pressure of 7000-14000 psi applied. Debris was removed via centrifugation at 15000rpm for 1 hour at 4 °C.

Purification was undertaken using a AKTAPrime (GE Healthcare) fitted with a HisTrap Crude FF 5 ml column with flow of 6 ml $min^{-1}$, a 50 ml gradient with a 100% target and set to aliqout 1.8 ml samples. Sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE) and subsequent tandem MS showed the target protein only. Fractions were pooled and placed in SnakeSkin Dialysis Tubing, 3.5K MWCO for dialysis overnight with TEV protease added.

The dialysis sample was incubated with nickel beads 3 ml and the flowthrough collected. Beads were washed 4 times by 5 ml buffer (20 mM HEPES pH7.5, 500 mMNaCl, 20 mM Imidazole, 5% Glycerol, 0.5 mM TCEP) and eluted with 2 times 5 ml elution buffer (20 mM HEPES pH7.5, 500 mMNaCl, 500 mM Imidazole, 5% Glycerol, 0.5 mM TCEP).

Gel filtration (size exclusion chromatography) was carried out using AKTA Express (GE Healthcare) on a HiLoad 16/600 Superdex 75 pg column with a flow of 1 ml $min^{-1}$, and aliquoted into volumes of 3 ml.

Protein was concentrated to 13 mg $ml^{-1}$ using 10 kDa centrifugal filter.

#### B.3.1.0.1 Crystallisation

NUDT22 crystallisation was carried out by Laura Diaz Saez and Romain Talon.

Swiss CI 3 Drop plates were set-up using 150nl drops using a ratio of 2:1, 1:1 and 1:2 of protein:reservoir solution. The crystallisation condition across the plate were found during previous optimisation: 0.1 M HEPES pH 7.5, 0.3 M NaKPO, 15% PEG Smear High (50% v/v PEG Smear High, Molecular Dimensions) and 20% ethylene glycol.

## B.4 NUDT7 purification

NUDT7 purification was initially carried out by Michael Fairhead (SGC), I carried out the purification used for crystals this thesis.

NUDT7 was cloned into pNIC28-Bsa4 with a TEV-cleavable N-terminal hexa-histidine tag. After transformation into *E.coli* (BL21(DE3)-R3), the colonies were used to setup 100 ml overnight cultures were incubated at 37 °C and 180 rpm, using LB media supplemented with 50 µg $l^{-1}$ kanamycin and 34 µg $l^{-1}$ chloramphenicol.

For culture growth, TB medium (FroMedium), was supplemented with 20 g $l^{-1}$ glycerol, 50 µg $l^{-1}$ kanamycin and 34 µg $l^{-1}$ chloramphenicol, 1 ml $l^{-1}$ 10% antifoam, 1 ml $l^{-1}$ *E.coli* metals, as defined in appendix B.5.1, 0.01 M ammonium sulphate, 0.01 M magnesium sulphate and 20 ml $l^{-1}$ 25% glucose. Cultures were grown for four hours at 37 °C and 180rpm, then the temperature was decreased to 18 °C. 300 µl $l^{-1}$ 1 M IPTG was added to the cultures to induce protein expression. The cultures were incubated for another 20 hours.

Cells were spun at 4,700 rpm for 20 minutes at 4 °C, then resuspended in lysis buffer (50 mM HEPES, 500 mMNaCl, 20 mM Imidazole, 5% Glycerol, 0.5 mM TCEP, 0.5 mg $ml^{-1}$ lysozyme and 1 µg $ml^{-1}$ benzonase) supplemented with a cocktail of protease inhibitors. After a 30 minute room temperature incubation, 1% Triton X-100 was added and the cells were frozen at −80 °C.

On thawing, cells were centrifuged for one hour at 4,700 rpm at 4 °C and the supernatant applied to a His GraviTrap column (GE healthcare) equilibrated with binding buffer (10 mM HEPES, 5% glycerol, 500 mM NaCl and 0.5 mM TCEP equilibrated to pH 7.5). After washing with binding buffer supplemented to 20 mM imidazole, NUDT7 was eluted with binding buffer supplemented to 500 mM imidazole. The eluted protein was applied to a PD-10 desalting column (GE Healthcare) and eluted with binding buffer supplemented to 20 mM imidazole. The N-terminal affinity tag was removed by TEV cleavage overnight and uncleaved protein was removed by applying it again to a His GraviTrap column. A similar method has been reported in [94].

## B.5 Purification details

### B.5.1 *E.coli* metals

Preparation of *E. coli* metals used alongside media for increased viability of *E.coli*:

- 27 g $l^{-1}$ $FeCl_2 \cdot 6\,H_2O$
- 2 g $l^{-1}$ $ZnCl_2 \cdot 4\,H_2O$
- 2 g $l^{-1}$ $ZnCl_2 \cdot 4\,H_2O$
- 2 g $l^{-1}$ $CoCl_2 \cdot 6\,H_2O$
- 2 g $l^{-1}$ $Na_2MoO_4 \cdot 2\,H_2O$,
- 1 g $l^{-1}$ $CaCl_2 \cdot 2\,H_2O$
- 1 g $l^{-1}$ $CuCl_2$
- 0.5 g $l^{-1}$ $H_3BO_3$
- 100 ml 37% HCl

# C

# Superposed modelling

Figure C.1 shows the structure of the site for *T. cruzi* FPPS metal binding site, shown without extra water molecules which were added to structure shown in main text.

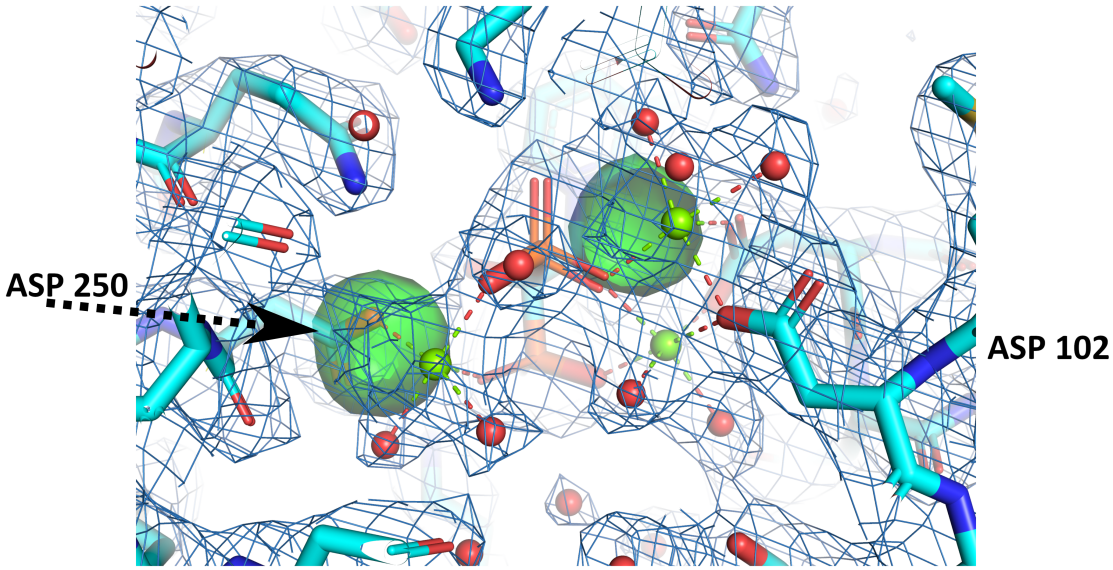
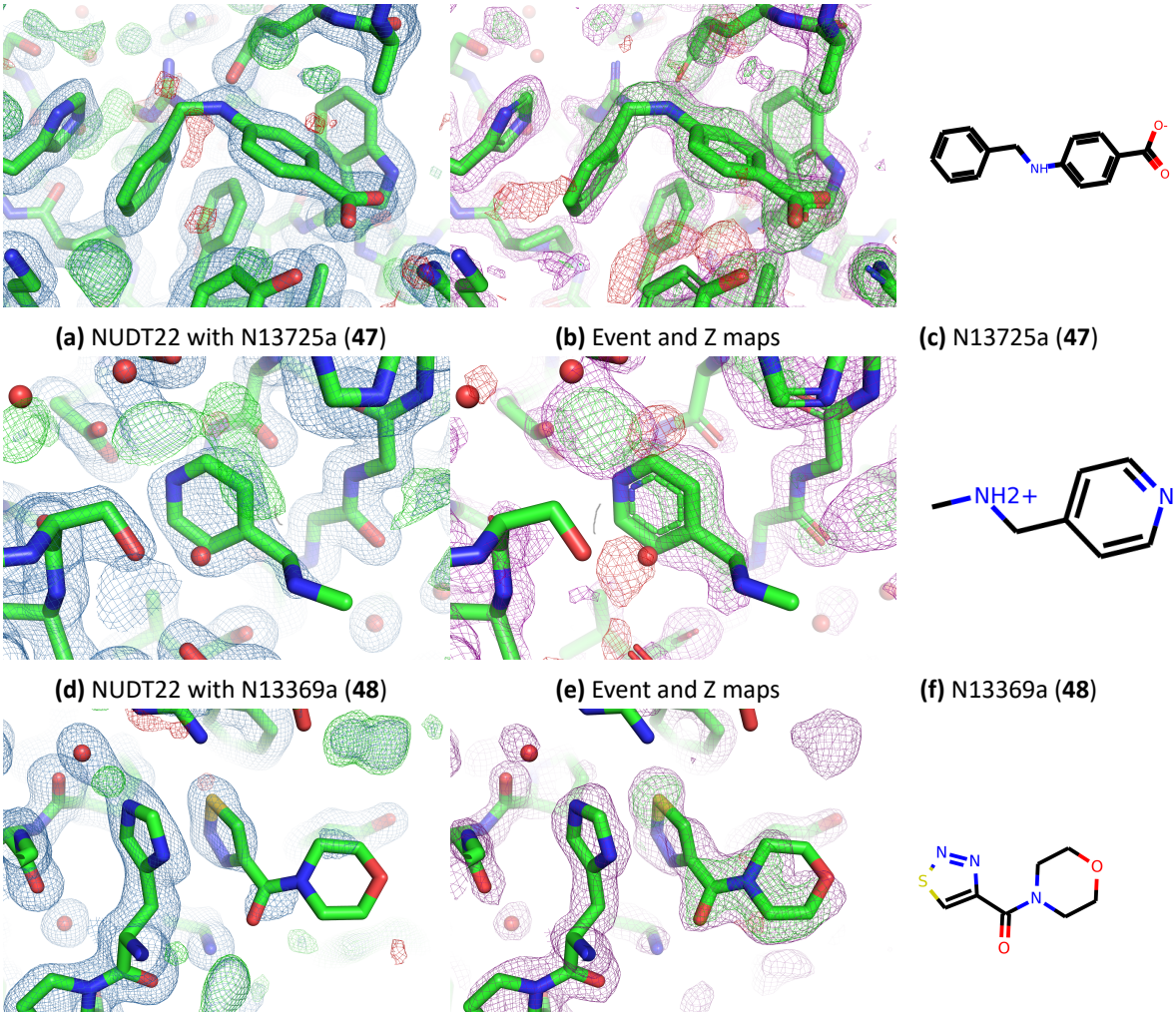



**Figure C.1:** *T. cruzi* FPPS metal binding site, shown without extra water molecules which were added to structure shown in main text.

# D

# Repeat datasets of protein-ligand complexes for assessing refinement programs

In this section we show the additional repeat datasets of protein-ligand complexes discussed in chapter 3. Figures D.1 and D.2 show the electron density maps, event maps, Z maps and compounds studied. Figures D.3 and D.4 show the distribution of occupancy and B factor for repeat datasets of protein-ligand complexes.

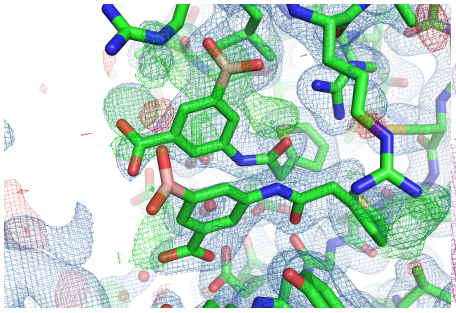



**(a)** NUDT22 with N13725a (**47**) **(b)** Event and Z maps **(c)** N13725a (**47**)

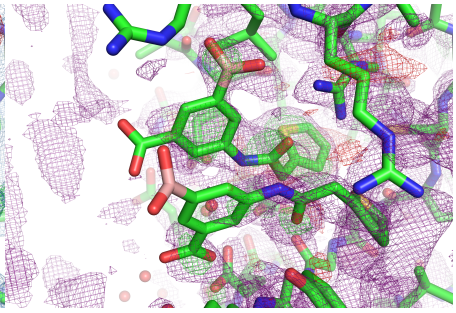
**(d)** NUDT22 with N13369a (**48**) **(e)** Event and Z maps **(f)** N13369a (**48**)

**(g)** NUDT22 with N13663a (**49**) **(h)** Event and Z maps **(i)** N13663a (**49**)

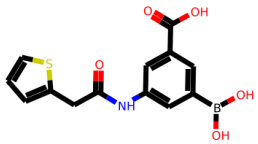
**Figure D.1:** Binding evidence for the ligands in repeat datasets of protein-ligand complexes. $2mF_{obs} - DF_{calc}$ electron density maps contoured at 1 $\sigma$, $mF_{obs} - DF_{calc}$ electron density maps contoured at 3 $\sigma$. Event maps contoured at twice the background density correction value. Z maps countered at $\pm 4.0\sigma$.



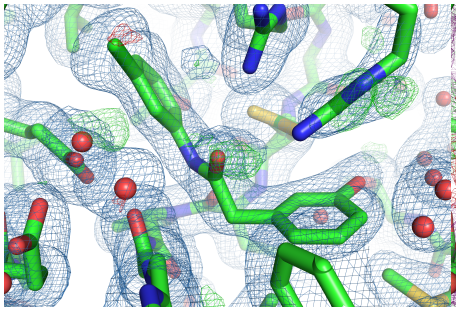
**(a)** NUDT7 with Two copies of OX210 (**50**)



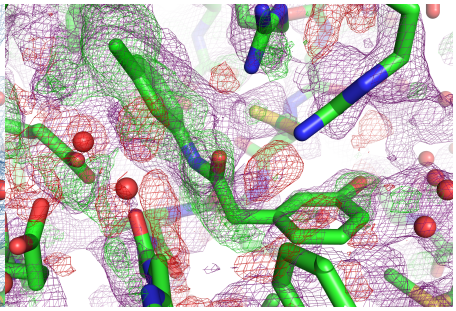
**(b)** Event and Z maps



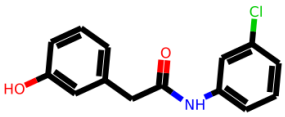

**(c)** OX210 (**50**)



**(d)** NUDT7 with NUOOA000181a (**51**)

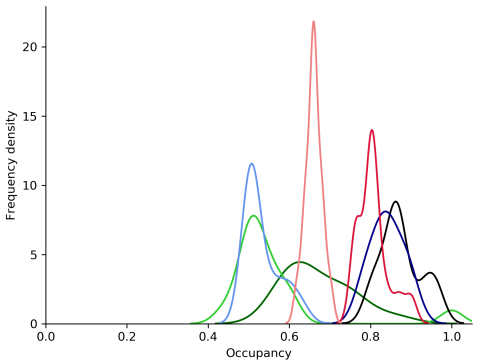


**(e)** Event and Z maps

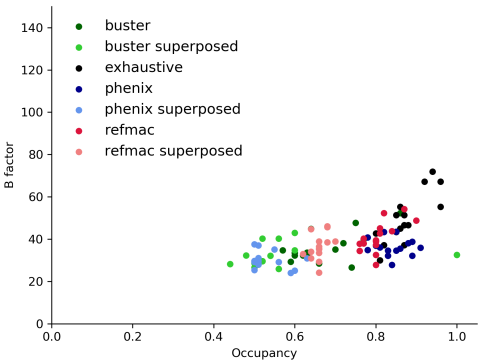



**(f)** NUOOA000181a (**51**)

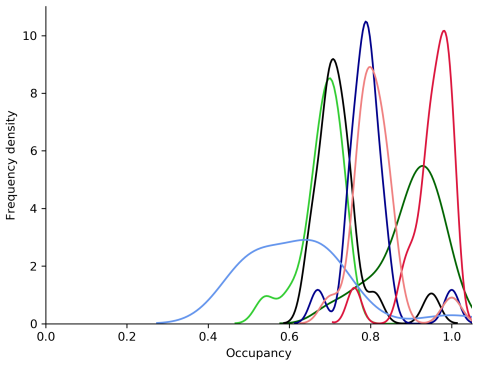
**Figure D.2:** Binding evidence for the ligands in repeat datasets of protein-ligand complexes. $2mF_{obs} - DF_{calc}$ electron density maps contoured at 1 $\sigma$, $mF_{obs} - DF_{calc}$ electron density maps contoured at 3 $\sigma$. Event maps contoured at twice the background density correction value. Z maps countered at $\pm 4.0\sigma$. Compounds shown in **(c)** and **(f)**



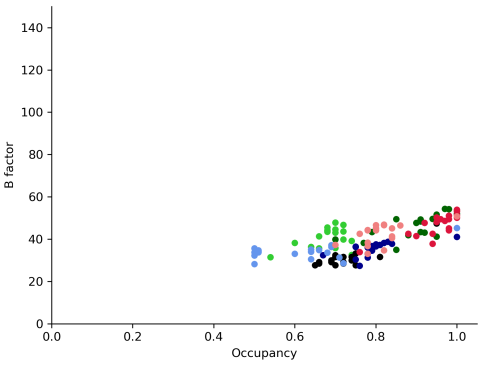

**(a)** NUDT22 with N13725a: n=15 (**47**)



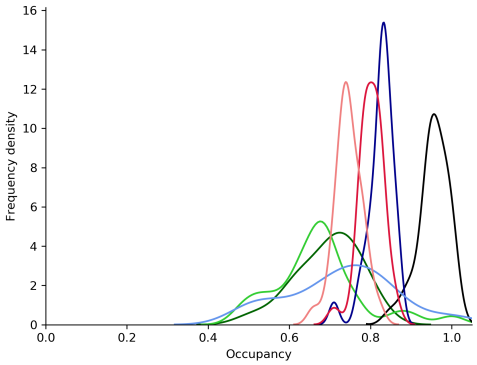

**(b)** NUDT22 with N13725a (**47**)



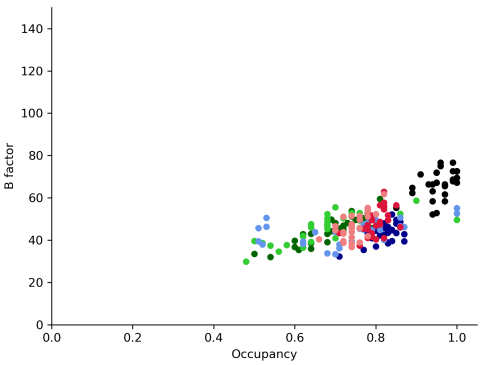

**(c)** NUDT22 with N13369a: n=18 (**48**)




**(d)** NUDT22 with N13369a (**48**)

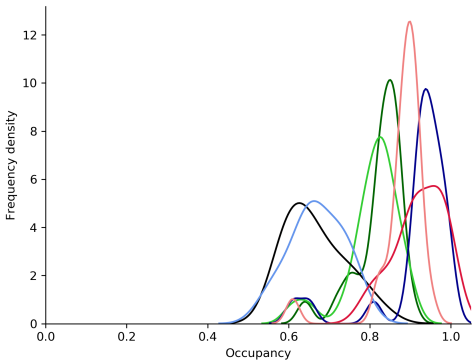



**(e)** NUDT22 with N13663a: n=29 (**49**)

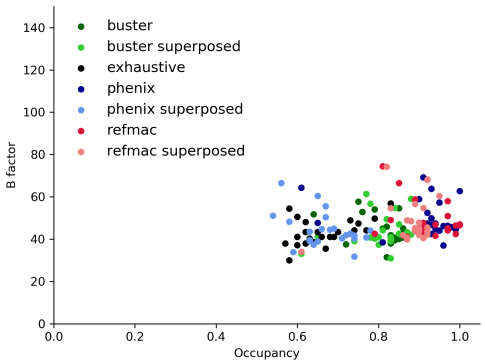



**(f)** NUDT22 with N13663a (**49**)

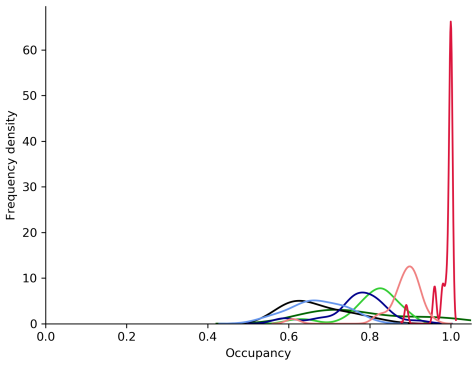
**Figure D.3:** Distribution of occupancy and B factor for repeat datasets of protein-ligand complexes. Figures **(a)** **(c)** and **(e)** show kernel density plots showing the distribution of occupancy for three different protein-ligand complexes, Figures **(b)** **(d)** and **(f)** show scatter pltos of ligand mean B factor and occupancy.



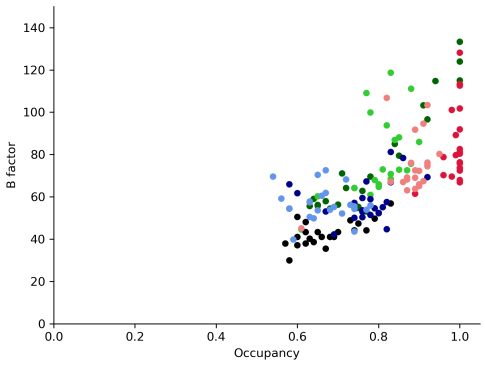

**(a)** NUDT7 with OX210: n=23 (**50**)



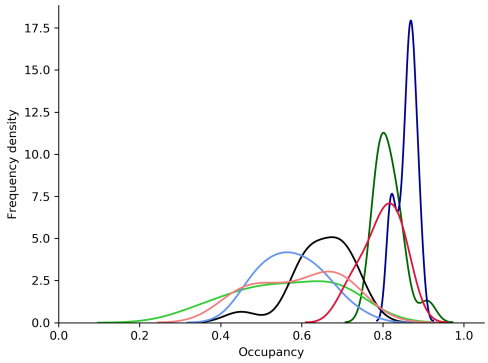

**(b)** NUDT7 with OX210 (**50**)



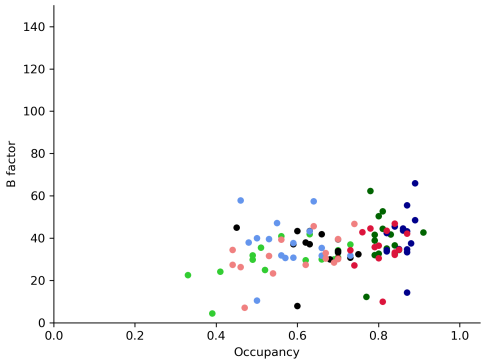

**(c)** NUDT7 with OX210 (**50**), second ligand: n=23




**(d)** NUDT7 with OX210 (**50**) second ligand copy




**(e)** NUDT7 with NUOOA000181a: n=15 (**51**)




**(f)** NUDT7 with NUOOA000181a (**51**)

**Figure D.4:** Distribution of occupancy and B factor for repeat datasets of protein-ligand complexes. Figures **(a) (c)** and **(e)** show kernel density plots showing the distribution of occupancy for three different protein-ligand complexes, Figures **(b) (d)** and **(f)** show scatter pltos of ligand mean B factor and occupancy.

# E

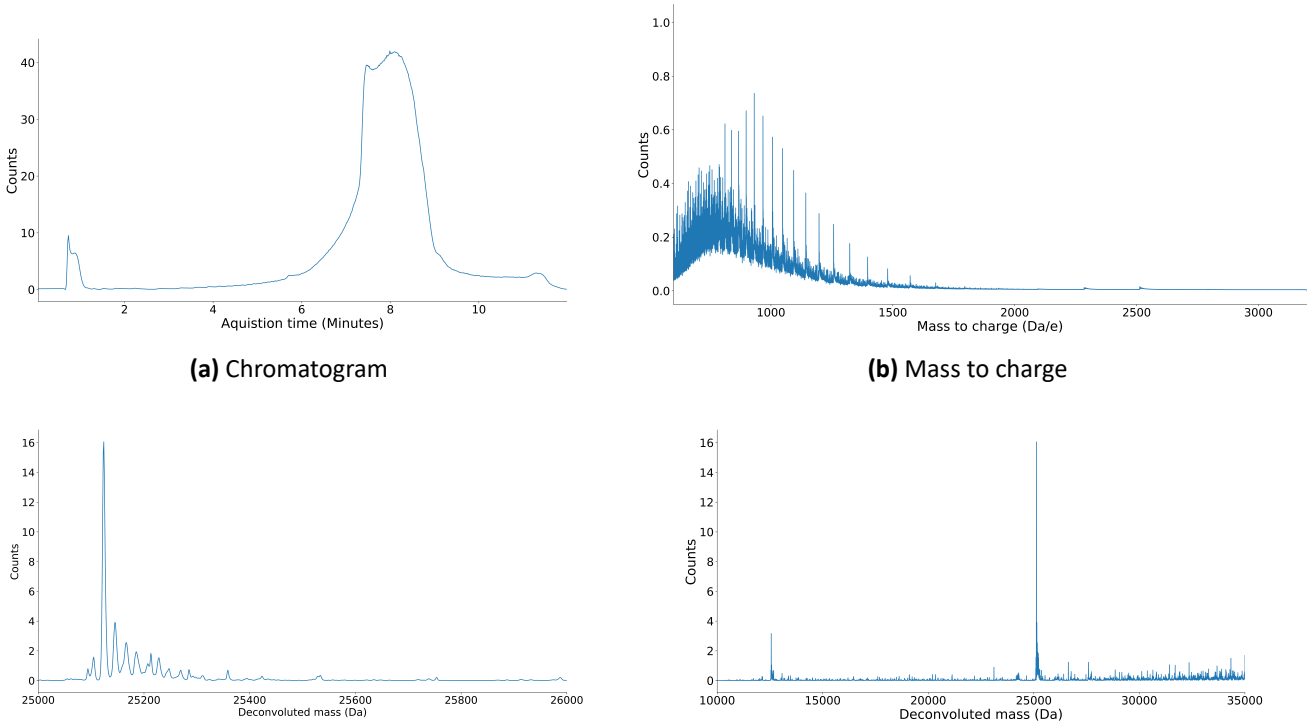
# Intact mass spectrometry of NUDT7

## E.1 Alternate sample preparation: Titanium dioxide spin columns

Titanium dioxide has been shown to separate PEG 4000 Da from protein samples [252].

Here we adapt the method outlined in [252] alongside the initial preparation of a titanium dioxide spin column using an existing SGC method: 5 µg of titanium dioxide beads (Titansphere TiO 10 µm medium, GL-Sciences cat no. 520-75010 supplied via Hichrom) are wieghed out. A disc of the filter is punched out (3M Empore $C_{18}$ High performance extraction (filter) disc 2215) using a cannelation needle and packed it into a 10 µl pipette tip using another blunt needle such that it works as a frit. A 10 µl pipette tip is packed with the TiO medium. A pair of fine forceps were used to puncture a hole about 4 mm across in the lid of a 1.5 ml microfuge tube and twisted to make it round. The loaded pipette tip should fit into this hole as far as the second bevel, with the tip suspended about 5 mm from the bottom of the tube. 40 µl of mobile phase (10% Acetonitrile, 0.1% Trifluoroacetic acid (TFA)) was added into the top part of the tip and spun in a micro-centrifuge for 1 minute at 4000 rpm (1100 rcf).

Altering the [252] method for PEG containing samples: Dissolved the sample in 10 µl of 20 mg $ml^{-1}$ 2,5-Dihydroxybenzoic acid (DHB) with 0.1% TFA. Wetted the tip in 10% Acetonitrile, 0.1% TFA. Loaded the sample onto titanium dioxide loaded pipette




**(a)** Chromatogram

**(b)** Mass to charge

**(c)** Deconvoluted mass , near 25 124 Da peak

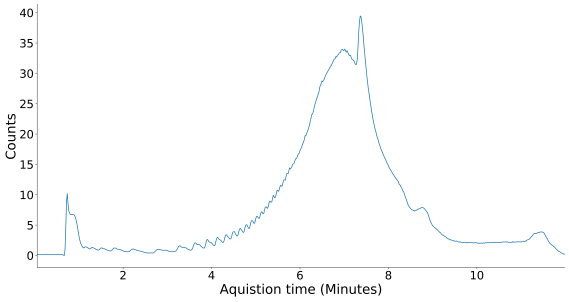
**(d)** Deconvoluted mass

**Figure E.1:** Adding 0.5% PEG 10 kDa to NUDT7 reduced signal to noise ratio of MS. Unlabelled 1 mg ml$^{-1}$ NUDT7 with 0.5% PEG 10 kDa

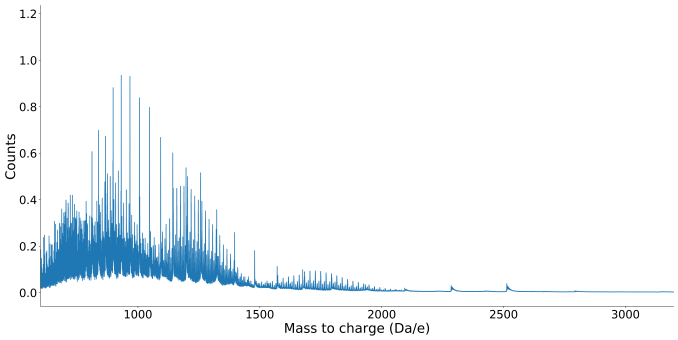
tip. Washed by micro centrifuge at 1,100 rcf for 1 minute, with 40 µl of 20 mg ml$^{-1}$ DHB, 10% Acetonitrile, 0.1% TFA: 3 times. Washed by micro centrifuge at 1,100 rcf for 1 minute, 10% Acetonitrile, 0.1% TFA: 7 times. Eluted into collection eppendorf with 40 µl of 80% Acetonitrile, 0.1% TFA. Speed vacuumed till dry. Reconstituted in mass spec buffer: 30% methanol 0.1% formic acid.

## E.2 Signal to noise ratio

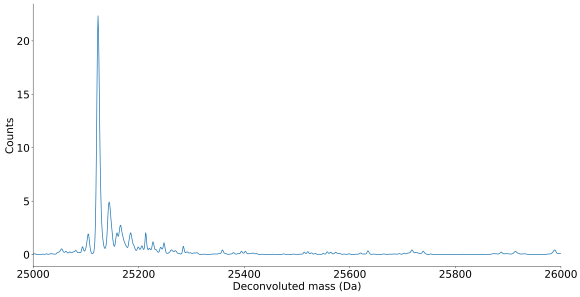
Additional figures showing the signal to noise ratio in NUDT7 MS samples with PEG added in figures E.1 and E.2.



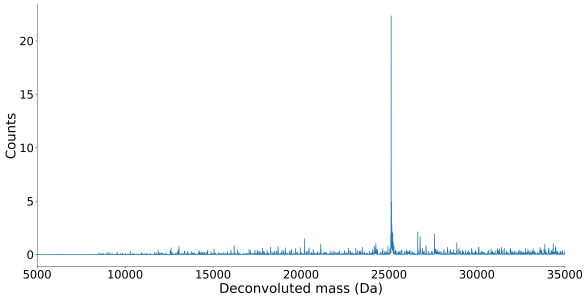

**(a)** Chromatogram of 1 mg ml$^{-1}$ NUDT7A with 5% PEG 10 kDa




**(b)** Mass to charge




**(c)** Deconvoluted mass, near 25 124 Da peak




**(d)** Deconvolution

**Figure E.2:** Adding 5% PEG 10 kDa to NUDT7 reduced signal to noise ratio of MS.Unlabelled 1 mg ml$^{-1}$ NUDT7 with 5% PEG 10 kDa

# References


[1] Nicholas M Pearce et al. “A Multi-Crystal Method for Extracting Obscured Signal from Crystallographic Electron Density”. In: *Nature Communications* (2016).

[2] Fei Long et al. “AceDRG: A stereochemical description generator for ligands”. In: *Acta Crystallographica Section D: Structural Biology* 73.2 (2017), pp. 112–122.

[3] P Emsley et al. “Features and development of Coot”. In: *Acta Crystallographica Section D: Biological Crystallography* 66.4 (2010), pp. 486–501.

[4] O S Smart et al. *grade, version 1.2.13*. http://www.globalphasing.com. 2011.

[5] Nigel W. Moriarty, Ralf W. Grosse-Kunstleve, and Paul D. Adams. “Electronic ligand builder and optimization workbench (eLBOW): A tool for ligand coordinate and restraint generation”. In: *Acta Crystallographica Section D: Biological Crystallography* 65.10 (2009), pp. 1074–1080.

[6] Marcin Wojdyr et al. “DIMPLE - a pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands.” In: *Acta Crystallographica Section A* 69 (2013), s299.

[7] E. Blanc et al. “Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT”. In: *Acta Crystallographica Section D: Biological Crystallography* 60.12 I (2004), pp. 2210–2221.

[8] Garib N. Murshudov et al. “REFMAC5 for the refinement of macromolecular crystal structures”. In: *Acta Crystallographica Section D: Biological Crystallography* (2011).

[9] Peter E. Wright and H. Jane Dyson. “Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm”. In: *Journal of Molecular Biology* 293.2 (1999), pp. 321–331.

[10] Christopher N. Johnson et al. “Fragment-to-Lead Medicinal Chemistry Publications in 2016”. In: *Journal of Medicinal Chemistry* 61.5 (2018), pp. 1774–1784.

[11] Richard J. Hall, Christopher W. Murray, and Marcel L. Verdonk. “The Fragment Network: A Chemistry Recommendation Engine Built Using a Graph Database”. In: *Journal of Medicinal Chemistry* 60.14 (2017), pp. 6440–6450.

[12] Shipra Malhotra and John Karanicolas. “When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode?” In: *Journal of Medicinal Chemistry* 60.1 (2017), pp. 128–145.

[13] Thomas D. Pollard. “MBOC technical perspective: A guide to simple and informative binding assays”. In: *Molecular Biology of the Cell* 21.23 (2010), pp. 4061–4067.

[14] Iva Navaratilova et al. “Fragment screening by surface plasmon resonance”. In: *ACS Medicinal Chemistry Letters* 1.1 (2010), pp. 44–48.

[15] Ernesto Freire, Obdulio L. Mayorga, and Martin Straume. “Isothermal Titration”. In: *Analytical Chemistry* 62.18 (1990), 950A–959A.

[16] Moran Jerabek-Willemsen et al. “Molecular Interaction Studies Using Microscale Thermophoresis”. In: *ASSAY and Drug Development Technologies* 9.4 (2011), pp. 342–353. URL: http://www.liebertonline.com/doi/abs/10.1089/adt.2011.0380.

[17] Suzanne B. Shuker et al. “Discovering high-affinity ligands for proteins: SAR by NMR”. In: *Science* 274.5292 (1996), pp. 1531–1534.

[18] Richard R Neubig et al. “Update on Terms and Symbols in Quantitative Pharmacology”. In: *Pharmacological Reviews* 55.4 (2003), pp. 597–606.

[19] Tim Bugg. *Introduction to Enzyme and Coenzyme Chemistry: Second Edition*. 1997, pp. 1–292.

[20] W. W. Cleland. “The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations”. In: *BBA - Biochimica et Biophysica Acta* 67.C (1963), pp. 104–137.

[21] Avick Kumar Ghosh et al. “Covalent Inhibition in Drug Discovery”. In: *ChemMedChem* 14.9 (2019), pp. 889–906.

[22] Naruo Yoshimura et al. “EKB-569, a new irreversible epidermal growth factor receptor tyrosine kinase inhibitor, with clinical activity in patients with non-small cell lung cancer with acquired resistance to gefitinib”. In: *Lung Cancer* 51.3 (2006), pp. 363–368.

[23] Jeffrey A. Engelman et al. “PF00299804, an irreversible pan-ERBB inhibitor, is effective in lung cancer models with EGFR and ERBB2 mutations that are resistant to gefitinib”. In: *Cancer Research* 67.24 (2007), pp. 11924–11932.

[24] Ken A. Brameld et al. “Discovery of the Irreversible Covalent FGFR Inhibitor 8-(3-(4-Acryloylpiperazin-1-yl)propyl)-6-(2,6-dichloro-3,5-dimethoxyphenyl)-2-(methylamino)pyrido[2,3-d]pyrimidin-7(8H)-one (PRN1371) for the Treatment of Solid Tumors”. In: *Journal of Medicinal Chemistry* 60.15 (2017), pp. 6516–6527.

[25] Jacob Lauwring Andersen et al. “Dimethyl fumarate is an allosteric covalent inhibitor of the p90 ribosomal S6 kinases”. In: *Nature Communications* 9.1 (2018), pp. 1–8. URL: http://dx.doi.org/10.1038/s41467-018-06787-w.

[26] Joseph Grimsby et al. “Allosteric activators of glucokinase: Potential role in diabetes therapy”. In: *Science* 301.5631 (2003), pp. 370–373.

[27] Serge Mignani et al. “Why and how have drug discovery strategies in pharma changed? What are the new mindsets?” In: *Drug Discovery Today* 21.2 (2016), pp. 239–249.

[28] S. A. Waldman and A. Terzic. “Systems-based discovery advances drug development”. In: *Clinical Pharmacology and Therapeutics* 93.4 (2013), pp. 285–287.

[29] John G. Moffat et al. “Opportunities and challenges in phenotypic drug discovery: An industry perspective”. In: *Nature Reviews Drug Discovery* 16.8 (2017), pp. 531–543. URL: http://dx.doi.org/10.1038/nrd.2017.111.

[30] Jörg Eder, Richard Sedrani, and Christian Wiesmann. “The discovery of first-in-class drugs: Origins and evolution”. In: *Nature Reviews Drug Discovery* (2014).

[31] Ricardo MacArron et al. “Impact of high-throughput screening in biomedical research”. In: *Nature Reviews Drug Discovery* 10.3 (2011), pp. 188–195. URL: http://dx.doi.org/10.1038/nrd3368.

[32] Robert J. Young and Paul D. Leeson. “Mapping the Efficiency and Physicochemical Trajectories of Successful Optimizations”. In: *Journal of Medicinal Chemistry* 61.15 (2018), pp. 6421–6467.

[33] Paul D. Leeson and Brian Springthorpe. “The influence of drug-like concepts on decision-making in medicinal chemistry”. In: *Nature Reviews Drug Discovery* 6.11 (2007), pp. 881–890.

[34] Hongyu Zhao et al. “Discovery of Potent, Highly Selective, and Orally Bioavailable Pyridine Carboxamide c-Jun NH 2-Terminal Kinase Inhibitors”. In: (2006).

[35] Christopher W Murray, David R Newell, and Patrick Angibaud. *A successful collaboration between academia, biotech and pharma led to discovery of erdafitinib, a selective FGFR inhibitor recently approved by the FDA*. 2019.

[36] William P Jencks. *On the attribution and additivity of binding energies*. Tech. rep. 7. 1981, pp. 4046–4050.

[37] Suhman Chung et al. “Impact of linker strain and flexibility in the design of a fragment-based inhibitor”. In: *Nature Chemical Biology* 5.6 (2009), pp. 407–413.

[38] Andrew J Whitehouse et al. “Development of Inhibitors against Mycobacterium abscessus tRNA (m 1 G37) Methyltransferase (TrmD) Using Fragment-Based Approaches”. In: (2019).

[39] Daniel A. Erlanson. *Fragment-based lead discovery: a chemical update*. 2006.

[40] H L Silvestre et al. “Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery”. In: *Proceedings of the National Academy of Sciences* 110.32 (2013), pp. 12984–12989. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1304045110.

[41] Rupp Bernhard. *Biomolecular crystallography: principles, practice and applications to structural biology*. 2010.

[42] Jin Liu and Ruth Nussinov. “Allostery: An Overview of Its History, Concepts, Methods, and Applications”. In: *PLOS Computational Biology* 12.6 (2016), e1004966. URL: http://dx.plos.org/10.1371/journal.pcbi.1004966.

[43] Alexandr P. Kornev and Susan S. Taylor. “Dynamics-Driven Allostery in Protein Kinases”. In: *Trends in Biochemical Sciences* 40.11 (2015), pp. 628–647. URL: http://dx.doi.org/10.1016/j.tibs.2015.09.002.

[44] Steve W Lockless et al. “Evolutionarily conserved networks of residues mediate allosteric communication in proteins”. In: *Nature structural biology* 10.1 (2003), pp. 59–69. URL: http://www.ncbi.nlm.nih.gov/pubmed/12483203.

[45] E Benedetti. “Statistical and energetic analysis of side-chain conformations in oligopeptides”. In: *… journal of peptide …* 22 (1983), pp. 1–15. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1399-3011.1983.tb02062.x/abstract.

[46] Hesam N Motlagh et al. *The ensemble nature of allostery*. 2014.

[47] Supriyo Bhattacharya et al. “Conserved Mechanism of Conformational Stability and Dynamics in G-Protein-Coupled Receptors”. In: *Journal of Chemical Theory and Computation* 12.11 (2016), pp. 5575–5584.

[48] Dimitar V. Pachov et al. “Coupled Motions in $\beta$2AR:G$\alpha$s Conformational Ensembles”. In: *Journal of Chemical Theory and Computation* 12.3 (2016), pp. 946–956.

[49] Julian Wölcke and Dirk Ullmann. “Miniaturized HTS technologies - uHTS”. In: *Drug Discovery Today* 6.12 (2001), pp. 637–646.

[50] James Inglese et al. “High-throughput screening assays for the identification of chemical probes”. In: *Nature Chemical Biology* 3.8 (2007), pp. 466–479.

[51] Octerloney B. McDonald et al. “A scintillation proximity assay for the raf/MEK/ERK kinase cascade: High-throughput screening and identification of selective enzyme inhibitors”. In: *Analytical Biochemistry* 268.2 (1999), pp. 318–329.

[52] Rati Verma et al. “Ubistatins inhibit proteasome-dependent degradation by binding the ubiquitin chain”. In: *Science* 306.5693 (2004), pp. 117–120.

[53] Marc Ferrer et al. “A Fully Automated [35S]GTP$\gamma$S Scintillation Proximity Assay for the High-Throughput Screening of Gi-Linked G Protein-Coupled Receptors”. In: *Assay and Drug Development Technologies* 1.2 (Apr. 2003), pp. 261–273. URL: http://www.liebertpub.com/doi/10.1089/15406580360545071.

[54] W. Frank An and Nicola Tolliday. “Cell-based assays for high-throughput screening”. In: *Molecular Biotechnology* 45.2 (2010), pp. 180–186.

[55] Marco F. Schmidt and Jörg Rademann. “Dynamic template-assisted strategies in fragment-based drug discovery”. In: *Trends in Biotechnology* 27.9 (2009), pp. 512–521.

[56] Helen Plant et al. “High-throughput hit screening cascade to identify respiratory syncytial virus (RSV) inhibitors”. In: *Journal of Biomolecular Screening* 20.5 (2015), pp. 597–605.

[57] Regine S Bohacek, Colin McMartin, and Wayne C Guida. “The art and practice of structure-based drug design: A molecular modeling perspective”. In: *Medicinal Research Reviews* 16.1 (1996), pp. 3–50.

[58] Andy Barker et al. “Expanding medicinal chemistry space”. In: *Drug Discovery Today* 18.5-6 (2013), pp. 298–304. URL: http://dx.doi.org/10.1016/j.drudis.2012.10.008.

[59] James T Metz, Jeffrey R Huth, and Philip J Hajduk. “Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups”. In: *Journal of Computer-Aided Molecular Design* 21.1-3 (2007), pp. 139–144.

[60] R. Sink et al. “False Positives in the Early Stages of Drug Discovery”. In: *Current Medicinal Chemistry* 17.34 (Nov. 2010), pp. 4231–4255.

[61] Susan L. McGovern et al. “A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening”. In: *Journal of Medicinal Chemistry* 45.8 (2002), pp. 1712–1722.

[62] Michael M. Hann, Andrew R. Leach, and Gavin Harper. “Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery”. In: *Journal of Chemical Information and Computer Sciences* 41.3 (2001), pp. 856–864.

[63] Andrew R Leach and Michael M Hann. “Molecular complexity and fragment-based drug discovery: Ten years on”. In: *Current Opinion in Chemical Biology* 15.4 (2011), pp. 489–496. URL: http://dx.doi.org/10.1016/j.cbpa.2011.05.008.

[64] Miles Congreve et al. “A ’Rule of Three’ for fragment-based lead discovery?” In: *Drug Discovery Today* 8.19 (2003), pp. 876–877.

[65] Ulrich Schieborr et al. “How much NMR data is required to determine a protein-ligand complex structure?” In: *ChemBioChem* 6.10 (2005), pp. 1891–1898.

[66] B Shuker, S et al. “Discovering high-affinity ligands for proteins: SAR by NMR”. In: *Science* 274.November (1996), pp. 1531–1534. URL: http://www.ncbi.nlm.nih.gov/pubmed/18051441.

[67] Paul N Mortenson et al. “Fragment-to-Lead Medicinal Chemistry Publications in 2017”. In: (2018).

[68] Moriz Mayer and Bernd Meyer. “Characterization of ligand binding by saturation transfer difference NMR spectroscopy”. In: *Angewandte Chemie - International Edition* 38.12 (1999), pp. 1784–1788.

[69] T. Scherf and J. Anglister. “A T1 rho-filtered two-dimensional transferred NOE spectrum for studying antibody interactions with peptide antigens”. In: *Biophysical Journal* 64.3 (1993), pp. 754–761. URL: http://dx.doi.org/10.1016/S0006-3495(93)81436-X.

[70] Christian Ludwig and Ulrich L. Guenther. “Ligand based NMR methods for drug discovery”. In: *Frontiers in Bioscience* 14.12 (2009), pp. 4565–4574.

[71] Claudio Dalvit et al. “Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water”. In: *Journal of Biomolecular NMR* 18.1 (2000), pp. 65–68.

[72] Claudio Dalvit et al. “Fluorine-NMR experiments for high-throughput screening: Theoretical aspects, practical considerations, and range of applicability”. In: *Journal of the American Chemical Society* 125.25 (2003), pp. 7696–7703.

[73] Jia Gao et al. “Automated NMR fragment based screening identified a novel interface blocker to the LARG/RhoA complex”. In: *PLoS ONE* 9.2 (2014).

[74] Olivier Cala and Isabelle Krimm. “Ligand-orientation based fragment selection in STD NMR screening”. In: *Journal of Medicinal Chemistry* 58.21 (2015), pp. 8739–8742.

[75] Toshihiko Sugiki et al. “Current NMR techniques for structure-based drug discovery”. In: *Molecules* 23.1 (2018), pp. 1–27.

[76] Andreas Boettcher et al. “Fragment-based screening by biochemical assays: Systematic feasibility studies with trypsin and MMP12”. In: *Journal of Biomolecular Screening* 15.9 (2010), pp. 1029–1041.

[77] D. E. Koshland. “Application of a Theory of Enzyme Specificity to Protein Synthesis”. In: *Proceedings of the National Academy of Sciences* 44.2 (1958), pp. 98–104.

[78] G. V. Semisotnov et al. “Study of the “molten globule” intermediate state in protein folding by a hydrophobic fluorescent probe”. In: *Biopolymers* 31.1 (1991), pp. 119–128.

[79] Patrick Garidel et al. “A rapid, sensitive and economical assessment of monoclonal antibody conformational stability by intrinsic tryptophan fluorescence spectroscopy”. In: *Biotechnology Journal* 3.9-10 (2008), pp. 1201–1211.

[80] Morgane J.J. Moreau et al. “Rapid determination of protein stability and ligand binding by differential scanning fluorimetry of GFP-tagged proteins”. In: *RSC Advances* 2.31 (2012), pp. 11892–11900.

[81] Daniel A. Erlanson et al. “Twenty years on: The impact of fragments on drug discovery”. In: *Nature Reviews Drug Discovery* 15.9 (2016), pp. 605–619. URL: http://dx.doi.org/10.1038/nrd.2016.109.

[82] Ben J Davis and Daniel A Erlanson. “Learning from our mistakes: The ’unknown knowns’ in fragment screening”. In: *Bioorganic and Medicinal Chemistry Letters* 23.10 (2013), pp. 2844–2852. URL: http://dx.doi.org/10.1016/j.bmcl.2013.03.028.

[83] Anthony M. Giannetti. *From experimental design to validated hits: A comprehensive walk-through of fragment lead identification using surface plasmon resonance*. 1st ed. Vol. 493. Elsevier Inc., 2011, pp. 169–218. URL: http://dx.doi.org/10.1016/B978-0-12-381274-2.00008-X.

[84] Iva Navratilova and David G. Myszka. “Investigating Biomolecular Interactions and Binding Properties Using SPR Biosensors”. In: 2006, pp. 155–176.

[85] Minh-dao Dao Duong-Thi et al. “Weak affinity chromatography as a new approach for fragment screening in drug discovery”. In: *Analytical Biochemistry* 414.1 (2011), pp. 138–146. URL: http://dx.doi.org/10.1016/j.ab.2011.02.022.

[86] Minh Dao Duong-Thi et al. “Comparison of weak affinity chromatography and surface plasmon resonance in determining affinity of small molecules”. In: *Analytical Biochemistry* 461 (2014), pp. 57–59. URL: http://dx.doi.org/10.1016/j.ab.2014.05.023.

[87] John E. Ladbury, Gerhard Klebe, and Ernesto Freire. *Adding calorimetric data to decision making in lead discovery: A hot tip*. Jan. 2010.

[88] Philippe Dumas et al. “Extending ITC to Kinetics with kinITC”. In: *Methods in Enzymology*. Vol. 567. Academic Press Inc., 2016, pp. 157–180.

[89] Joel Tellinghuisen and John D Chodera. “Systematic errors in isothermal titration calorimetry : Concentrations and baselines”. In: *Analytical Biochemistry* 414.2 (2011), pp. 297–299. URL: http://dx.doi.org/10.1016/j.ab.2011.03.024.

[90] Moran Jerabek-Willemsen et al. “MicroScale Thermophoresis: Interaction analysis and beyond”. In: *Journal of Molecular Structure* 1077 (2014), pp. 101–113. URL: http://dx.doi.org/10.1016/j.molstruc.2014.03.009.

[91] Renato A. Bauer. “Covalent inhibitors in drug discovery: From accidental discoveries to avoided liabilities and designed therapies”. In: *Drug Discovery Today* 20.9 (2015), pp. 1061–1073. URL: http://dx.doi.org/10.1016/j.drudis.2015.05.005.

[92] Christian Jöst et al. “Promiscuity and selectivity in covalent enzyme inhibition: A systematic study of electrophilic fragments”. In: *Journal of Medicinal Chemistry* 57.18 (2014), pp. 7590–7599.

[93] Jonathan M. Ostrem et al. “K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions”. In: *Nature* 503.7477 (Nov. 2013), pp. 548–551. URL: http://www.nature.com/articles/nature12796.

[94] Efrat Resnick et al. “Rapid covalent-probe discovery by electrophile fragment screening”. In: (2018).

[95] Scott B. Ficarro et al. “Leveraging gas-phase fragmentation pathways for improved identification and selective detection of targets modified by covalent probes”. In: *Analytical Chemistry* 88.24 (2016), pp. 12248–12254.

[96] John B. Fenn et al. *Electrospray ionization for mass spectrometry of large biomolecules*. 1989.

[97] Steven A. Hofstadler and Kristin A. Sannes-Lowery. *Applications of ESI-MS in drug discovery: Interrogation of noncovalent complexes*. July 2006.

[98] Liliana Pedro and Ronald J. Quinn. “Native mass spectrometry in fragment-based drug discovery”. In: *Molecules* 21.8 (2016).

[99] Federico Riccardi Sirtori et al. “Establish an automated flow injection ESI-MS method for the screening of fragment based libraries: Application to Hsp90”. In: *European Journal of Pharmaceutical Sciences* (2015).

[100] Kurt Hilpert et al. “Design and Synthesis of Potent and Highly Selective Thrombin Inhibitors”. In: *Journal of Medicinal Chemistry* 37.23 (1994), pp. 3889–3901.

[101] Christophe L M J Verlinde, Gabrielle Rudenko, and Wim G J Hol. “In search of new lead compounds for trypanosomiasis drug design: A protein structure-based linked-fragment approach”. In: *Journal of Computer-Aided Molecular Design* 6.2 (1992), pp. 131–147.

[102] B Shuker, S et al. “Discovering high-affinity ligands for proteins: SAR by NMR”. In: *Science* 274.November (1996), pp. 1531–4. URL: http://www.ncbi.nlm.nih.gov/pubmed/18051441.

[103] Ellene H Mashalidis, Steffen Śledå Pawełand Lang, and Chris Abell. “A three-stage biophysical screening cascade for fragment-based drug discovery”. In: *Nature Protocols* 8.11 (2013), pp. 2309–2324.

[104] Johannes Schiebel et al. “Six Biophysical Screening Methods Miss a Large Proportion of Crystallographically Discovered Fragment Hits: A Case Study”. In: *ACS Chem Biol* 11 (2016), pp. 1693–1701.

[105] Jerome Wielens et al. “Parallel screening of low molecular weight fragment libraries: Do differences in methodology affect hit identification?” In: *Journal of Biomolecular Screening* 18.2 (2013), pp. 147–159.

[106] *Maybridge Ro3 Diversity Fragment Library*. URL: https://www.maybridge.com/portal/alias__Rainbow/lang__en/tabID__230/DesktopDefault.aspx.

[107] J D O’Hea et al. “BART: Development of a Sample Exchange System for MX Beamlines”. In: *International Conference on Accelerator and Large Experimental Control Systems* Geneva, Switzerland (2018), pp. 1919–1922. URL: http://jacow.org/icalepcs2017/papers/thpha200.pdf.

[108] Patrick M Collins et al. “Gentle, fast and effective crystal soaking by acoustic dispensing”. In: (2016), pp. 246–255.

[109] Graeme Winter and Katherine E. McAuley. “Automated data collection for macromolecular crystallography”. In: *Methods* 55.1 (2011), pp. 81–93. URL: http://dx.doi.org/10.1016/j.ymeth.2011.06.010.

[110] Graeme Winter. “xia2: an expert system for macromolecular crystallography data reduction”. In: *Journal of Applied Crystallography* 43 (2009), pp. 186–190.

[111] Patrick M. Collins et al. “Achieving a Good Crystal System for Crystallographic X-ray Fragment Screening”. In: (Sept. 2018). URL: http://www.preprints.org/manuscript/201809.0383/v1.

[112] Disha Patel, Joseph D. Bauman, and Eddy Arnold. “Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery”. In: *Progress in Biophysics and Molecular Biology* 116.2-3 (2014), pp. 92–100. URL: http://dx.doi.org/10.1016/j.pbiomolbio.2014.08.004.

[113] Vicki L. Nienaber et al. “Discovering novel ligands for macromolecules using X-ray crystallographic screening”. In: *Nature Biotechnology* 18.10 SUPPL. (2000), pp. 1105–1108.

[114] Jia Tsing Ng et al. “Using textons to rank crystallization droplets by the likely presence of crystals”. In: *Acta Crystallographica Section D: Biological Crystallography* 70.10 (2014), pp. 2702–2718.

[115] Tobias Krojer et al. “The XChemExplorer graphical workflow tool for routine or large-scale protein-ligand structure determination”. In: *Acta Crystallographica Section D: Structural Biology* 73 (2017), pp. 267–278.

[116] J. W. Pflugrath. “Practical macromolecular cryocrystallography”. In: *Acta Crystallographica Section F:Structural Biology Communications* 71 (2015), pp. 622–642.

[117] Helen M. Berman et al. “The Protein Data Bank”. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235.

[118] Helen Berman, Kim Henrick, and Haruki Nakamura. “Announcing the worldwide Protein Data Bank”. In: *Nature Structural Biology* 10.12 (2003), pp. 980–980. URL: http://www.nature.com/doifinder/10.1038/nsb1203-980.

[119] “Crystallography: Protein data bank”. In: *Nature New Biology* 233.42 (Oct. 1971), pp. 223–223.

[120] Airlie J. McCoy et al. “Phaser crystallographic software”. In: *Journal of Applied Crystallography* 40.4 (2007), pp. 658–674.

[121] A. W. Hung et al. “Route to three-dimensional fragments using diversity-oriented synthesis”. In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 6799–6804. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1015271108.

[122] Sarah L. Kidd et al. “Recent Applications of Diversity-Oriented Synthesis Toward Novel, 3-Dimensional Fragment Collections”. In: *Frontiers in Chemistry* 6.October (2018), p. 460. URL: https://www.frontiersin.org/article/10.3389/fchem.2018.00460/full.

[123] Oakley B Cox and Et Al. “A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain”. In: *Chem. Sci.* 7.3 (2016), pp. 2322–2330. URL: http://xlink.rsc.org/?DOI=C5SC03115J.

[124] Gyorgy M Keseru et al. “Design principles for fragment libraries – Maximizing the value of learnings from Pharma fragment based drug discovery (FBDD) programs for use in academia”. In: *Journal of Medicinal Chemistry* (2016), acs.jmedchem.6b00197. URL: http://pubs.acs.org/doi/abs/10.1021/acs.jmedchem.6b00197.

[125] Bradley C. Doak et al. “Design and evaluation of the performance of an NMR screening fragment library”. In: *Australian Journal of Chemistry* 66.12 (2013), pp. 1465–1472.

[126] James Seidler et al. “Identification and prediction of promiscuous aggregating inhibitors among known drugs”. In: *Journal of Medicinal Chemistry* 46.21 (2003), pp. 4477–4486.

[127] Brian Y. Feng et al. “High-throughput assays for promiscuous inhibitors”. In: *Nature Chemical Biology* 1.3 (2005), pp. 146–148.

[128] Jonathan B Baell and Georgina A Holloway. “New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays”. In: *Journal of Medicinal Chemistry* 53.7 (2010), pp. 2719–2740.

[129] Stephen D. Roughley and Allan M. Jordan. “The medicinal chemist’s toolbox: An analysis of reactions used in the pursuit of drug candidates”. In: *Journal of Medicinal Chemistry* 54.10 (2011), pp. 3451–3479.

[130] Markus Hartenfeller et al. “A collection of robust organic synthesis reactions for in silico molecule design”. In: *Journal of Chemical Information and Modeling* 51.12 (2011), pp. 3093–3098.

[131] John Spencer et al. “Microwave-mediated synthesis and manipulation of a 2-substituted-5- aminooxazole-4-carbonitrile library”. In: *Tetrahedron Letters* 53.13 (2012), pp. 1656–1659. URL: http://dx.doi.org/10.1016/j.tetlet.2012.01.081.

[132] Andreas C. Joerger et al. “Exploiting Transient Protein States for the Design of Small-Molecule Stabilizers of Mutant p53”. In: *Structure* 23.12 (2015), pp. 2246–2255. URL: http://dx.doi.org/10.1016/j.str.2015.10.016.

[133] Marc O’Reilly et al. “Crystallographic screening using ultra-low-molecular-weight ligands to guide drug design”. In: *Drug Discovery Today* 24.5 (2019), pp. 1081–1086. URL: https://doi.org/10.1016/j.drudis.2019.03.009.

[134] Daniel J Wood et al. “FragLites - Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation”. In: *Journal of Medicinal Chemistry* 62.7 (2019), pp. 3741–3752.

[135] B. W. Matthews. “Solvent content of protein crystals”. In: *Journal of Molecular Biology* 33.2 (1968), pp. 491–497.

[136] B W Matthews. “X-ray Crystallographic Studies of Proteins”. In: *Annual Review of Physical Chemistry* 27.1 (1976), pp. 493–493.

[137] Christian X. Weichenberger and Bernhard Rupp. “Ten years of probabilistic estimates of biocrystal solvent content: New insights via nonparametric kernel density estimate”. In: *Acta Crystallographica Section D: Biological Crystallography* 70.6 (2014), pp. 1579–1588.

[138] Jark Bttcher et al. “Key factors for successful generation of proteinfragment structures: Requirement on protein, crystals, and technology”. In: *Methods in Enzymology* (2011).

[139] Frederik Rainer Ehrmann et al. “Soaking suggests ”alternative facts”: Only cocrystallization discloses major ligand-induced interface rearrangements of a homodimeric tRNA-binding protein indicating a novel mode-of-inhibition”. In: *PLoS ONE* 12.4 (2017), pp. 1–30.

[140] Bhushan Nagar et al. “Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571)”. In: *Cancer Research* 62.15 (2002), pp. 4236–4243.

[141] Disha Patel, Joseph D. Bauman, and Eddy Arnold. “Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery”. In: *Progress in Biophysics and Molecular Biology* 116.2-3 (2014), pp. 92–100. URL: http://dx.doi.org/10.1016/j.pbiomolbio.2014.08.004.

[142] Sergio Ruiz-Carmona et al. “Dynamic undocking and the quasi-bound state as tools for drug discovery”. In: *Nature Chemistry* 9.3 (2017), pp. 201–206. URL: http://dx.doi.org/10.1038/nchem.2660.

[143] Susan Leung et al. “SuCOS is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses”. In: *ChemRxiv* (2019), pp. 1–47. URL: https://chemrxiv.org/articles/SuCOS%7B%5C_%7Dis%7B%5C_%7DBetter%7B%5C_%7Dthan%7B%5C_%7DRMSD%7B%5C_%7Dfor%7B%5C_%7DEvaluating%7B%5C_%7DFragment%7B%5C_%7DElaboration%7B%5C_%7Dand%7B%5C_%7DDocking%7B%5C_%7DPoses/8100203.

[144] Alexander McPherson and Jose A Gavira. “Introduction to protein crystallization”. In: *Acta Crystallographica Section F* 70.1 (2014), pp. 2–20.

[145] Jia Tsing Ng et al. “Lessons from ten years of crystallization experiments at the SGC”. In: *Acta Crystallographica Section D: Structural Biology* 72.2 (2016), pp. 224–235.

[146] Guido A. Malawski et al. “Identifying protein construct variants with increased crystallization propensity–A case study”. In: *Protein Science* 15.12 (2006), pp. 2718–2728.

[147] Marc C. Deller and Bernhard Rupp. “Models of protein-ligand crystal structures: Trust, but verify”. In: *Journal of Computer-Aided Molecular Design* 29.9 (2015), pp. 817–836.

[148] Swagata Dasgupta et al. “Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers”. In: *Proteins: Structure, Function and Genetics* 28.4 (1997), pp. 494–514.

[149] Alexandra Cousido-Siah et al. “Crystal packing modifies ligand binding affinity: The case of aldose reductase”. In: *Proteins: Structure, Function and Bioinformatics* 80.11 (2012), pp. 2552–2561.

[150] Elspeth F Garman and Robin Leslie Owen. “Cryocooling and radiation damage in macromolecular crystallography”. In: *Acta Crystallographica Section D: Biological Crystallography* 62.1 (2006), pp. 32–47.

[151] Daniel A Keedy et al. “Crystal cryocooling distorts conformational heterogeneity in a model michaelis complex of DHFR”. In: *Structure* 22.6 (2014), pp. 899–910. URL: http://dx.doi.org/10.1016/j.str.2014.04.016.

[152] Daniel A. Keedy et al. “An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering”. In: *eLife* 7 (2018), pp. 1–36.

[153] Marcus Fischer, Brian K Shoichet, and James S Fraser. “One Crystal, Two Temperatures: Cryocooling Penalties Alter Ligand Binding to Transient Protein Sites.” In: *Chembiochem : a European journal of chemical biology* 16.11 (2015), pp. 1560–1564.

[154] Matthew Warkentin et al. “Global radiation damage: Temperature dependence, time dependence and how to outrun it”. In: *Journal of Synchrotron Radiation* (2013).

[155] M Weik et al. “Specific chemical and structural damage to proteins produced by synchrotron radiation”. In: *Proceedings of the National Academy of Sciences* 97.2 (2000), pp. 623–628. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.97.2.623.

[156] Raimond B G Ravelli and Sean M. McSweeney. “The ’fingerprint’ that X-rays can leave on structures”. In: *Structure* 8.3 (2000), pp. 315–328.

[157] Emanuela Fioravanti et al. “Specific radiation damage to acidic residues and its relation to their chemical and structural environment”. In: *Journal of Synchrotron Radiation* 14.1 (2007), pp. 84–91.

[158] Jessica L. Thomaston et al. “XFEL structures of the influenza M2 proton channel: Room temperature water networks and insights into proton conduction”. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.51 (2017), pp. 13357–13362.

[159] Masayoshi Nakasako et al. “Water-protein interactions from high-resolution protein crystallography”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 359.1448 (2004), pp. 1191–1206.

[160] Stephen K Burley et al. *Protein Crystallography*. Vol. 1607. 2017, pp. 627–641. URL: http://link.springer.com/10.1007/978-1-4939-7000-1.

[161] T. Geoff G. Battye et al. “iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM”. In: *Acta Crystallographica Section D: Biological Crystallography* 67.4 (2011), pp. 271–281.

[162] S. Fili et al. “Human insulin polymorphism upon ligand binding and pH variation: The case of 4-ethylresorcinol”. In: *IUCrJ* 2 (2015), pp. 534–544.

[163] J. W. Pflugrath. “The finer things in X-ray diffraction data collection”. In: *Acta Crystallographica Section D: Biological Crystallography* 55.10 (1999), pp. 1718–1725.

[164] Harold R. Powell. “X-ray data processing”. In: *Bioscience Reports* 37.5 (2017), pp. 1–14.

[165] Ethan A Merritt. “To B or not to B: A question of resolution?” In: *Acta Crystallographica Section D: Biological Crystallography* 68.4 (2012), pp. 468–477.

[166] Raimond B G Ravelli and Sean M McSweeney. “The ’fingerprint’ that X-rays can leave on structures”. In: *Structure* 8.3 (2000), pp. 315–328.

[167] Alexander Wlodawer et al. “Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination”. In: *FEBS Journal* 280.22 (2013), pp. 5705–5736.

[168] Verner Schomaker and K.N. Trueblood. “On the rigid-body motion of molecules in crystals”. In: *Acta Cryst* (1968), pp. 63–76.

[169] Jay Painter and Ethan A Merritt. “Optimal description of a protein structure in terms of multiple groups undergoing TLS motion research papers”. In: *Acta Crystallographica - Section D Biological Crystallography* 62 (2006), pp. 439–450.

[170] A. J.C. Wilson. *Determination of absolute from relative X-ray intensity data [3]*. 1942.

[171] Garry L. Taylor. “Introduction to phasing”. In: *Acta Crystallographica Section D: Biological Crystallography* 66.4 (2010), pp. 325–338.

[172] George M. Sheldrick. “A short history of SHELX”. In: *Acta Crystallographica Section A: Foundations of Crystallography* 64.1 (2007), pp. 112–122.

[173] M. G. Rossmann and D. M. Blow. “The detection of sub-units within the crystallographic asymmetric unit”. In: *Acta Crystallographica* 15.1 (1962), pp. 24–31.

[174] Bi Cheng Wang. “Resolution of phase ambiguity in macromolecular crystallography”. In: *Methods in Enzymology* 115.C (1985), pp. 90–112.

[175] Wayne A Hendrickson. “Determination of Macromolecular Determination Structures from Anomalous Diffraction of Synchrotron Radiation”. In: *Science* 254.5028 (1991), pp. 51–58.

[176] G. M. Sheldrick. “Phase annealing in SHELX-90: direct methods for larger structures”. In: *Acta Crystallographica Section A* 46.6 (1990), pp. 467–473.

[177] Richard J. Morris and Gérard Bricogne. “Sheldrick’s 1.2 Å rule and beyond”. In: *Acta Crystallographica - Section D Biological Crystallography* 59.3 (2003), pp. 615–617.

[178] A. L. Patterson. “A Direct Method for the Determination of the Components of Interatomic Distances in Crystals”. In: *Zeitschrift für Kristallographie - Crystalline Materials* 90.1-6 (Jan. 1935), pp. 517–542.

[179] Randy J. Read. “Pushing the boundaries of molecular replacement with maximum likelihood”. In: *Acta Crystallographica - Section D Biological Crystallography* 57.10 (2001), pp. 1373–1382.

[180] Airlie J. McCoy et al. “Likelihood-enhanced fast translation functions”. In: *Acta Crystallographica Section D: Biological Crystallography* 61.4 (2005), pp. 458–464.

[181] Giovanna Scapin. “Molecular replacement then and now”. In: *Acta Crystallographica Section D: Biological Crystallography* 69.11 (2013), pp. 2266–2275.

[182] Gábor Bunkóczi, Björn Wallner, and Randy J. Read. “Local error estimates dramatically improve the utility of homology models for solving crystal structures by molecular replacement”. In: *Structure* 23.2 (2015), pp. 397–406.

[183] Huw T. Jenkins. “Fragon: Rapid high-resolution structure determination from ideal protein fragments”. In: *Acta Crystallographica Section D: Structural Biology* 74 (2018), pp. 205–214.

[184] Garry L. Taylor. “Introduction to phasing”. In: *Acta Crystallographica Section D: Biological Crystallography* 66.4 (2010), pp. 325–338.

[185] George M. Sheldrick. “Experimental phasing with SHELXC/D/E: Combining chain tracing with density modification”. In: *Acta Crystallographica Section D: Biological Crystallography* 66.4 (2010), pp. 479–485.

[186] Kevin Cowtan. “Recent developments in classical density modification”. In: *Acta Crystallographica Section D: Biological Crystallography* 66.4 (2010), pp. 470–478.

[187] Manfred S Weiss, Tom Sicker, and Rolf Hilgenfeld. “Soft X-rays, high redundancy, and ways & means proper scaling: A new procedure for automated protein structure determination via SAS”. In: *Structure* 9.9 (2001), pp. 771–777.

[188] James Doutch et al. “Challenges of sulfur SAD phasing as a routine method in macromolecular crystallography”. In: *Journal of Synchrotron Radiation* 19.1 (2012), pp. 19–29.

[189] William A Barton et al. “Highly efficient selenomethionine labeling of recombinant proteins produced in mammalian cells”. In: *Protein science* 15 (2006), pp. 2008–2013.

[190] Hao Wu et al. “Structure of human chorionic gonadotropin at 2.6 åresolution from MAD analysis of the selenomethionyl protein”. In: *Structure* 2.6 (1994), pp. 545–558.

[191] Kay Diederichs and P. Andrew Karplus. “Improved R-factors for diffraction data analysis in macromolecular crystallography”. In: *Nature Structural Biology* 4.4 (1997), pp. 269–275.

[192] G. N. Ramachandran and S. Raman. “Syntheses for the deconvolution of the Patterson function. I. General principles”. In: *Acta Crystallographica* 12.12 (1959), pp. 957–964.

[193] P. Main. “A theoretical comparison of the $\beta$,$\gamma$’ and 2Fo–Fc syntheses”. In: *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 35.5 (Sept. 1979), pp. 779–785. URL: http://scripts.iucr.org/cgi-bin/paper?S0567739479001789.

[194] Laura E. Zawadzke and Jeremy M. Berg. “The structure of a centrosymmetric protein crystal”. In: *Proteins: Structure, Function, and Bioinformatics* 16.3 (1993), pp. 301–305.

[195] R. J. Read. “Improved Fourier coefficients for maps using phases from partial structures with errors”. In: *Acta Crystallographica Section A Foundations of Crystallography* 42.3 (May 1986), pp. 140–149. URL: http://scripts.iucr.org/cgi-bin/paper?S0108767386099622.

[196] V. Luzzati. “Traitement statistique des erreurs dans la determination des structures cristallines”. In: *Acta Crystallographica* 5.6 (1952), pp. 802–810.

[197] P Therese Lang et al. “Protein structural ensembles are revealed by redefining X-ray electron density noise.” In: *Proc. Natl. Acad. Sci. USA* 111.1 (2014), pp. 237–42.

[198] T. N. Bhat. “Calculation of an OMIT map”. In: *Journal of Applied Crystallography* 21.3 (1988), pp. 279–281.

[199] Oleg V. Sobolev et al. “Polder maps: Improving OMIT maps by excluding bulk solvent”. In: *Acta Crystallographica Section D: Structural Biology* 73.2 (2017), pp. 148–157.

[200] T A Jones and S W Cowan. “Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models”. In: *Acta Crystallographica Section A* (1991), pp. 110–119.

[201] Ian J. Tickle. “Statistical quality indicators for electron-density maps”. In: *Acta Crystallographica Section D: Biological Crystallography* 68.4 (2012), pp. 454–467.

[202] Nicholas M. Pearce, Tobias Krojer, and Frank Von Delft. “Proper modelling of ligand binding requires an ensemble of bound and unbound states”. In: *Acta Crystallographica Section D: Structural Biology* 73 (2017), pp. 256–266.

[203] Nicholas Pearce. “Multi-dataset electron density analysis methods for X-ray crystallography”. PhD thesis. 2016.

[204] Axel T. Brunger. “Free R value: a nove statistical quantity for assessing the accuracy of crystal structures”. In: *Nature* 355 (1992), pp. 472–475.

[205] James M. Holton et al. “The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures”. In: *The FEBS journal* 281.18 (2014), pp. 4046–4060.

[206] Paul Emsley and Kevin Cowtan. “Coot: Model-building tools for molecular graphics”. In: *Acta Crystallographica Section D: Biological Crystallography* 60.12 I (2004), pp. 2126–2132.

[207] Anastassis Perrakis, Richard Morris, and Victor S Lamzin. “Automated protein model building combined with iterative structure refinement.” In: *Nature structural biology* 6.5 (1999), pp. 458–463.

[208] Thomas C Terwilliger. “SOLVE and RESOLVE: Automated Structure Solution and Density Modification”. In: *Methods in Enzymology* 374.1997 (2003), pp. 22–37.

[209] Kevin Cowtan. “The Buccaneer software for automated model building. 1. Tracing protein chains”. In: *Acta Crystallographica Section D: Biological Crystallography* 62.9 (2006), pp. 1002–1011.

[210] Nathaniel Echols et al. “Automating crystallographic structure solution and refinement of protein-ligand complexes”. In: *Acta Crystallographica Section D: Biological Crystallography* 70.1 (2014), pp. 144–154.

[211] Garib N. Murshudov, Alexei A. Vagin, and Eleanor J. Dodson. “Refinement of macromolecular structures by the maximum-likelihood method”. In: *Acta Crystallographica Section D: Biological Crystallography* 53.3 (1997), pp. 240–255.

[212] Dale. E. Tronrud. “Introduction to macromolecular refinement”. In: *Acta Crystallographica Section D Biological Crystallography* 60.12 (Dec. 2004), pp. 2156–2168. URL: http://scripts.iucr.org/cgi-bin/paper?S090744490402356X.

[213] Maksymilian Chruszcz et al. “Analysis of solvent content and oligomeric states in protein crystals-does symmetry matter?” In: *Protein Science* 17.4 (2008), pp. 623–632.

[214] Dale E. Tronrud. “TNT refinement package”. In: *Methods in Enzymology* (1997).

[215] Jian Sheng Jiang and Axel T. Brünger. *Protein Hydration Observed by X-ray Diffraction. Solvation Properties of Penicillopepsin and Neuraminidase Crystal Structures*. 1994.

[216] Alexei A. Vagin et al. “REFMAC5 dictionary: Organization of prior chemical knowledge and guidelines for its use”. In: *Acta Crystallographica Section D: Biological Crystallography* 60.12 I (2004), pp. 2184–2195.

[217] Paul D. Adams et al. “PHENIX: A comprehensive Python-based system for macromolecular structure solution”. In: *Acta Crystallographica Section D: Biological Crystallography* 66.2 (2010), pp. 213–221.

[218] Richard A Engh and Robert Huber. “Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement”. In: *Acta Cryst* 47 (1991), pp. 392–400.

[219] Frank H. Allen. “The Cambridge Structural Database: A quarter of a million crystal structures and rising”. In: *Acta Crystallographica Section B: Structural Science* 58.3 PART 1 (2002), pp. 380–388.

[220] Jason C. Cole, Ilenia Giangreco, and Colin R. Groom. “Using more than 801 296 small-molecule crystal structures to aid in protein structure refinement and analysis”. In: *Acta Crystallographica Section D: Structural Biology* 73 (2017), pp. 234–239.

[221] Mariusz Jaskolski et al. “Stereochemical restraints revisited: How accurate are refinement targets and how much should protein structures be allowed to deviate from them?” In: *Acta Crystallographica Section D: Biological Crystallography* 63.5 (2007), pp. 611–620.

[222] Brandl M Flensburg C Keller P Paciorek W Roversi P Sharff A Smart O S Vonrhein C Womack T O Bricogne G. Blanc E. *BUSTER version 2.10.3*. 2017.

[223] Pavel V. Afonine et al. “Towards automated crystallographic structure refinement with phenix.refine”. In: *Acta Crystallographica Section D: Biological Crystallography* 68.4 (2012), pp. 352–367.

[224] George M Sheldrick. “Crystal structure refinement with SHELXL”. In: *Acta Crystallographica Section C: Structural Chemistry* 71.Md (2015), pp. 3–8.

[225] Thomas Simonson and Gregory L Warren. “Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination”. In: *Acta Crystallographica Section D Biological Crystallography* 54 (1998), pp. 905–921. URL: http://scripts.iucr.org/cgi-bin/paper?S0907444998003254.

[226] Axel T. Brunger. “Version 1.2 of the crystallography and nmr system”. In: *Nature Protocols* 2.11 (2007), pp. 2728–2733.

[227] Tobias Krojer. *Update on batch deposition of XChem structures*. 2018. URL: https://openlabnotebooks.org/update-on-batch-deposition-of-xchem-structures/ (visited on 12/07/2019).

[228] Gozde Kar et al. “Allostery and population shift in drug discovery”. In: *Current Opinion in Pharmacology* 10.6 (2010), pp. 715–722. URL: http://dx.doi.org/10.1016/j.coph.2010.09.002.

[229] David D Boehr, Ruth Nussinov, and Peter E Wright. “The role of dynamic conformational ensembles in biomolecular recognition”. In: *Nature Chemical Biology* 5.11 (2009), pp. 789–796. URL: http://www.nature.com/doifinder/10.1038/nchembio.232.

[230] Paul Sherwood, Bernard R. Brooks, and Mark SP Sansom. “Multiscale methods for macromolecular simulations”. In: *Current Opinion in Structural Biology* 18.5 (2008), pp. 630–640.

[231] Lewis E Kay and Antony Mitttermaier. “New Tools Provide New Insights in NMR Studies of Protein Dynamics”. In: *Science* 312.April (2006), pp. 224–228.

[232] Carlos Bustamante. “In singulo Biochemistry: When Less Is More ”. In: *Annual Review of Biochemistry* 77.1 (2008), pp. 45–50. URL: http://www.annualreviews.org/doi/10.1146/annurev.biochem.012108.120952.

[233] Rahel A Woldeyes, David A Sivak, and James S Fraser. *E pluribus unum, no more: From one crystal, many conformations*. 2014.

[234] Janet L. Smith et al. “Structural Heterogeneity in Protein Crystals”. In: *Biochemistry* 25.18 (1986), pp. 5018–5027.

[235] Mark A. DePristo, Paul I.W. De Bakker, and Tom L. Blundell. “Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography”. In: *Structure* 12.5 (2004), pp. 831–838.

[236] John Kuriyan et al. “Effect of anisotropy and anharmonicity on protein crystallographic refinement. An evaluation by molecular dynamics”. In: *Journal of Molecular Biology* (1986).

[237] Elena J. Levin et al. “Ensemble Refinement of Protein Crystal Structures: Validation and Application”. In: *Structure* (2007).

[238] Thomas C Terwilliger et al. “Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard”. In: *Acta Crystallographica Section D: Biological Crystallography* 64.1 (2007), pp. 61–69.

[239] Thomas C. Terwilliger et al. “Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models”. In: *Acta Crystallographica Section D: Biological Crystallography* 63.5 (2007), pp. 597–610.

[240] Matteo Pellegrini et al. “Highly constrained multiple-copy refinement of protein crystal structures”. In: *Proteins: Structure, Function and Genetics* 29.4 (1997), pp. 426–432.

[241] B. Tom Burnley et al. “Modelling dynamics in protein crystal structures by ensemble refinement”. In: *eLife* 1 (2012), e00311.

[242] Jay Painter and Ethan A Merritt. “Optimal description of a protein structure in terms of multiple groups undergoing TLS motion research papers”. In: *Acta Crystallographica - Section D Biological Crystallography* 62 (2006), pp. 439–450.

[243] Henry Van Den Bedem et al. “Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers”. In: *Acta Crystallographica Section D: Biological Crystallography* 65.10 (2009), pp. 1107–1117.

[244] Daniel A Keedy, James S Fraser, and Henry van den Bedem. “Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit”. In: *PLoS Computational Biology* 11.10 (2015), pp. 1–22.

[245] P Therese Lang et al. “Automated electron-density sampling reveals widespread conformational polymorphism in proteins”. In: *Protein Science* 19.7 (2010), pp. 1420–1431.

[246] Rod Chalk et al. “High-Throughput Mass Spectrometry Applied to Structural Genomics”. In: *Chromatography* 1.4 (2014), pp. 159–175.

[247] C S Ho et al. “Electrospray ionisation mass spectrometry: principles and clinical applications.” In: *The Clinical biochemist. Reviews* 24.1 (2003), pp. 3–12. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1853331.

[248] John B Fenn et al. “Electrospray Ionisation for Mass Spectrometry of Large Biomolecules”. In: *Science* 246.6 (1989).

[249] A. G. Ferrige et al. “Maximum entropy deconvolution in electrospray mass spectrometry”. In: *Rapid Communications in Mass Spectrometry* 5.8 (1991), pp. 374–377.

[250] A. G. Ferrige et al. “Disentangling electrospray spectra with maximum entropy”. In: *Rapid Communications in Mass Spectrometry* 6.11 (1992), pp. 707–711.

[251] Daniel P. Donnelly et al. “Best practices and benchmarks for intact protein analysis for top-down mass spectrometry”. In: *Nature Methods* 16.7 (2019), pp. 587–594. URL: http://dx.doi.org/10.1038/s41592-019-0457-0.

[252] Cheng Zhao and Peter B O'Connor. “Removal of polyethylene glycols from protein samples using titanium dioxide”. In: *Analytical Biochemistry* 365.2 (2007), pp. 283–285.

[253] Steven L. Cohen and Brian T. Chait. “Mass spectrometry as a tool for protein crystallography”. In: *Annu. Rev. Biophys. Biomol. Struct.* 30 (2001), pp. 67–85.

[254] Dimitar B Nikolov et al. “Crystal structure of a TFIIB-TBP-TATA- element ternary complex and the TFIIB can interact with TBP-associated factors , other class II initiation factors , and”. In: *Nature* 377 (1995), pp. 119–128.

[255] Joanne E Nettleship et al. “Sample preparation and mass-spectrometric characterization of crystal-derived protein samples”. In: *Acta Crystallographica Section D: Biological Crystallography* 61.5 (2005), pp. 643–645.

[256] Krystal Cole et al. “A linear relationship between crystal size and fragment binding time observed crystallographically: Implications for fragment library screening using acoustic droplet ejection”. In: *PLoS ONE* 9.7 (2014), pp. 1–8.

[257] Nicholas M. Pearce et al. “A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density”. In: *Nature Communications* 8 (Apr. 2017), p. 15123. URL: http://www.nature.com/doifinder/10.1038/ncomms15123.

[258] Sandra B. Gabelli et al. “Structure and mechanism of the farnesyl diphosphate synthase from Trypanosoma cruzi: Implications for drug design”. In: *Proteins: Structure, Function, and Bioinformatics* 62.1 (2005), pp. 80–88. URL: http://doi.wiley.com/10.1002/prot.20754.

[259] Julio A. Urbina et al. “Trypanosoma cruzi contains major pyrophosphate stores, and its growth in vitro and in vivo is blocked by pyrophosphate analogs”. In: *Journal of Biological Chemistry* 274.47 (1999), pp. 33609–33615.

[260] Andrea Montalvetti et al. “Bisphosphonates Are Potent Inhibitors of Trypanosoma cruzi Farnesyl Pyrophosphate Synthase”. In: *Journal of Biological Chemistry* 276.36 (2001), pp. 33930–33937.

[261] Luciana R Garzoni et al. “Antiparasitic activity of risedronate in a murine model of acute Chagas’ disease”. In: *International Journal of Antimicrobial Agents* 23.3 (2004), pp. 286–290.

[262] Luciana R. Garzoni et al. “Selective in vitro effects of the farnesyl pyrophosphate synthase inhibitor risedronate on Trypanosoma cruzi”. In: *International Journal of Antimicrobial Agents* 23.3 (2004), pp. 273–285.

[263] Sundeep Khosla et al. “Benefits and risks of bisphosphonate therapy for osteoporosis”. In: *Journal of Clinical Endocrinology and Metabolism* 97.7 (2012), pp. 2272–2282.

[264] Jean Michel Rondeau et al. “Structural basis for the exceptional in vivo efficacy of bisphosphonate drugs”. In: *ChemMedChem* 1.2 (2006), pp. 267–273.

[265] K L Kavanagh et al. “The molecular mechanism of nitrogen-containing bisphosphonates as antiosteoporosis drugs”. In: *Proceedings of the National Academy of Sciences* 103.20 (2006), pp. 7829–7834.

[266] Jacob D. Durrant et al. “Non-Bisphosphonate Inhibitors of Isoprenoid Biosynthesis Identified via Computer-Aided Drug Design”. In: *Chemical Biology and Drug Design* 78.3 (2011), pp. 323–332.

[267] Sinéad Heuston et al. “Isoprenoid biosynthesis in bacterial pathogens”. In: *Microbiology (United Kingdom)* 158.6 (2012), pp. 1389–1401.

[268] Andreas L Marzinzik et al. “Discovery of novel allosteric non-bisphosphonate inhibitors of farnesyl pyrophosphate synthase by integrated lead finding”. In: *ChemMedChem* 10.11 (2015), pp. 1884–1891.

[269] Wolfgang Jahnke et al. “Allosteric non-bisphosphonate FPPS inhibitors identified by fragment-based discovery”. In: *Nature Chemical Biology* 6.9 (2010), pp. 660–666. URL: http://www.nature.com/doifinder/10.1038/nchembio.421.

[270] Jaeok Park et al. “Human farnesyl pyrophosphate synthase is allosterically inhibited by its own product”. In: *Nature Communications* 8 (2017), p. 14132. URL: http://www.nature.com/doifinder/10.1038/ncomms14132.

[271] Heping Zheng et al. “CheckMyMetal: A macromolecular metal-binding validation tool”. In: *Acta Crystallographica Section D: Structural Biology* 73 (2017), pp. 223–233.

[272] Katarzyna B. Handing et al. “Characterizing metal-binding sites in proteins with X-ray crystallography”. In: *Nature Protocols* 13.5 (2018), pp. 1062–1090. URL: http://dx.doi.org/10.1038/nprot.2018.018.

[273] Heping Zheng et al. “Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server”. In: *Nature Protocols* 9.1 (2014), pp. 156–170.

[274] Isabella Ascone and Richard Strange. “Biological X-ray absorption spectroscopy and metalloproteomics”. In: *Journal of Synchrotron Radiation* 16.3 (2009), pp. 413–421.

[275] Schrödinger, LLC. “The {PyMOL} Molecular Graphics System, Version~1.8”. Nov. 2015.

[276] Chuan Hsiang Huang et al. “Binding of nitrogen-containing bisphosphonstes (N-BPs) to the Trypanosoma cruzi fernesyl diphosphate synthase homodimer”. In: *Proteins: Structure, Function and Bioinformatics* 78.4 (2010), pp. 888–899.

[277] Mikko Laitaoja, Jarkko Valjakka, and Janne Jänis. “Zinc coordination spheres in protein structures”. In: *Inorganic Chemistry* 52.19 (2013), pp. 10983–10991.

[278] Heping Zheng et al. “Data mining of metal ion environments present in protein structures”. In: *Journal of Inorganic Biochemistry* 102.9 (Sept. 2008), pp. 1765–1776.

[279] Astrid Sigel, Helmut Sigel, and Roland K.O. Sigel. *Metal ions in biological systems*. Vol. 44. 1990, pp. 1–299.

[280] Janine L. Gray, Frank von Delft, and Paul Brennan. “Targeting the Small GTPase Superfamily through their Regulatory Proteins”. In: *Angewandte Chemie International Edition* (2019).

[281] Krister Wennerberg, Kent L. Rossman, and Channing J. Der. “The Ras superfamily at a glance”. In: *Journal of Cell Science* 118.5 (2005), pp. 843–846.

[282] Anja Schmidt and Alan Hall. “Guanine nucleotide exchange factors for Rho GTPases: Turning on the switch”. In: *Genes and Development* 16.13 (2002), pp. 1587–1609.

[283] Jl Bos, Holger Rehmann, and Alfred Wittinghofer. “GEFs and GAPs: Critical Elements in the Control of Small G Proteins”. In: *Cell* 129 (2007), pp. 865–877. URL: http://www.sciencedirect.com/science/article/pii/S0092867407006551.

[284] Nigel W. Moriarty, Ralf W. Grosse-Kunstleve, and Paul D. Adams. “Electronic ligand builder and optimization workbench (eLBOW): A tool for ligand coordinate and restraint generation”. In: *Acta Crystallographica Section D: Biological Crystallography* 65.10 (2009), pp. 1074–1080.

[285] A. A. Vagin, G. N. Murshudov, and B. V. Strokopytov. “BLANC : the program suite for protein crystallography ”. In: *Journal of Applied Crystallography* 31.1 (1998), pp. 98–102.

[286] Nathan S Babcock, Daniel A Keedy, and James S Fraser. “Model Selection for Biological Crystallography”. In: (2018).

[287] Helen Ginn. “Vagabond: a new project for macromolecular model refinement”. In: *Acta Crystallographica Section A* 74.a1 (July 2018), a454. URL: https://doi.org/10.1107/S0108767318095466.

[288] J. Michael Bradshaw et al. “Prolonged and tunable residence time using reversible covalent kinase inhibitors”. In: *Nature Chemical Biology* 11.7 (2015), pp. 525–531.

[289] Michael S. Cohen et al. “Biochemistry: Structural bioinformatics-based design of selective, irreversible kinase inhibitors”. In: *Science* 308.5726 (2005), pp. 1318–1321. URL: https://science.sciencemag.org/content/308/5726/1318.full.pdf.

[290] D A Matthews et al. “Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes”. In: 96.September (1999), pp. 11000–11007.

[291] Zhengying Pan et al. “Discovery of selective irreversible inhibitors for Bruton’s tyrosine kinase”. In: *ChemMedChem* 2.1 (2007), pp. 58–61.

[292] Annissa J. Huhn et al. “Selective Covalent Targeting of Anti-Apoptotic BFL-1 by Cysteine-Reactive Stapled Peptide Inhibitors”. In: *Cell Chemical Biology* 23.9 (2016), pp. 1123–1134. URL: http://dx.doi.org/10.1016/j.chembiol.2016.07.022.

[293] Apirat Chaikuad et al. “The Cysteinome of Protein Kinases as a Target in Drug Development”. In: *Angewandte Chemie - International Edition* 57.16 (2018), pp. 4372–4385.

[294] Khadijeh Rajabi, Alison E. Ashcroft, and Sheena E. Radford. “Mass spectrometric methods to analyze the structural organization of macromolecular complexes”. In: *Methods* 89 (2015), pp. 13–21. URL: http://dx.doi.org/10.1016/j.ymeth.2015.03.004.

[295] Georgina Berridge et al. “High-performance liquid chromatography separation and intact mass analysis of detergent-solubilized integral membrane proteins”. In: *Analytical Biochemistry* 410.2 (2011), pp. 272–280. URL: http://dx.doi.org/10.1016/j.ab.2010.11.008.

[296] Elisabeth Gasteiger et al. “Protein Analysis Tools on the ExPASy Server”. In: *The Proteomics Protocols Handbook Protein Identification and Analysis Tools on the ExPASy Server* (2005), pp. 571–607. URL: http://www.expasy.org/tools/..

[297] Dan L Crimmins, Sheenah M Mische, and Nancy D Denslow. “Chemical Cleavage of Proteins in Solution”. In: *Current Protocols in Protein Science* (2005), pp. 11.4.1–11.4.11.

[298] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *{SciPy}: Open source scientific tools for {Python}*. URL: http://www.scipy.org/.

[299] Thomas M Annesley. “Ion suppression in mass spectrometry”. In: *Clinical Chemistry* 49.7 (2003), pp. 1041–1044.

[300] Jinguo Kang, Larry A Hick, and William E Price. “Identification and reduction of ion suppression effects on pharmacokinetic parameters by polyethylene glycol 400”. In: *Rapid Communications in Mass Spectrometry* (2007), pp. 4065–4072.

[301] Markus Gerstel, Charlotte M. Deane, and Elspeth F. Garman. “Identifying and quantifying radiation damage at the atomic level”. In: *Journal of Synchrotron Radiation* 22.2 (2015), pp. 201–212.

[302] E Lesellier and C West. “Description and comparison of chromatographic tests and chemometric methods for packed column classification”. In: *Journal of Chromatography A* 1158.1-2 (2007), pp. 329–360.

[303] Andrey A Lebedev et al. “JLigand: A graphical tool for the CCP4 template-restraint library”. In: *Acta Crystallographica Section D: Biological Crystallography* 68.4 (2012), pp. 431–440.

[304] Luca Ronda et al. “Exploring methionine $\gamma$-lyase structure-function relationship via microspectrophotometry and X-ray crystallography”. In: *Biochimica et Biophysica Acta - Proteins and Proteomics* 1814.6 (2011), pp. 834–842. URL: http://dx.doi.org/10.1016/j.bbapap.2010.06.017.

[305] Roberto Orru et al. “Snapshots of enzymatic baeyer-villiger catalysis: Oxygen activation and intermediate stabilization”. In: *Journal of Biological Chemistry* 286.33 (2011), pp. 29284–29291.

[306] M D Altose et al. “Comparing protein-ligand interactions in solution and single crystals by Raman spectroscopy”. In: *Proceedings of the National Academy of Sciences* 98.6 (2001), pp. 3006–3011.

[307] John E McGeehan et al. “Raman-assisted crystallography of biomolecules at the synchrotron: Instrumentation, methods and applications”. In: *Biochimica et Biophysica Acta - Proteins and Proteomics* 1814.6 (2011), pp. 750–759. URL: http://dx.doi.org/10.1016/j.bbapap.2010.07.021.

[308] Luca Ronda et al. “From protein structure to function via single crystal optical spectroscopy”. In: *Frontiers in Molecular Biosciences* 2.April (2015), pp. 1–14. URL: http://journal.frontiersin.org/article/10.3389/fmolb.2015.00012/abstract.

[309] Henry Eyring. “Steric Hinderance and Collison Diameters”. In: *J. Amer. Chem. Soc.* 54 (1932), pp. 3191–3203.

[310] Henry Van Den Bedem et al. “Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers”. In: *Acta Crystallographica Section D: Biological Crystallography* 65.10 (2009), pp. 1107–1117.

[311] Dominique Bourgeois and Antoine Royant. “Advances in kinetic protein crystallography”. In: *Current Opinion in Structural Biology* 15.5 (2005), pp. 538–547.

[312] Ricardo J G B Campello, Davoud Moulavi, and Joerg Sander. “Density-Based Clustering Based on Hierarchical Density Estimates”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172.

[313] Jeffrey S Mugridge et al. “Structural basis of mRNA-cap recognition by Dcp1–Dcp2”. In: *Nature Publishing Group* 23.11 (2016), pp. 7–8.

[314] Maurice J. Bessman, David N. Frick, and Suzanne F. O’Handley. “The MutT proteins or ’Nudix’ hydrolases, a family of versatile, widely distributed, ’housecleaning’ enzymes”. In: *Journal of Biological Chemistry* 271.41 (1996), pp. 25059–25062.

[315] Clément Charenton et al. “Structure of the active form of Dcp1–Dcp2 decapping enzyme bound to m7GDP and its Edc3 activator”. In: *Nature Structural & Molecular Biology* 23.11 (2016), pp. 982–986. URL: http://www.nature.com/doifinder/10.1038/nsmb.3300.

[316] Janet Newman et al. “Towards rationalization of crystallization screening for small- To medium-sized academic laboratories: The PACT/JCSG+ strategy”. In: *Acta Crystallographica Section D: Biological Crystallography* 61.10 (2005), pp. 1426–1431.

[317] Terese Bergfors. “Seeds to crystals”. In: *Journal of Structural Biology* 142.1 (2003), pp. 66–76.