Discovery by Virtual Screening of Ethionamide Boosters for Tuberculosis Treatment

NATALIE JOAN TATUM

Discovery by Virtual Screening of Ethionamide Boosters for Tuberculosis Treatment

Natalie Joan Tatum

THESIS SUBMITTED TO THE DEPARTMENT OF CHEMISTRY

OF THE UNIVERSITY OF DURHAM

FOR A DOCTOR OF PHILOSOPHY

2015

DECLARATION

The work described in this thesis was undertaken at the Department of Chemistry, Durham University between October 2011 and March 2015. All work presented is my own, except where specifically stated otherwise. No part of the work presented herein has previously been submitted for a degree at this, or any other, university.

STATEMENT OF COPYRIGHT

ABSTRACT

Tuberculosis remains the world's deadliest communicable bacterial disease with an unacceptably high death rate. In 2013 an estimated 1.5 million people died as a direct result of TB, and nine million new cases were reported.[1] Multi-drug resistant (MDR) and extensively drug-resistant (XDR) tuberculosis cases are on the rise and without novel approaches to combat their spread, tuberculosis will continue to claim the lives of millions worldwide. One such novel approach is to rejuvenate the use of the second-line antibiotic ethionamide.

Ethionamide is a structural analogue of the first-line pro-drug isoniazid, which is used widely and to which there is growing resistance. Ethionamide was introduced in the 1960s and primarily used in cases of drug-resistant TB due to its severe adverse effects. This makes ethionamide an exploitable target for small-molecule booster drugs.

Expression of the enzyme responsible for ethionamide activation, EthA, is regulated by a transcriptional repressor EthR which can be inhibited to improve ethionamide activation and so reduce ethionamide treatment doses and bring an old drug new life in the clinic. EthR inhibitors are currently in development; here, chemoinformatic pipelining and virtual screening in GOLD were used to identify hits with novel scaffolds for hit-to-lead efforts from an initial library of over six million drug-like molecules.

Thermal shift assays were used to identify EthR-binding molecules and SPR was utilised to confirm and potentially quantify binding affinities. Herein are reported the co-crystal structures of several hit molecules, used to confirm and characterise the EthR-ligand complexes.

Through the application of computational, biophysical and crystallographic methods, this thesis presents several novel scaffolds for development against EthR. These novel hits will be developed to expand our arsenal against the growing, global problem of drug-resistant TB.

[1] World Health Organisation, Global Tuberculosis Report 2014.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank several people, without whom this work presented herein would not have been possible.

Firstly, my supervisor Dr. Ehmke Pohl, for the best opportunity, your full confidence and support, your honest guidance and encouraging advice, and for the occasional free coffee.

Drs. John Liebeschuetz, Jason Cole, and Colin Groom at the CCDC for the opportunity to undertake this PhD, and for their unwavering support over the course of it; for sharing their expertise and wisdom; and for their freely given time and encouragement. Additionally, Dr. Tjelvar Olsson, for his support and coding assistance, and to Dr. Oliver Korb for running the virtual screening on Darwin.

To Dr. Alain Baulard and Prof. Nicolas Willand for their technical assistance and invaluable discussions. To Drs. Steven Cobb and Victoria Money for giving me the opportunity to dip my toes into other projects, allowing me to step out of my comfort zone and learn some new skills.

Thanks to Ian Edwards, without whom we'd all be lost. Thank you for your protein production and purification expertise and everything else too. To Erin Dickinson, who did more than just "pipette some stuff;" thanks for your protein production efforts, and that big stack of crystal trays in the rock store – you rock.

Thanks to the assorted members of 209/229/235 both past and present for making work feel like play most of the time.

To Jack, we've just about managed to keep each other sane through the last two years. I love you.

Finally to my dad, my mum, James, my absurd collection of extended family and friends: your support has meant the world to me. Thank you from the bottom of my heart.

CONTENTS

List of Figures vii

List of Tables xi

Introduction 1

Chapter I: Structure-Based Drug Design 2

1.1 Drug Discovery and Development 2

1.1.1 The 20th Century Rollercoaster 2

1.1.2 Contemporary Practice 4

1.2 Computational Tools and Methods 6

1.2.1 Chemoinformatics and Chemical Space 6

1.2.1.1 Metrics of Lipophilicity 6

1.2.1.2 Lipinksi & Veber: Oral Bioavailability Trends 8

1.2.1.3 Guidelines Vs. Rules 9

1.2.2 Docking and Virtual Screening 10

1.2.2.1 From Molecular Graphics to DOCK (1978-1989) 10

1.2.2.2 The 1990s: Honing Docking Algorithms 12

1.2.2.3 Post Millennium: Towards a Realistic Model 14

1.3 Docking and Chemoinformatic Software in SBDD 16

1.3.1 Docking and Virtual Screening: GOLD 17

1.3.1.1 The Genetic Algorithm 17

1.3.1.2 Anatomy of a GOLD Run 19

1.3.1.3 Scoring: GOLD Fitness Functions 20

1.3.1.4 GOLD Optional Docking Constraints 27

1.3.1.5 Protein and Ligand Flexibility in GOLD 28

1.3.2 Chemoinformatics: KNIME 29

1.3.3 Chemical Space: Mogul and the CSD 30

1.4 Implementation of SBDD Techniques 30

1.5 Conclusion 33

Chapter II: EthR, A Virtual Screening Target for TB Co-Drugs 34

2.1 Target Discovery 36

2.2 Target Validation & Screening 38

2.3 Lead Optimisation of EthR Inhibitors 39

2.4 Further Drug Development on EthR Inhibitors 42

2.5 Conclusion 44

Chapter III: Virtual Screening Protocol Development 45
3.1 Protein Characterisation 45
3.1.1 Crystal Structures of EthR 45
3.1.2 Relibase+ and CavBase 46
3.1.3 Ensemble Docking with EthR Structures 49
3.2 Acquisition, Filtering and Clustering of a Ligand Database 52
3.2.1 The ZINC Database "DrugsNow" Set 52
3.2.2 Filtering in KNIME 52
3.2.3 Clustering the Filtered Ligand Set 54
3.3 Preliminary Docking Tests: Training Sets and Parameter Optimisation 56
3.3.1 "Decoy" Set for Probing GOLD Parameters 56
3.3.2 Optimising GOLD Parameters for EthR 57
3.3.3 Running the Virtual Screening Protocol 58
Chapter IV: Post-Screening Filtering Protocol 60
4.1 Retrieval of Data 60
4.2 Filtering 60
4.2.1 Filter by Fitness 61
4.2.2 Hydrogen Bond Geometry Filter 62
4.2.3 Calculation of Additional Descriptors 63
4.2.4 Iterative Filter Design 65
4.2.5 Visual Inspection Aided by Mogul 68
Chapter V: Thermal Shift Assays on Potential EthR Inhibitors 72
5.1 Thermal Shift Assays on EthR 73
5.2 Expression of EthR for Biophysical Assays 75
5.3 Protocol Implementation: SYPRO Orange Excited at 505-535 nm 75
5.4 Protocol Optimisation: SPYRO Orange Excited at 455-485 nm 77
5.5 Results: Biophysical Identification of Potential EthR Binders 80
5.5.1 Temperature-Dependent Thermal Shifts 82
5.5.2 Small-Shift Compounds 83
5.6 Summary 83
Chapter VI: Surface Plasmon Resonance Studies 84
6.1 Confirmation and Quantification of EthR-Binding by SPR 85
6.2 Method: Immobilisation of EthR 85
6.2.2 Method: Determination and Quantification of Compound Binding 86

6.3 Results 87
6.4 Summary 91
Chapter VII: Co-Crystal Structures of Potential EthR Inhibitors 92
7.1 Introduction 92
7.2 Native EthR Crystallisation Methods 92
7.3 Native EthR 92
7.4 Co-Crystallisation Methods 94
7.5 Synchrotron Data Collection and Assessment of Ligand Density 94
7.5.1 Compound 3 95
7.5.2 Compound 10 98
7.5.3 Compound 15 101
7.5.4 Compound 25 102
7.5.5 Compound 42 103
7.5.6 Compound 57 106
7.5.7 Compound 60 108
7.5.8 Compound 74 112
7.5.9 Compound 80 113
7.5.10 Compound 85 114
7.6 Summary **116**
Conclusion 119
References 122
Appendices 132

LIST OF FIGURES

Chapter I: Structure-based Drug Discovery

page

Figure 1.1 Schematic of typical drug discovery and development pipeline. 4

Figure 1.2 Schematic of genetic operations migration, crossover and mutation in a genetic algorithm for docking. 18

Figure 1.3 Schematic flowchart of the GOLD genetic algorithm. 20

Figure 1.4 Graphical representation of the block function before and after Gaussian smoothing. 23

Figure 1.5 Example KNIME pipeline for Chemoinformatic data calculation. 30

Figure 1.6 Graphs showing publications and citations in "virtual screening" field. 29

Figure 1.7 Chemical structure of dorzolamide. 31

Figure 1.8 Dorzolamide bound to carbonic anhydrase II. 32

Figure 1.9 Zanamivir bound to neuraminidase subtype 1 alongside docking poses from original study. 33

Chapter II: EthR, A Virtual Screening Target for TB Co-Drugs

Figure 2.1 TB and XDR-TB incidence map 35

Figure 2.2 Chemical structures of isoniazid and ethionamide 36

Figure 2.3 Summary of ethionamide mode of action 36

Figure 2.4 Crystal structure of EthR dimer (PDB: 1U9O) 37

Figure 2.5 Ligands BDM14500, BDM31343 and BDM31381 bound to EthR 39

Figure 2.6 Overlay of BDM14500 and BDM14801 co-crystal structures 40

Figure 2.7 BDM14950 bound to EthR 41

Figure 2.8 Chemical and co-crystal structure of lead, BDM41906

Figure 2.9 Chemical structures of two new EthR inhibitors in the N-phenylphenoxyacetamide class 42

Figure 2.10 Hit fragments and fragment-linked inhibitors of EthR 43

Chapter III: Virtual Screening Protocol Development

Figure 3.1 Differences in side-chain positions at the top of the EthR channel 48

Figure 3.2 Alignment of the eleven EthR crystal structures used for ensemble docking and depiction of search space. 50

Figure 3.3 Depiction of original Asn179 position with histogram of cross-docking results. 51

Figure 3.4 Depiction of flipped Asn179 residue with histogram of cross-docking results. 51
Figure 3.5 KNIME workflow for calculation of descriptors and filtering of the Drugs Now ZINC subset. 53
Figure 3.6 Schematic overview of the clustering pipeline. 54

Chapter IV: Post-Screening Filtering Protocol

Figure 4.1 Graph of re-clustered virtual screening output. 61
Figure 4.2 Heatmap and histogram distributions of occluded ligand donor and acceptor atom count, and ligand clash count. 63
Figure 4.3 Histograms of scoring fitness contributions from ligand torsion and ligand clash terms. 64
Figure 4.4 Histogram of PLP ligand correction term contribution to CHEMPLP scoring function. 64
Figure 4.5 Histograms of TPSA and AlogP within screening output. 65
Figure 4.6 Graph demonstrating exponential reduction of overall set by iterative filter design. 67
Figure 4.7 Poses of ZINC08980600, ZINC15789306 and ZINC06952722 which share a Murcko scaffold. 69
Figure 4.8 Poses of ZINC01226689 and ZINC00889208. 69
Figure 4.9 Poses of ZINC0073180 and ZINC0072059. 70
Figure 4.10 Poses of ZINC1077920 and ZINC15754437. 71

Chapter V: Thermal Shift Assays on Potential EthR Inhibitors

Figure 5.1 Protein melting curve derived from differential scanning fluorimetry as exemplified by native EthR. 72
Figure 5.2 Ligands BDM31343, BDM31381 and BDM14801 and the G106W mutant of EthR. 73
Figure 5.3 Melting curves for unbound EthR with high background fluorescence obscuring any signal at Ex/Em: 510/510 nm. 76
Figure 5.4 Melting curves for EthR with positive control ligands BDM31343 and BDM41906. 77
Figure 5.5 Melting curve of unbound EthR excited with SYPRO Orange at Ex/Em: 485/510 nm. 78
Figure 5.6 Partial table output from NAMI with grid screen of protein and SYPRO Orange concentrations. 79

Figure 5.7	Melting curves for varying concentrations of unbound EthR with 5X SYPRO Orange at Ex/Em: 485/510 nm.	79
Figure 5.8	Cohort of thermal shift assay hits taken forward.	80
Figure 5.9	Cohort of hits which showed ΔTm greater than 3°C at 400 μM.	82
Figure 5.10	Cohort of hits which showed ΔTm greater than 1°C at 400 μM.	83

Chapter VII: Surface Plasmon Resonance Studies

Figure 6.1	Schematic diagram of surface plasmon resonance.	84
Figure 6.2	SPR sensorgram of EthR immobilisation on CM5 chip	86
Figure 6.3	Example sensorgram and corresponding dose response curve	86
Figure 6.4	SPR sensorgram of compound 10 binding to EthR	87
Figure 6.5	Dose-response curve of compound 10 binding to EthR	87
Figure 6.6	Dose-response of compounds 85 and 25	88
Figure 6.7	Dose-response of compounds 3, 4 17, 31, 42 and 48	89
Figure 6.8	Dose-response of compounds 50, 57, 58, 60, 64 and 80	90

Chapter VIII: Co-Crystal Structures of EthR and Potential Inhibitors

Figure 7.1	Crystal growth of EthR in four different conditions	93
Figure 7.2	Unbiased density to which compound 3 was fitted	96
Figure 7.3	Comparison of computational and experimental ligand conformations for compound 3	96
Figure 7.4	Compound 3 bound to EthR with hydrogen bonding shown	97
Figure 7.5	Unbiased density for EthR co-crystallised with compound 10	98
Figure 7,6	One molecule of compound 10 bound to EthR	99
Figure 7.7	Compound 10 bound to EthR in two simultaneous binding modes	100
Figure 7.8	Compound 15 and the unbiased ligand density of the co-crystal structure, two which compound 15 could not be fitted	101
Figure 7.9	Compound 25 bound to EthR.	
Figure 7.10	Compound 42 bound to EthR, which makes no observable hydrogen bonding interactions; 2Fo-Fc map shown	104
Figure 7.11	Comparison of computational and experimental ligand conformations for compound 42	106
Figure 7.12	Compound 57 bound to EthR, with 2Fo-Fc map inset	107
Figure 7.13	Close-up image of putative weak hydrogen bond with M142	107
Figure 7.14	Comparison of computational and experimental ligand conformations for compound 57	108

Figure 7.15	Unbiased density map of EthR binding site for compound 60 co-crystallisation structure	109
Figure 7.16	Ranked virtual screening poses of compound 60	109
Figure 7.17	Each potential ligand orientation of compound 60 modelled in density with unbiased Fo-Fc and 2Fo-Fc maps shown	110
Figure 7.18	Final modelling of compound 60 with both binding modes simultaneously; density and ligand interactions shown	111
Figure 7.19	Compound 74 and the corresponding unbiased ligand density in the co-crystal structure	112
Figure 7.20	Both potential orientations of compound 74 modelled to density with unbiased Fo-Fc and 2Fo-Fc maps shown	113
Figure 7.21	Compound 80 and unbiased density in co-crystal structure	114
Figure 7.22	Compound 85 modelled with 2Fo-Fc density shown	115
Figure 7.23	Comparison of computational and experimental ligand conformations of compound 85	115

Conclusions

Figure 12.1	Chemical structure of BDM41906.	119

LIST OF TABLES

Chapter I: Structure-based Drug Discovery

		page
Table 1.1	PLP interaction types in the CHEMPLP scoring function	27

Chapter III: Virtual Screening Protocol Development

Table 3.1	List of EthR structures in the PDB circa Summer 2012, and their crystallisation conditions	46
Table 3.2	List of EthR structures available in Relibase+ 3.1	48
Table 3.3	Descriptors calculated in KNIME using the workflow in figure 3.5	53
Table 3.4	Proportioning the clusters for docking, in order to take roughly a third of the filtered set in total, as equally represented as possible	56

Chapter IV: Post-Screening Filtering Protocol

Table 4.1	Parameters of final filter, filter7	67

Chapter V: Thermal Shift Assays on Potential EthR Inhibitors

Table 5.1	Working codes of thermal shift 'hit' compounds with the original ZINC code	81

Chapter VI: Surface Plasmon Resonance Studies

Table 6.1	Summarised binding affinity data for each compound	91

Chapter VIII: Co-Crystal Structures of EthR and Potential Inhibitors

Table 7.1	Crystallisation screen for EthR, utilised for native and ligand co-crystallisation	93
Table 7.2	Refinement statistics for EthR with compound 10 before ligand placement, with one and with both ligand binding modes modelled	99
Table 7.3	Mogul torsion check results for compound 42	105

INTRODUCTION

The overriding aim of the work presented here was to use computational methods to identify small molecules with novel scaffolds capable of binding EthR, with a view to developing those hits into new inhibitor series.

Chapters one and two of this thesis serve as reviews, split between the computational aspects (chapter one) and the main biological focus (chapter two) which are entwined throughout this thesis. Chapter one charts the rise and implementation of computer software for the design, development and screening of drug molecules *in silico*. This includes a detailed look at the software package used for virtual screening against EthR, GOLD. Conversely, chapter two describes the previous development of EthR inhibitors, culminating in the current lead compound from a 1,2,4-oxadiazole series and several avenues of fragment-based drug discovery research.

Chapter three describes the computational methods applied in the design and implementation of a virtual screening protocol. Beginning with an initial library of over six million compounds, detailed knowledge of the protein was used to filter down to a cohort of 1.3 million candidates which were clustered and sampled in order to screen a rich and diverse set of chemical compounds against EthR. In chapter four, post-screening work yielded a set of non-stringent filters based on score, physiochemical properties and protein-ligand interaction analysis; these filters reduced the potential screening set exponentially to a small cohort of only 85 compounds for testing against the protein.

Biophysical testing in the form of thermal shift assays (chapter five), surface plasmon resonance (chapter six) were used to identify and characterise the binding of hit molecules to EthR. Finally, co-crystallisation studies detailed in chapter seven demonstrate hit molecules bound to EthR.

CHAPTER I:
STRUCTURE-BASED DRUG DISCOVERY

This first chapter summarises the history and state of drug discovery and design, firmly placing the following work in the realm of pre-clinical research, followed by a detailed discussion of key computational tools and techniques employed later in this thesis for structure-based drug discovery. Finally this chapter will cover key successes of structure-based drug design targeting a variety of human diseases.

1.1 Drug Discovery and Development

The successes and failures of the pharmaceutical industry are often and understandably high-profile. Business run on improving human health is a business one naturally wishes to see succeed, but what form that success takes depends on if it is measured in human health improvements, or profits; the pharmaceutical industry has to balance both. In the recent past, discoveries and developments have come thick and fast; as we turned the corner into the new millennium, the state of the pharmaceutical industry was vastly different from its post-WWII high and it has become necessary to change the way drug design and development are approached. This section gives a short history of the evolution of the pharmaceutical industry, and demonstrates the current pipeline for drug development and discovery.

1.1.1 The 20th Century Rollercoaster

Modern drug treatment, which can be accurately called chemotherapy, began in the early years of the 1900s but was revolutionised during the Second World War with the discovery and application of antibiotics. Up to the 1960s the pharmaceutical industry boomed with drugs to fight infectious bacterial diseases, becoming what is typically referred to as "Big Pharma" – large, transnational corporations which were the cornerstone of post-war economies. However, from the 1970s onward, a "low-hanging fruit"[1] effect became apparent, where it seemed the easy targets of discovery on the metaphorical drug tree had been found and only the harder-to-reach "high fruit" remained. With these easy wins gone priorities for drug companies changed as developing new antibiotics became a challenge with high financial risk. With a culture of complacency and a need to maintain high profits, in a world where vaccinations were close to eradicating infectious diseases, and where chronic and age-related diseases were on the rise thanks to an increased life expectancy in industrial nations, attention

turned from antibiotics to chemotherapy for more chronic illnesses such as cancer, diabetes and cardiovascular disease. Since the 1980s antibacterial drug approvals have dropped – to only two in the period 2008-2010[2] – despite numerous infectious diseases developing antibiotic resistance.

The recent Ebola virus outbreak from West Africa in 2014[3] has highlighted the risks and ease of disease spread in a globalised world, even with a virus which is, relative to the likes of influenza, difficult to contract. Moreover, intercontinental migration, tourism, and population displacement not only increase disease spread but in so doing can increase the spread of antibiotic-resistant disease. In a 2014 review of antibacterial development,[2] Stewart T. Cole reports that in Europe, the ECDC (European Centre for Disease Control) estimated multi-drug resistant bacterial infections such as tuberculosis were responsible for 25,000 deaths annually, costing the EU in excess of €1.5 billion. Cole goes on to report that studies in the USA indicated that MRSA (methicillin-resistant *Staphylococcus aureus*) claimed more lives annually than emphysema, HIV/AIDS, Parkinson's disease and homicide *combined*. With such a global antibiotic crisis there are questions as to why new anti-infective research is so stagnant and unyielding of new chemical entities. The problem is, unsurprising, complex, and many reasons and solutions have been suggested.

In the specific case of antibiotics (particularly broad-spectrum drugs of which only two[4] have been approved in the last 40 years), the work became cost ineffective compared to more lucrative areas and a gradual shift towards these chronic, non-communicable diseases resulted in a lack of expertise in the antibiotic area. However, this did not provide long-term security for BigPharma[5] – the costs of drug development soared towards the end of the twentieth century, until in 2010 the cost of developing a new chemical entity reached an reported estimate of $2bn.[1] Of course the cost of making a drug is a highly disputed figure, with most estimates professed by drug companies such as Eli Lilly being in the billion dollar range. A simple study by Forbes utilised a rough yardstick of R&D investment of several major drug companies, divided by the number of drugs successfully brought to market in a given period; this placed costs between $4bn and $11bn.[6] Alternatively, a study by Light and Warburton[7] suggested that reported "costs" are perhaps "padded" with legal and non-research expenditures (such as marketing and insurance), and their study places a more modest median figure of $43.4 million on development. One interesting point made by Light and Warburton is that such high cost estimates in the billion dollar range may be self-fulfilling prophecies.

Whatever the real cost figure, there is strong investment in drug discovery and development and yet attrition (failure) rates for small molecules stand at 93% (for drugs entering clinical trials but not, ultimately, reaching the market). In the light of such difficulties, many recent reviews[1,2,5,8–13] have covered the scientific and financial challenges faced in the industry and how advances could be made to mitigate these costly losses, including shifts in scientific approach and partnerships between academia and industry. However, one clear overriding conclusion can be drawn: more time, care and a better distribution of efforts must be made on the pre-clinical development phase, where currently only a third of financial investment and time is spent.

1.1.2 Contemporary Practice

A typical drug development pipeline, from target discovery to approved drug on the market, has two main phases: pre-clinical, which covers everything at the molecular level before human clinical trials; and clinical, which involves proof-of-concept, efficacy and safety trials in patients (figure 1.1). The whole drug discovery and development process takes, on average, ten to fifteen years.

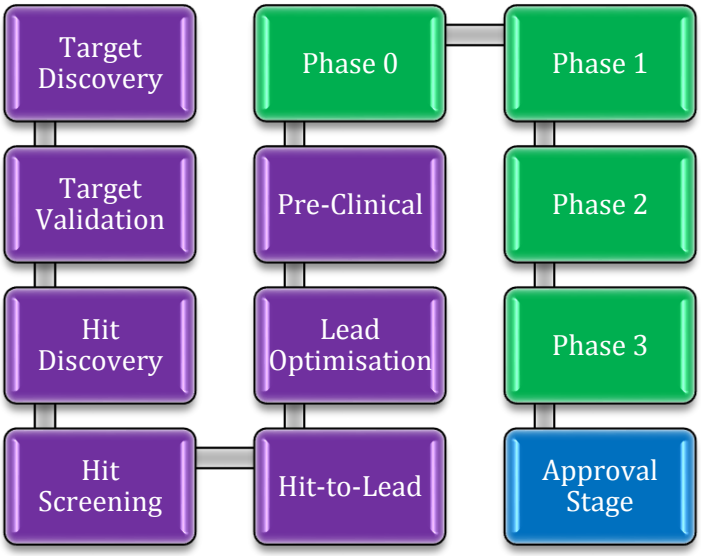


Figure 1.1: Schematic representation of the typical drug discovery and development pipeline. The pre-clinical (purple), clinical (green) and approval (blue) stages are colour-coded for clarity.

The pre-clinical phase is the last step before an application is made for clinical trials, in the form of an Investigational New Drug Application (in the US, with the FDA) or a Clinical Trial Application (in Europe with the EMA).[6] The pre-clinical stage is therefore the last opportunity to gather data for (or against) investing in clinical trials, involving extensive *in vivo* testing on animal models. Pharmacokinetic and pharmacodynamic (PK/PD) data[9] is acquired from non-human subjects at this point, typically elucidating

what actions the body takes on the drug (PK) and what effect the drug has on the body (PD). It is often beneficial for multiple candidates to undergo pre-clinical PK/PD tests.

As this phase is the last laboratory-based (as opposed to clinic-based) step in the development pipeline, the pre-clinical step has been cited as the point at which most improvement can be made to reduce attrition rates in the clinic; this includes not only improving the data collection,[8] but also the methods. Effective *in vivo* models are essential to qualify both the target and the candidate molecule, determine therapeutic doses and justify entry into clinical trials.[1]

Before drug approval there are four phases to a typical clinical trial: phases 0 to 3. The PK/PD studies in non-human models are repeated on a small number of patients (typically less than ten) in sub-optimal doses; this is the first test on patients with the disease to be treated, and determines bioavailability and half-life of the drug in humans. Phase I involves a larger (~100) cohort of patients who are subjected to a range of doses – still typically sub-therapeutic - to check efficacy and determine the therapeutic dose to be used in phase II. For this very reason most drugs (barring rare extreme reactions)[14] progress through to phase II, which is the point at which a true proof-of-concept test occurs. Phase II involves a few hundred patients, dosed at the now-determined therapeutic level. The efficacy and safety of the drug is put to the test, and it is at this point at which large numbers of drugs fail. An in-house study at Pfizer in 2012[9] analysed 44 programs which reached phase II trials in the period 2005-2009 in order to determine what lessons could be learned. The study itself was revealing; of the 44 case studies, only a third were deemed to have achieved "proof-of-concept" with the target modulated by the administered drug. A smaller number of the case studies (25%) were able to modulate the target but were deemed unsafe, and a disturbingly larger percentage – 43% - did not "adequately demonstrate" the mechanism and so failed proof-of-concept.[9]

At phase III the drug in trial is now proven to have an effect (if indeed, it has passed through to phase III), and trials are conducted on thousands of patients. It is clear that phase II is the key hurdle and that, for all the care in design and optimisation in pre-clinical drug discovery, it is rare to develop a drug which succeeds in safely treating patients. Yet for the 93% of drugs entering clinical trials which fail there are still 7% of small molecule drugs which are successful and subsequently approved.[1]

1.2 Computational Tools and Methods

The drug discovery work presented in the following chapters involves a variety of computational and biophysical techniques; the latter will be explained in context in their relevant chapters, however for the sake of brevity when explaining the protocols implemented and choices made, the computational tools and methods will be discussed in detail here. This primarily focuses on the development of docking algorithms, and the software GOLD which was used for this purpose.

1.2.1 Chemoinformatics and Chemical Space

To return to the discussion of attrition rates in clinical trials, much discussion has been made about the medicinal chemistry approaches[13] which dictate hit-to-lead programmes, and how this may influence success or failure. A trend has been identified by which hit molecules are "inflated" through the optimisation process in both molecular weight and lipophilicity in order to increase potency and cell permeability; this is typically to their detriment, reflected by attrition rates, and termed 'molecular obesity'. A review by Haan[15] coined the phrase in 2011 and demonstrated that, for molecules which pass their proof-of-concept test and yet are still pulled from clinical trials, reason for failure typically lies with toxicological issues; that is, the drugs in question work on their target, but fail their patient. Haan points to molecular obesity as a potential cause, simply adding mass rather than reappraising the entire compound.

Multiple chemoinformatic metrics and rules-of-thumb for quantifying a drug's chemical properties have grown around trying to guide drug design. With reference to key indicators, those metrics discussed in detail will be lipophilicity, measured by logP, Lipinski's 'rules' which are a guideline for oral bioavailability, and Veber's augments to Lipinski's 'rules,' which added considerations of polar surface area and the number of rotatable bonds in a molecule.

1.2.1.1 Metrics of Lipophilicity

Lipophilicity is tied to the ability of a drug molecule to pass through the cell membrane, which is comprised of lipids. One particular method of describing the lipophilicity of a molecule is the partition co-efficient *P*, which is the ratio of concentrations of the compound to be found in a two-compartment, closed system at equilibrium, comprised of octanol (organic) and water (aqueous). It is more typically used in the logarithmic form *logP*.

Acquisition of experimental *logP* data is rare, as for a database of compounds it would be understandably time-consuming. Therefore, in the 1970s a general method of calculating *logP* values was introduced and has been subsequently developed into a plethora of methods for *logP* calculation, with structure-based methods being the fastest and easiest to calculate for large databases of compounds. Structure-based (2D) methods for *logP* calculation fall into two subcategories: fragmental methods (such as CLogP), and atom-based methods (such as AlogP and XlogP); these named examples will be described in more detail.

Fragmental methods reduce a molecule to fragments with correction factors, a typical drawback being that some fragments which are poorly (or not at all) defined prevent a calculation; or that there are arbitrary fragments which may not represent the fragment as a whole.

CLogP utilises an equation common to all fragmental methods (equation 1.1), and derives terms from data on simple molecules; it is the most widely used calculation.

$$\log P = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_j F_j$$

Equation 1.1: Fragment-based methods for logP calculation utilise fragmental constants (f) for a fragment, and the incidence (a) of that fragment in the query molecule. The second term utilises a correction factor F, and its frequency (b).

Utilising the 2D molecular structure of a molecule, CLogP defines a fragment as an atom or group of atoms bounded by "isolating carbons", which are bound to heteroatoms by only single or aromatic bonds; it will be defined as a polar fragment if the isolating carbon is bound to anything other than hydrogen. Tables of fragmental constants for CLogP were derived using experimental *logP* values on known structures,[16] and consequent regression analysis on constituent functional groups. Constants for carbon and hydrogen were derived from small alkanes and, for more complex fragments, correction factors are used. Additionally, interaction factors account for electronic effects through bonds, making correct atom and bond assignment of the input structure essential.

Atom-based methods use single atoms rather than fragments and sum the contributions of different types of atoms. Ambiguity is avoided as a large number of atom environments are classified, but atom-based models do not account for long-range interactions.

ALogP is based on the Ghose-Crippen approach which operates without a correction factor, according to equation 1.2. In this approach there are 120 atom types classified, derived from a training set of nearly 900 structures; for example carbons are

classified by hybridisation state and depending upon the nature of their neighbour. ALogP is also known as ALogP98, and is a refined version of the Ghose-Crippen Approach utilising an additional 68 atomic definitions from a training set of 9000 structures.

$$\log P = \sum a_k N_k$$

Equation 1.2: ALogP sums the products of atom (a) type (k) by the number of times (N) the type (k) appears in the query molecule.

Contrastingly, XLogP includes a correction factor. A total of 90 atom types (compared to a larger 186 atom types for ALogP) and 10 correction factors are used, the latter of which compensate for some intramolecular interactions such as internal hydrogen bonds. The calculation for XLogP is given below as equation 1.3.

$$\log P = \sum_{i=1}^{M} a_i A_i + \sum_{j=1}^{N} c_j C_j$$

Equation 1.3: XLogP contains two terms. The first summation is the contribution (a) and occurrence (A) of the atom (i). The second summation are contributions (c) and occurrence (C) of the necessary correction factor.

A study by Mannhold *et al.*[16] in 2008 compared up to 30 different *logP* calculation methods on a more than 96,000 compounds (266 public, 882 from Nycomed and over 95,000 from Pfizer). They reported only seven methods which produced reasonable accuracy for the larger datasets, and of those they cited ALogP and XLogP as among the most consistent.

1.2.1.2 Lipinski and Veber: Oral Bioavailability Trends

GlaxoSmithKline and Pfizer have developed internal rules to control molecular obesity and problem compounds. As molecular weight and *logP* are convoluted in the theory of molecular obesity, the GSK 4/400 rule arose from a study of 30,000 compounds which showed those having a CLogP < 4 and a molecular weight < 400 displayed a more favourable ADMET (absorption, delivery, metabolism, excretion and toxicity) profile. This is similar to the Pfizer 3/75 rule, which utilises a CLogP < 3 and a total polar surface area of greater than 75 $Å^3$; compounds fitting this profile showed a reduced *in vivo* toxicity profile compared to those of the inverse properties.

However by far the most famous of the 'drug-like' metrics is Lipinski's "Rule of 5" (RO5). In 1997 Lipinksi *et al*. studied a subset of 2254 compounds which had entered phase II trials and therefore had good physiochemical properties. The aim of the study was to identify properties correlated with oral bioavailability, and the results showed most were small and somewhat lipophilic. Therefore, Lipinksi *et al*. proposed that generally an orally active drug molecule would conform to at least three of the following four criteria, firstly, a CLogP < 5; secondly, the drug would have a molecular weight lower than 500 g mol^{-1}; thirdly, the molecule would contain no more than 10 hydrogen bond acceptors; and finally, the molecule would contain no more than 5 hydrogen bond donors. Lipinski intended the RO5 to be a conservative indicator, in that they described 90% of the drugs which achieved phase II clinical trials.

In 2004 Lipinski responded to what he called the "rule-of-five revolution" with a short perspective, which addressed misinterpretations of the original 1997 paper as well as offering tool-like (ie. probe) and lead-like (ie. fragment) property guidelines. Lipinksi reiterated that "passing the RO5 is no guarantee that a compound is drug-like," as his RO5 had been applied as such by many. He acknowledged Veber *et al*.'s observation that compounds with greater than 10 rotatable bonds correlated to lower oral bioavailability in rats, adding that for *in vitro* screening ligand affinity "on average decreases 0.5 kcal for each two rotatable bonds." Veber *et al*. also indicate a polar surface area of equal or less than 140 $Å^2$ associated with good oral bioavailability. These metrics, derived from a study of compounds with good ADMET profiles, yielded powerful tools for drug development.

1.2.1.3 Guidelines vs. Rules

It seems obvious to say, but there is no clear rule-of-thumb or definitive metric which can be used to define what makes a good drug - or even a drug likely to succeed. Lipinski *et al*.'s approach, followed by Veber *et al*.'s are correlations of success and physiochemical property, but outliers will always exist. For hit identification screening, it is better to sample a wide range of chemical space in order to identify novel chemotypes, however databases of commercially available structures such as ZINC contain 35 million compounds, a practically untestable number.

Therefore in terms of hit identification by screening it is best to use such metrics described here to guide a search towards a starting point with beneficial properties, and trust that potent compounds will be identified if the protocols have been implemented correctly. For medicinal chemists the challenge is then to optimise the structure with the dangers of 'molecular obesity' in mind.

1.2.2 Docking and Virtual Screening

The advent of protein crystallography, with the solution and publication of the structures of myoglobin[17] and haemoglobin,[18] opened the doors to a new realm of rational, structure-based drug design. Even before sufficiently sophisticated molecular graphics, pioneering work used scale models of haemoglobin and prior knowledge of ligands to develop novel compounds effecting oxygen liberation.[19,20] As computer processing power began to increase it became possible to solve the more complex problems with algorithms, first for representing molecular shape and surface,[21,22] and on, to calculating protein-ligand interactions,[23] grid-based search methods, and shape complementarity,[24] until finally the first automated molecular docking program from the Kuntz group was released in 1989.[25] This section will describe evolution of docking from the development of the initial algorithms for automated rigid docking in the 1980s, through the development of flexible docking in the 1990s, and into the 2000s and the post-millennial work to improve the accuracy of scoring functions, and to more accurately represent the biophysical realities in what is a very derivative, isolated method of binding simulation.

1.2.2.1 From Molecular Graphics to DOCK (1978-1989)

Docking as a method of discovering and designing a drug molecule which interacts with its protein target in three dimensions relies upon three core aspects: firstly, accurately positioning a ligand in its binding conformation with the receptor; secondly, in scoring or assigning the energy of such binding; and thirdly, representing this on screen. In 1978 Greer and Bush[21] described a method for calculating the molecular surface to be represented as a contour map, based on the proposition by Richards[26] the year before, utilising water position and implied van der Waals interactions where water is excluded in the structure. In the following five years molecular graphics were developed[27–29] capable of displaying the structures and surfaces on a workstation screen with real-time manipulation, and so with representations came the ability to conduct manual docking without a scale model taking up a bench.

The first automated docking program, DOCK,[25,30] traces its roots back to the development of molecular graphics more literally than most; it was within the Kuntz group at UCSF, where these graphics were developed, that the first algorithm for calculation of protein-ligand interactions was written and published in 1982. A version of molecular docking had occurred before, for example with Levinthal *et al*.[19] and the design of novel ligands for haemoglobin; however, the new geometric approach from

Kuntz *et al*.[23] was the first step towards automation of the docking process, and making library screening feasible even later down the line.

Earliest versions of DOCK utilised a geometric approach. DOCK described the protein binding surface using spheres,[25] which were derived from tangents from one solvent-accessible surface[22] point meeting another; a resulting sphere has its centre at the meeting point and a radius to the surface. Various factors prune the number of spheres, such as only the smallest possible sphere generated being kept, and only one sphere per receptor atom. Spheres are grouped where they overlap and so the largest cluster of spheres should correspond to the binding site (spheres of radius greater than 5 Å are discarded as being likely to project into solvent). Ligand spheres are created by inverting the tangent, projecting into the ligand structure rather than out as in the receptor. Spheres are paired, by matching centre-to-centre distances within a distance threshold, and a minimum of four such matches are required to yield a ligand orientation. The algorithm repeats to find multiple potential orientations. Scoring was done simply by determining to what extent the ligand atoms violated the lower bounds of the van der Waals radii of the atoms, summed over all ligand and receptor atoms. When compared to modern scoring functions (such as those described in section 1.3.1.2 for GOLD), this is highly simplistic but the best compromise considering the limited computing power at the time.

From this point onward progress was swift. DOCKER,[31] a program from Blundell group, is a manual polypeptide-protein docking program which, though manual, allows for two molecules up to 400 atoms in size, one of which can be flexible, with real-time movement. The algorithm calculates van der Waals and accessible surface areas, with an interaction calculation and energy minimisation to a local minimum within the calculated energy. Though not widely applied, it does represent the first program of note for such a purpose.

In 1983 Connolly[22] published a method of calculating the solvent accessible surface of a macromolecule. The Connolly method uses a probe sphere which represents a solvent molecule, and the algorithm positions this "solvent molecule" tangentially to all protein atoms at thousands of positions; each "solvent molecule" which does not overlap van der Waals radii with a protein atom identifies a point on a solvent-accessible grid-point map of the surface. The area and co-ordinates of the resulting points are calculated to give a continuous surface which can be displayed in the molecular viewer in three-dimensions in real-time. It allowed areas and volumes to be calculated and, as it did not rely on atomistic wire-mesh or space-fill models, it was faster and showed only those surfaces of relevance for docking.

Work by DesJarlais *et al*.[32] in 1986 built on early versions of DOCK by introducing some limited flexibility of substrate. By splitting the ligand into smaller, rigid fragments to be docked separately and rejoining the fragments later in the docking process for energy minimisation, geometries which "may not have been found" by manual docking were found, and were close to the original test-case X-ray crystal structures used. Crucially, the matching of ligand to receptor site was by shape complementarity, a method DesJarlais *et al*. expanded upon in a later paper.[24]

Shape complementarity was an extension of the sphere-based geometric mapping described above, by which the spheres are used to generate a surface which describes the ligand shape and the receptor shape, and these are matched. For rigid molecules, this allowed much faster computation, and DOCK introduced a new scoring function to reward good contacts (ideal van der Waals distances) and penalise poor contacts (violating van der Waals distances).[24] Eventually, this expanded to include non-bonded terms from an AMBER force field[33] to simulate electrostatic as well as shape complementarity; this would provide a more robust scoring function which could be used to select molecules from a database for testing.[25]

Finally, the first version of DOCK was made publically available in 1989, although the paper was not published until three years later to coincide with the improved 2.0 release.[25] This program was the first of its kind and was the culmination of almost a decade of development.

1.2.2.2 The 1990s: Honing Docking Algorithms

Moore's Law derives from an observation made by Gordon E. Moore,[34] head of the Intel Corporation, in 1965; he noted that the number of transistors on a dense integrated circuit (such as those found in computer hardware) double every two years. This has since been a self-fulfilling prophecy, as it is both true and used as a design guideline in industry. In a simpler form, it means that computer processing power doubles every two years, and this also holds true. In 1989, the Intel 489 processor had 1.2 million transistors and a speed of 25 MHz; by 1999, the Intel Pentium III processor used 28 million transistors and had a speed of 733 MHz. This exponential increase in computing power through the 1990s allowed great strides forward in docking algorithms, and the main question to be answered was how to treat a ligand as flexible and retain tractable speeds for screening large databases. Multiple methods grew from the DOCK implementation, and some key developments in ligand flexibility and novel docking algorithm approaches will be discussed here.

The previous work by DesJarlais *et al*.[32] on fragmental approaches to ligand flexibility (detailed above) was a starting point for incremental methods, published by Leach and Kuntz in 1992.[35] Their method, dubbed "Directed DOCK" utilises an *anchor fragment* – a portion of the molecule which has the least conformational flexibility – and the assumption is made that ligands will preferentially adopt low energy conformations. Multiple orientations are generated for the anchor through a conformational search, the results of which are pruned and clustered, and then representatives from each cluster form the centre-point of the systematic conformational exploration of the rest of the molecule. This method was later implemented in FlexX.[36]

The Directed DOCK method still considered the protein to be completely rigid, however Leach went on to describe a method of introducing some flexibility to the receptor in 1994,[37] by which the protein backbone remains rigid but side chains can adopt positions from library of conformers derived from observed orientations in protein crystal structures.

Two key programs were made available in the 1990s, which are widely used two decades later: these are GOLD and FlexX, and they utilise different approaches to the docking problem. FlexX implements an incremental approach, based on the work by Leach and Kuntz described above.

GOLD utilises a genetic algorithm,[38,39] which is a non-deterministic approach to docking. This means that despite identical starting points and as the conformations go through a combinatorial optimisation, the docking process is random and the results will never by definition be identical. This is achieved by encoding rotatable bond torsion angles in 8 byte strings, and concatenation of these strings correspond to a conformation of the molecule. A randomly generated initial population of these strings (ie. conformations) undergo various types of alteration – by swapping substrings (crossover) or switching bit values to their opposite number (mutation) – and are scored.[40] This is discussed in greater detail in section 1.3.1 with relevance to the GOLD implementation.

In contrast to GOLD, FlexX[36,41] is based on the Directed DOCK[35] approach with a scoring function derived from a *de novo* design software, LUDI.[42] LUDI addressed the conformational flexibility problems of the 1980s development by using a library of fragments, placed and scored in an active site, which was mapped for interactions using non-bonding contact information from the Cambridge Structural Database (CSD).[43] By this method, LUDI would generate completely novel compounds from its fragment and bridging molecule library. FlexX[36] combined the incremental methods[35] with the scoring

from LUDI[42] and added conformational searching with torsional preferences determined by observations from the CSD.

By the mid-1990s, flexible ligand docking was feasible, and could be achieved by multiple methods across a range of programs. It was now possible to screen libraries of compounds (and given the growing computer processing power, without supercomputer access) and some limited protein flexibility was possible. To bridge the gap between flexible side-chains with a rigid backbone, and with full molecular dynamics simulations, Knegtel *et al.*[44] introduced ensemble docking in 1997 as a method of representing dynamic protein movement and conferring a kind of receptor flexibility.

Knegtel *et al*. noted that, with multiple co-crystal structures of different ligands with the same receptor, there could be sometimes seen "modest but significant" changes in conformation; and, though a limited number of models could not represent the full range of conformational space available to a protein in solution, could provide a better model for ligand discovery.[44]

Two methods of utilising protein ensembles were suggested for use in DOCK 3.5: an "energy-weighted" method, and a "geometry-weighted" method.[44] The "energy-weighted" method of ensemble docking calculates the interaction energies for every atom of each protein structure, and a weighted potential is calculated from this to average over all macromolecule structures. Attractive potentials are given priority over repulsive potentials, unless the repulsive potential is present in all structures; this method allows for small local variations. Contrastingly, the "geometry-weighted" method calculates the mean position and a variance for every atom over all structures, allowing greater flexibility and allowing ligands to take advantage of greater site plasticity, but this method gives higher ambiguity in the effects of binding on the protein structure.

The ensemble docking method gave the option in docking to avoid the "computationally demanding" task of free energy calculations on fully flexible protein-ligand complexes with explicit solvent (ie. molecular dynamics) and instead take advantage of multiple experimental crystal structures of receptor structures.

1.2.2.3 Post Millennium: Towards A Realistic Model

Prior to 2000, docking strategies typically consisted of a flexible ligand and a largely rigid protein, perhaps with a co-factor which was considered part of the protein model, but little else. Ensemble docking[44–46] had been introduced but was as yet undeveloped as a method of representing protein dynamics, and factors such as water and covalent adduct formation were not common-place implementations. As computational power continued

its upward trajectory, attention shifted from pose generation algorithms (which were fairly well-honed thanks to the 1990s) and onto the task of more accurately simulating the protein environment.

Over ten years since the first docking program DOCK was released, DOCK 4.0[30] implemented flexible docking by incremental methods (like FlexX) and random search approaches. Alternatively, another program DARWIN[47] utilised CHARMM[48] force fields as a scoring function along with a genetic algorithm for conformational optimisation (see section 1.3.1.3 for discussion of scoring functions). With desktop units housing increasing computational processing power, a variety of programs and methods were being developed and implemented all over the scientific community. In 2004, another widely used program Glide[49] was developed which uses an initial approximation of the binding orientation to narrow down docking candidates; this is followed by a refinement using random sampling.

Water molecules can play significant and essential roles in binding. To utilise water molecules in molecular docking, a variety of methods have been implemented. For example, FlexX[36] determines favourable potential water positions in the binding site during a pre-processing phase of the docking experiment;[50] then, during the incremental docking, water molecules in these putative positions are evaluated for the ability to form hydrogen bonds with the ligand. Any inclusion of water is used to further optimise the ligand position.

Alternatively, GOLD[51] implemented water models which (contrary to FlexX) are fully rotatable, can be bound or displaced, and if displaced then confer a reward in the scoring function (equation 1.4.)[52] The original fitness (σ_o, discussed in section 1.3.1.3) is altered by a summation over all atoms (ligand, protein and other waters) which interact with each specific water molecule in binding. The summation term utilises the occupancy of the water ($o(w)$ is either 1, and on, or 0, and off) the intrinsic binding affinity of water, $\sigma_i(w)$; and an energy penalty, as movement of water incurs a loss of entropy as a rigid-body (σ_p).

$$\text{Fitness} = \sigma_o + \sum_w o(w)(\sigma_p + \sigma_i(w))$$

Equation 1.4: The original fitness function (σ_o) is altered by the summation over all ligand, protein and other waters the water molecule interacts with. GOLD's fitness functions are discussed in section 1.3.1.3.[52]

The authors conclude in their validation of GOLD's water implementation that including water molecules where relevant could improve docking score correlations with experimental affinities.[52]

Two highly comprehensive reviews in 2005[53,54] demonstrated that with the plethora of docking programs available, it had become difficult, though necessary, to compare methodologies and performance. Mohan *et al*.[54] stress the importance of understanding the strengths, approximations and limitations of docking software. The need for standardised test data was highlighted, as well by Cole *et al.*,[53] and this need was later addressed by the CCDC/Astex set[55] and the DUD set (directory of useful decoys). Decoys will be discussed in more detail in chapter four.[56]

As a final point in the cherry-picking of post-millennial developments in docking, Durrant and McCammon[57] described in 2011 the growing role of molecular dynamics in drug discovery. Though not yet feasible for database virtual screening, molecular dynamics are now a powerful technique for identifying allosteric binding sites and enhancing docking methodologies (for example for simulating hits from initial virtual screens) with atomistic models including explicit solvent. However, even simulations of 100 ns take weeks on a multi-processor machine; powerful as molecular dynamics simulations are, they are far from a reasonable solution as yet.

Despite over three decades of development, at the end of the 2000s the "docking problem" remains; our ability to predict binding modes of molecules hinges not on our computational power, but on our fundamental knowledge of how proteins and ligands interact. We have ensemble docking and some modelling of flexible loop movements to bridge the gap between a rigid or semi-rigid protein model and full molecular dynamics simulations, and under replicative test conditions, the most popular docking programs (FlexX,[36] Glide, and GOLD[51]) are able to replicate binding modes and have been utilised in the identification of various small molecule inhibitors.[58–67] The problem resoundingly lies with the ability of these programs to predict the energy of binding; though *in silico* screening is (in terms of the scientific method) experimental, the results are only a model of the system and therefore must be tested *in vitro* using biophysical methods.

1.3 Docking and Chemoinformatic Software in SBDD

Many computational packages and strategies have been given as examples in the previous section. Here, software used in the course of the work presented in this thesis will be discussed in detail.

1.3.1 Docking and Virtual Screening: GOLD

1.3.1.1 The Genetic Algorithm

The problem of sampling the astronomical search spaces in docking have been touched upon, as well as strategies for finding global minima. Genetic algorithms (GAs) have already been mentioned in this chapter as one such strategy, explained as using concatenated 8-byte strings representing torsions of rotatable bonds, which undergo certain alterations to increase diversity and avoid local minima. Here, the GOLD implementation is described in greater detail.[38–40,68]

To begin, each conformation of the ligand (each concatenated 8-byte string) is designated a *chromosome*, and is therefore also a docking *pose*. This pose is not only the conformation but the position (x, y, z) of the molecule. An initial *population* of *chromosomes* (collection of concatenated strings) is created at random, and these are scored and ranked according to the chosen fitness function parameters.[38]

Populations in GOLD are arranged on "islands"; the default settings distribute populations of 100 chromosomes on each of the five islands. Therefore, GOLD creates an initial cohort of 500 ligand conformations in strings. The use of islands prevents conversion on any one conformation too early, as one of the possible genetic operators is *migration* of chromosomes between islands; strings can be transferred wholly from one cohort to another.

As in nature, selection pressures favour the fitter members of the population, which is maintained at 500 conformations. The *selection pressure* parameter in GOLD is the ratio of the likelihood of a *fit* chromosome to be chosen over an average chromosome, to become a *parent* chromosome.[68] As a default, this is set to 1.1, therefore the better scoring *child* pose is more likely to replace an average one in the parent population.

Niches describe the similarity between chromosomes; two individual chromosomes share a niche if their donor and acceptor atom co-ordinates are within a root mean square deviation (RMSD) of 1.0 Å. GOLD controls niche sizes to prevent convergence and increase diversity within the population. Therefore, if the number of individuals in a niche exceeds the parameter '*nichesize*', a suitably higher scoring child chromosome will replace the least fit chromosome in that over-filled niche, rather than the least fit chromosome in the entire population. The default value of *nichesize* in GOLD is set at 2. Overall, this means the initial population is of 500 individuals in populations of 100 to each of the five islands, and largest niche contains two chromosomes within 1.0 Å similarity.

The GOLD genetic algorithm contains three specific operators (figure 1.2) which dictate the change in the chromosome drawn from the population: *mutation*, *crossover*,

and *migration*. Which operator is used and how frequently depends on the weights assigned to them, determining which is selected as though on a 'roulette wheel'. In GOLD, *mutation* and *crossover* have weights of 95 and *migration* has a weight of 15 as determined from the development of the algorithm; this means *mutation* and *crossover* are equally likely to be picked as the operator, but both are vastly more likely than *migration*.

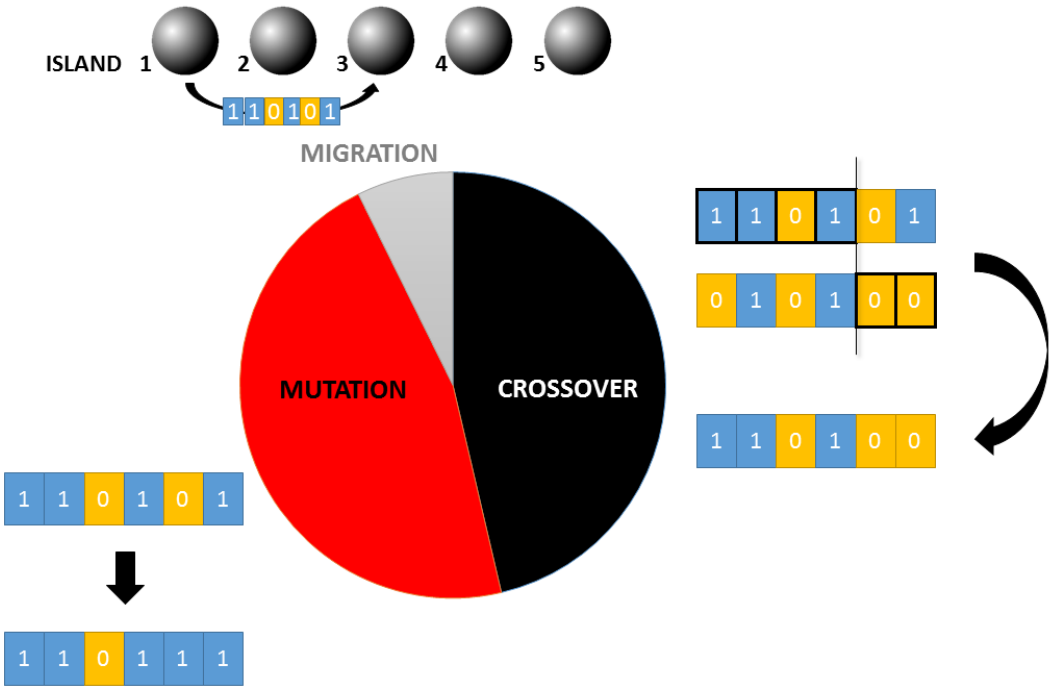


Figure 1.2: Examples of the three genetic operations, migration (movement between islands), crossover (swapping parts of strings encoding poses), and mutation (direct change of string).

Migration has already been explained as the movement of chromosomes between islands; if this scores higher than a chromosome in the population of the new island, it replaces the least fit. If the migrated chromosome is not fitter than any in the new island, it is discarded. Migration converges different populations on higher scoring poses.

Crossover works much as the natural counterpart. As the poses are encoded in strings, the crossover operator exchanges substrings to create a new child chromosome. Similarly, mutation introduces random changes to the parent chromosome by swapping the bit values for their opposite number (0 to 1 and 1 to 0). Both crossover and mutation increase diversity in the population as they will essentially swap or change randomly the conformation of a pose, allowing the rapid generation and scoring of a vast number of potential ligand conformations. These conformations eventually converge in a run to give just one result.

1.3.1.2 Anatomy of a GOLD Run

The number of GA operations is distinct to the number of GA runs: for example 100,000 *operations* such as migration or mutation will be conducted in one *run*, giving one *pose*. A docking experiment could conduct 10 GA *runs*, giving 10 *poses*. Schematic figure 1.3 demonstrates the typical process.

Each run consists of a number of operations, and GOLD will use information such as the number of rotatable bonds and the ligand flexibility to determine how many GA operations are required for optimal search and speed. Then, if search efficiency is selected to be 100%, GOLD will set up the optimal number of operations. In GOLD's virtual screening settings, this is reduced to 30%, which improves speed in an attempt to get a good answer, while compromising slightly on exhaustive accuracy.

The user is required to specify termination conditions when running a docking, and whichever is satisfied first will end the runs; these are *early termination* and *diverse solutions*. *Early termination* will stop the runs when a user-defined number of solutions have been found within a user-defined RMSD; for example 5 solutions within 1.5 Å of each other. Even if ten runs have been instructed, if five solutions are found, the docking will end. However, *early termination* can be misleading, as the poses may not represent the true global minimum. In virtual screening it is better to acquire an understanding of the range of possible binding poses of the ligand to the binding site, making *early termination* unfavourable.

The *diverse solutions* option, in contrast, will end the docking when a user-defined number of solutions within a cluster have been found, and this cluster will have an internal RMSD consensus; for example, 3 solutions to a cluster where all solutions are within 1.5 Å. The *diverse solutions* setting allows for a wider range of potential poses, but will take longer in run-time.

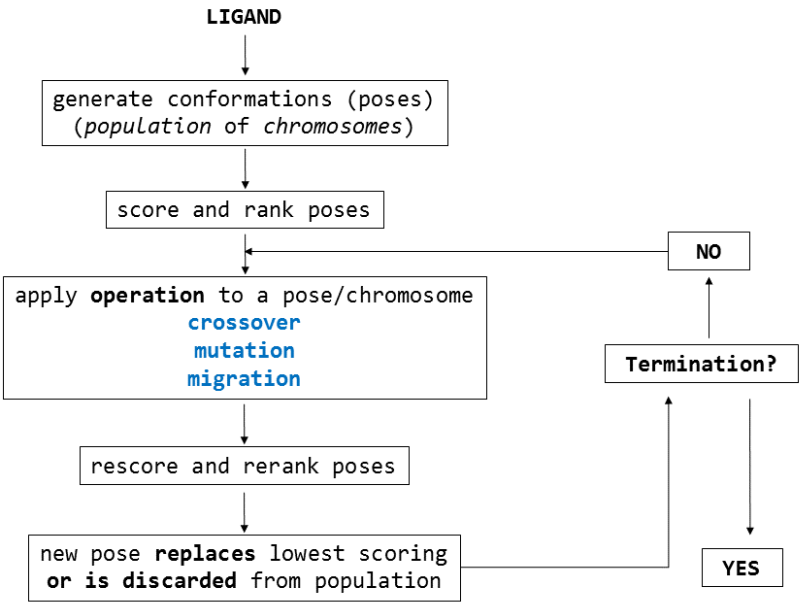


Figure 1.3: Schematic flowchart of the GOLD genetic algorithm.

1.3.1.3 Scoring: GOLD Fitness Functions

GOLD offers four fitness functions for the user to choose from when docking; a force-field function (GoldScore), one is knowledge-based (ASP), and two are empirical scoring functions (Chemscore and CHEMPLP).

Force-field functions assume binding is due to a sum of individual parameter scores and these parameters are themselves individual subsystems; for example, the internal torsional energy term would have no effect on the internal Van Der Waals (VdW) term, despite the torsion of the ligand influencing internal attractions and repulsions.[68]

Knowledge-based fitness functions use databases of known protein-ligand complexes, using the frequency of observations of an interaction to score the posed ligand interactions in context. Contrastingly, empirical functions use protein-ligand structures with corresponding binding affinity data create a scale factor, which is then used on the estimated total free energy. These functions are typically trained by regression to correlate scoring value with binding affinity, making them a prime choice for virtual screening.

Though each fitness function has a particular strength, it is the empirical function CHEMPLP which has been the default function in GOLD for versions 5.0 and later. Here, each of the four functions will be briefly described, with detail for CHEMPLP to highlight why it is the favoured fitness function for virtual screening and lead optimisation on EthR.

GoldScore

GoldScore is, as previously stated, a force-field fitness function. The GoldScore function has been optimised for ligand binding position, rather than for ligand binding affinity, making GoldScore inappropriate for predicting strongly binding ligands or binding constants (unpublished results, CCDC). GoldScore was not used as a scoring function in this work, and so will be only briefly summarised. At its simplest, GoldScore sums four component energy terms: external VdW energy, internal VdW energy, intermolecular hydrogen bond energy and internal torsional energy. Additional components from constrained values can be used, eg. for hydrogen bond contacts. For values such as hydrogen bond energies and atomic radii, empirically derived values are used which are standard to all GOLD fitness functions. The fitness score given for the pose is a negative sum of the component energy terms, with the external VdW term multiplied by 1.375 to encourage hydrophobic contacts between the ligand and the protein.

ASP: Astex Statistical Potential

As a knowledge-based fitness function, ASP uses a database of known complexes to derive atom-atom potentials. Interactions between specific ligand and protein atoms are scored based upon the frequency of their observation in the database set, a set which is typically derived from the PDB. The key advantage of ASP is that it can be targeted to certain protein classes, such as kinases, by changing the database from which the potentials are derived. In this situation, ligands which bind the receptor and make contacts seen in other kinases would score highly, making ASP customisable. However, when not tailored, the ASP effectively takes potentials from a set representative of the entire PDB, and in that case has accuracy comparable to GoldScore and ChemScore.

ChemScore

ChemScore is an empirical function derived from 82 protein-ligand complexes for which binding affinity data is available.[68,69] This particular function estimates the total free energy change based on physical contributions from hydrogen bonding, metal binding, lipophilic interactions and rotational energy, each with a scale factor derived from linear regression of the empirical data. The final ChemScore term includes clash and torsional penalties to discourage close contacts and strain in the ligand conformations. The ChemScore scoring function is therefore scaled by linear regression against known binding affinities to give a correlation between molecular recognition, high scores and predicted high binding affinity. The following breakdown of the ChemScore function is

taken from two papers by Eldridge *et al*.[70] and Baxter *et al*.,[69] as well as the GOLD 5.2 user guide.[68]

ChemScore estimates the change in total free energy according to equation 1.5, where each term is a product of a physical contribution to binding multiplied by the regression co-efficient (equation 1.6). To achieve the final ChemScore value, clash penalty and internal torsion terms are included to penalise close contacts and poor ligand conformations (equation 1.7).

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hbond} + \Delta G_{metal} + \Delta G_{lipo} + \Delta G_{rot}$$

Equation 1.5: The estimation of free energy change upon binding, where each term is a product defined by equation 1.6.

$$\begin{aligned} \Delta G_0 &= v_0 \\ \Delta G_{hbond} &= v_1 P_{hbond} \\ \Delta G_{metal} &= v_2 P_{metal} \\ \Delta G_{lipo} &= v_3 P_{lipo} \\ \Delta G_{rot} &= v_2 P_{rot} \end{aligned}$$

Equation 1.6: The breakdown in terms of equation 1.4, where each energy contribution is a physical term (P, discussed in detail below) scaled by a factor determined by regression on protein-ligand structures (v).

$$\begin{aligned} ChemScore = \Delta G_{binding} + P_{clash} + c_{internal} P_{internal} \\ + (c_{covalent} P_{covalent} + P_{constraint}) \end{aligned}$$

Equation 1.7: The ChemScore value is a combination of equation 1.4, plus clash and internal torsion terms. If covalent docking and/or docking constraints are used, additional terms in brackets are utilised.

To aid in scoring contacts, ChemScore uses block functions, which utilise an *ideal* and a *maximum* allowed value for a contact (such as a hydrogen bond, metal bond, or lipophilic interaction) and reduces the score contribution for deviations from the ideal (scoring 1.0) to the maximum (scoring 0.0). The block function (B) takes the form shown in equation 1.8, which is shown graphically in figure 1.4. This function is sometimes smoothed using a Gaussian function to give the graphical form shown in figure 1.4 (equation not shown).

$$B\ (x, x_{ideal}, x_{max}) = \begin{Bmatrix} 1 \text{ if } x \leq x_{ideal} \\ 1.0 - \dfrac{x - x_{ideal}}{x_{max} - x_{ideal}} \text{ if } x_{ideal} \leq x \leq x_{max} \\ 0 \text{ if } x > x_{max} \end{Bmatrix}$$

Equation 1.8: The block function allows ChemScore to scale contacts which are not ideal. The function B is applied to the P terms in the fitness function (equation 1.9).

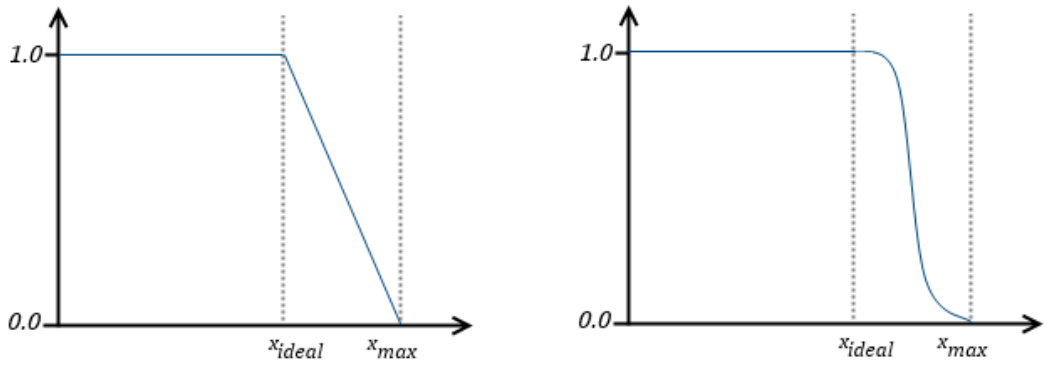


Figure 1.4: Graphical representation of the block function shown in equation 1.8 (left) and after Gaussian smoothing (right).[37,38]

Some of the individual terms will now be discussed, including the metal binding terms, clash and torsion terms, and lipophilic contact terms. Firstly, the hydrogen bond terms are computed over all potential donor-acceptor pairs and each term in summation – distance, angle and directionality – utilises the smoothed block functions above; this reduces the hydrogen bond contribution depending on deviation from the ideal values. This means a hydrogen bond of ideal geometry - with *ideal* hydrogen-acceptor distance, *ideal* D-H···A angle, and *ideal* directionality - would have a value of 1.0. There is no distinction between ionic and non-ionic hydrogen bonds, and water molecules are treat as part of the receptor.

The ChemScore parameters have various default figures for these *ideal* values. For distance, for example, the ideal distance between the donor hydrogen atom and the acceptor atom is 1.85 Å, with a tolerance for remaining *ideal* of 0.25 Å; therefore a hydrogen bond of ideal distance is 1.6 – 2.1 Å. However, the maximum deviation is 0.65 Å, and so a hydrogen bond will score 0.0 at 2.5 Å and beyond.

Similarly, for the hydrogen bond angle of D-H···A, the ideal angle is 180°, with a tolerance of 30°, and as such the ideal range begins at 150°. Hydrogen bonds with heavy atoms (D-H···A-X) have a larger tolerance of 70°. The maximum deviation is 100° and so an interaction of hydrogen with an acceptor with an angle smaller than 100° is not, strictly speaking, regarded as a hydrogen bond. This large tolerance will be discussed again in chapter five, as it becomes relevant to filtering on protein-ligand hydrogen bonds after virtual screening. The hydrogen bond co-efficient (v1, in equation 1.6) is set

to -3.34 as default, determined from Eldridge *et al*.'s development using the 82 complex training set for ChemScore.[70]

Metal binding terms are treat similarly to hydrogen bonds, in that they are calculated over all possible pairs, and the acceptor is a ligand ion capable of metal binding. Equation 1.9 shows the summation, in which all terms are subject to the Gaussian-smoothed block function (B) described previously with a Gaussian smearing sigma value (σ). The ideal distance is 2.6 Å, with 3.0 Å as the allowed maximum binding interaction. The associated regression-determined co-efficient, to bring the calculated values in line with the observed values from the training set, is set as -6.03 as determined by Eldridge *et al*..[70]

$$P_{metal} = \sum_{\substack{All\ ligand \\ acceptors}} \sum_{\substack{All\ protein \\ metals}} B(r_{aM}, R_{ideal}, R_{max}, \sigma_{metal})$$

$$\Delta G_{metal} = v_2\, P_{metal}$$

Equation 1.9: The metal-binding term ΔG_{metal} utilises the regression co-efficient v_2, and the physical contribution P, which sums over all ligand and metal ions with a Gaussian smoothed block function, B. Term r_{aM} represents the actual distance, with the ideal and maximum R distances given.

The lipophilic interaction takes a similar form, though the actual distance of lipophilic interactions is calculated per atom pair and given as r_{ll}. The ideal atom-atom distance is 4.1 Å by default, with a maximum separation of 7.1 Å. The coefficient v_3 is -0.117 as default.[70] The rotatable bond, clash penalty and torsion terms take different forms to define their physical (P) contribution to binding.

The rotatable bond term is used to approximate the loss in entropy which would be incurred by a single, acyclic bond becoming *non-rotatable* upon binding. This was implemented as Eldridge *et al*. found that a simple count of those rotatable bonds which became frozen was unreliable for estimating the contribution of flexibility, and changes in the composition of the training set would cause wide variations in the coefficients required. Their solution was a more complex term (equation 1.10) but a more stable one across the data; Eldridge *et al*. do concede, however, that the term is not "particularly satisfactory" in estimating the actual entropic penalty due to the approximations involved and restricting the definition to effects on rotatable bonds and flexibility without consideration of solvation effects or lipophilic interactions.[70]

$$P_{rot} = 1 + (1 - \frac{1}{N_{rot}}) \sum_{r} \frac{(P_{nl}(r) + P'_{nl}(r))}{2}$$

Equation 1.10: The rotatable-bond freezing term of ChemScore utilises the number of rotatable bonds (N_{rot}) in the ligand and the percentages of non-hydrogen atoms on either side of the rotatable bond that are not lipophilic.

The regression coefficent associated with the rotatable bond term has the default value in the ChemScore parameter file of 2.56.

There are three clash terms for different, previously described, interactions: one for hydrogen bonds, one for metal interactions, and a third for all other ligand protein interactions.[69] The purpose of the clash terms, in conjunction with the internal torsion term, is to prevent poor geometries in docking. Above, distances for interactions were given as ideal and maxima; the clash terms come into play when distances are shorter than the ideal. There are two values of r_{clash}: for contacts to protein sulphur atoms, the ideal distance is 3.35 Å, for all other contacts the ideal is 3.10 Å;[69] the ideal values for metal and hydrogen bond interaction distances have already been given. The three equations for hydrogen bond, metal interaction and 'other' clash terms are given below as equations 1.11, 1.12 and 1.13 respectively.

$$P_{clash-hbond} = \frac{20.0 \times (r_{hbond} - r)}{\Delta G_{hbond} \times r_{hbond}}$$

Equation 1.11: The hydrogen bond clash term for the interactions with a distance shorter than r_{hbond}. The value of 20.0 is an empirically-derived weighting term to notably penalise the overall score value. Without the weighting term, values of P_{clash} would be too small to impact the fitness in a discriminatory way.

$$P_{clash-metal} = \frac{20.0 \times (r_{metal} - r)}{\Delta G_{metal} \times r_{metal}}$$

Equation 1.12: The metal interaction clash term for the interactions with a distance shorter than the ideal.

$$P_{clash-other} = 1.0 + \frac{4.0 \times (r_{clash} - r)}{r_{clash}}$$

Equation 1.13: The other clash term, for which the ideal distance varies depending on the contact type. The weighting term here is smaller to allow for some close non-polar contacts but performs the same role as in the other clash terms.

The internal ligand torsion term uses pairs of atoms in rotatable bonds and their hybridisation states, along with clash terms to approximate ligand energy. The torsion

term is given as equation 1.14 below, where A, n and Φ vary for sp^3-sp^3, sp^3-sp^2 and sp^2-sp^2 bonds.[69]

$$P_{internal} = \sum_{\substack{All\ rotatable \\ bonds}} A_i\,(1 - \cos(n\Phi - \Phi_0)$$

Equation 1.14: The internal ligand torsion term depends upon the bond type for values of A, n and Phi.

Finally, terms for covalent bonding adapt the already described terms by (a) reducing the clash term, (b) adding torsion terms for the rotatable part of the protein-ligand covalent bond, and (c) adding a valence-angle bending term to penalise poor covalent linkage geometries.

All values discussed here are part of the ChemScore parameters file, which can be customised by editing the text file and instructing GOLD to use the altered version. However, all default terms were derived from extensive parameterisation by the program developers and as such rarely require changes; in the course of this work, no changes were made to these defaults.

CHEMPLP

CHEMPLP, derived from the *piecewise linear potential* (PLP) and select bonding terms from ChemScore, is the default scoring function in GOLD.[68,71] Like ChemScore, CHEMPLP is an empirical function (equation 1.15).

$$fitness_{PLP} = -\big(w_{PLP} \cdot f_{PLP} + w_{lig-clash} \cdot f_{lig-clash} + w_{lig-tors} \cdot f_{lig-tors} + f_{chem-cov} + w_{prot} \cdot f_{chem-prot} + w_{cons} \cdot f_{cons}\big)$$

$$fitness_{CHEMPLP} = fitness_{PLP} - (f_{chem-hb} + f_{chem-cho} + f_{chem-met})$$

Equation 1.15: GOLD's CHEMPLP scoring function utilises the PLP fitness, adjusted with ChemScore terms. PLP is used to model steric complementarity between the protein and the ligand; *lig-clash* is a heavy-atom clash penalty; *lig-tors* is a ligand torsion potential; *cons* handles constant contributions. Other terms (*chem-cov* for covalent docking and *chem-prot* for flexible side chains and waters) are included as necessary. Terms from ChemScore (hydrogen bonding, CHO intereactions and metal bonding) add additional scoring terms.

The PLP terms determine the steric complementarity between the ligand and the protein, using simple atom type designations: *donor*, *acceptor*, *donor/acceptor*, *nonpolar*, and *metal.* The interactions between protein and ligand atom types are then characterised according to table 1.1.

Table 1.1: PLP interaction types in the CHEMPLP scoring function.

	protein atom type				
ligand atom type	*donor*	*acceptor*	*don./acc.*	*nonpolar*	*metal*
donor	repulsive	hydrogen bond	hydrogen bond	buried	repulsive
acceptor	hydrogen bond	repulsive	hydrogen bond	buried	metal
donor/acceptor	hydrogen bond	hydrogen bond	hydrogen bond	buried	metal
nonpolar	buried	buried	buried	nonpolar	buried

Interactions between protein and ligand, as defined in table 1.1, depend on the type of atom. Metal atoms chelate only acceptors and atoms capable of acting as acceptors (or donors, ie. don./acc), and hydrophobic interactions can occur between two nonpolar atoms; however, polar atoms are typically buried by nonpolar atoms. Hydrogen bonds occur between two oppositely polar atoms, otherwise repulsion occurs.

CHEMPLP utilises ChemScore terms for hydrogen and metal bonds, and additionally terms for internal clash and torsional potentials, all described previously. A term is added in the presence of C-H donors, which can interact with oxygen acceptor atoms – this is the CHO potential. Terms can be added to CHEMPLP for covalent docking, flexible side chains and water molecules in addition to restraints (such as on hydrogen bond formations), making CHEMPLP highly customisable.

CHEMPLP was evaluated against the other fitness functions in GOLD and shown to be fastest - slightly faster than ChemScore on average docking times - and more successful in reproducing known binding conformations in a 298-complex test set within 2 Å: ChemScore up to 67%, GOLDScore up to 79% and CHEMPLP to 87%.[68,71] Overall, and across various speed and accuracy scenarios, CHEMPLP outperformed ChemScore and GOLDScore with notably higher success rates at lower average search times, and so was implemented as GOLD's default function.[51]

1.3.1.4 GOLD Optional Docking Constraints

GOLD is able to constrain various features, including distances, hydrophobic regions, and hydrogen bond constraints, depending upon the docking required

There are two kinds of hydrogen bond constraints, one which specifies both ligand and protein atoms, and one that specifies only a protein donor atom. The former 'Protein HBond' setting is not suitable for virtual screening, but the latter 'Hydrogen

Bond Constraint' allows ligands which make a favourable interaction to score more highly. The constraint itself is incorporated into the fitness function, with a user-defined minimum geometry weight score, which determines how good the interaction has to be to be considered a hydrogen bond between 0 and 1. This is defaulted at 0.005, meaning even weak hydrogen bonds are considered. A constraint weight is also used to penalise poses which do not form the hydrogen bond; this is deducted from the fitness score.

1.3.1.5 Protein and Ligand Flexibility in GOLD

Even at a basic level, GOLD does not treat the protein as completely rigid; this is a distinction from other software for docking such as Autodock Vina[72] and FlexX.[36] Though GOLD offers the possibility to enable flexible side chains, dock into multiple crystal structures (ensemble docking), or use soft potentials to compute flexible loops, when these functions are not in use, some flexibility is applied to the binding pocket regardless. Hydroxyl groups on serine, threonine and tyrosine are rotated during docking and optimised, maximising hydrogen bonding potential; NH_3^+ groups on lysine residues undergo the same treatment.

Flexible residues are allowed to undergo torsional rotations, which are distinct from the simple hydrogen donor optimisation which happens for Ser, Thr, Tyr and Lys residues. Side chains only move around acyclic bonds, and the specific rotamers must be defined. Rotamers can be defined by a library of crystallographically observed examples, or by defining specific ranges of torsional angle values. A maximum of ten flexible residues can be allowed in GOLD, as they increase the search space and computational strain. They can increase the number of false-positive results, therefore flexible residues should be used sparingly and only with absolute confidence.

Two specialised settings allow for more detailed consideration of protein flexibility: using soft potentials, and using ensemble docking. The former uses altered Van der Waals contributions to account for loop movements; 'ensemble docking' docks ligands to multiple protein models, which can take into account backbone movements.

GOLD does not alter ligand bond lengths or valence angles; this necessitates an optimal input model. Similarly, stereochemistry is not altered and therefore all stereoisomers should be generated and docked individually. Ideally small-molecule crystal structures would be used, however this is not always possible (and for libraries not including the CSD, likely impossible). In the course of the docking, the ligand conformation is changed to find the steric fit between that and the protein binding pocket. GOLD offers various options to fix or rotate certain parts of a molecule, including: detecting internal hydrogen bonds (both in the protein and the ligand); matching ring

conformations to templates from the CSD for accurate ligand geometries; rotating protonated carboxylic acids for donor and acceptor function; flipping *cis-trans* amide bonds and planar nitrogens; and fixing all, terminal, or specific rotatable bonds. This gives GOLD a full range of ligand flexibility options for docking.

1.3.2 Chemoinformatics: KNIME

The size and complexity of available small-molecule databases - such as ZINC which contains over 35 million compounds,[73] and the CSD[43] which contains over 700,000 crystal structures – would make virtual screening a monumental task of time and resources if not for our ability to apply our knowledge of what features constitute a good lead, fragment or drug in pre-processing filters. To aid in this, pipelining and data mining software has become invaluable.

KNIME is an open-source data mining and pipelining software.[74] In KNIME, nodes representing Java scripts are assembled in a user-friendly graphical user interface (GUI), into a workflow pipeline which parses data in tables.

For chemoinformatics, KNIME is especially powerful. Community contributions include a plethora of packaged chemistry nodes including CDK,[75] RDKit, nodes for the R statistical package, and nodes for adding user scripts in MATLAB and Python, among others.

For enacting on chemical structures 1D, 2D and 3D structures can be read-in to KNIME; for 3D structures the most common formats used are SDF or SYBYL MOL2 and, when written out from KNIME, additional data can be included in tags within SDF files. As the SDF format was used extensively, an annotated example can be found in Appendix A, the product of the example pipeline in figure 1.3.

An example of a KNIME pipeline is show as figure 1.5, in which the structure of testosterone is imported in the MOL2 format (previously retrieved by name and exported using Avogadro) and metrics for Lipinski Rule-of-Five compliance, XlogP and a number of other metrics (under the *Molecular Properties* node) are calculated. Testosterone is then exported in MOL2 format (identical to input) and SDF format, which retains calculated properties as additional tags in the file. Not shown are nodes which can filter columns and rows, nodes which concatenate tables, and nodes capable of complex clustering functions. Where relevant, these will be discussed in detail in context.

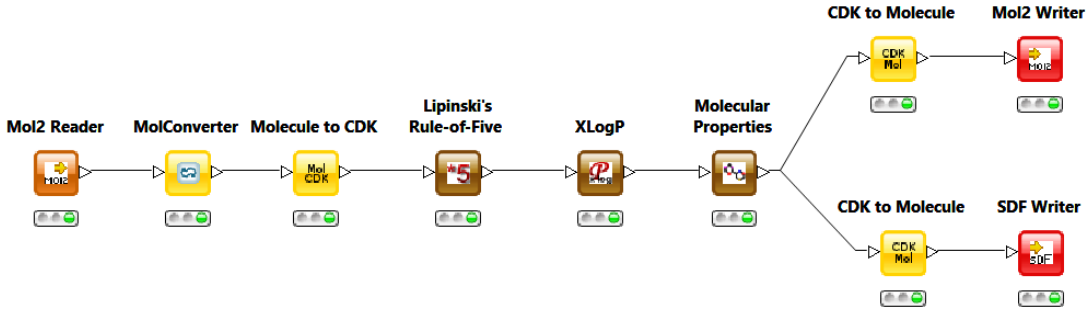


Figure 1.5: Example KNIME pipeline for chemoinfomatic data calculation. The input MOL2 file is converted first to SDF format, then to a proprietary CDK format[75] for the calculation of Lipinksi RO5 compliance and XLogP in standalone nodes, and a list of different properties in the Molecular Properties node (including volume, polar surface area, ALogP, and others). Finally, the table is converted into MOL2 format (which does not keep the properties in appended tags) and SDF format (which does).

1.3.3 Chemical Space: Mogul and the CSD

The Cambridge Structural Database[43,76] is the primary repository for small molecule crystal structures (and associated data). Currently containing over 700,000 small molecule crystal structures, with nearly 40,000 added annually, the CSD is a peerless archive of structural data. This wealth of information can be accessed, mined and applied using the CSD's associated software package; of particular relevance here is the knowledge-based library Mogul, which uses the structural data in the CSD to generate distributions of preferred molecular geometries.

Mogul has many applications, from generating restraint libraries for refining crystal structures, to validate conformations of deposited co-crystal ligand structures,[77] and for the design of new leads in drug development programmes. In the course of computational protein-ligand modelling Mogul, as a quantitative and rapid method of pose-evaluation, can be especially powerful for the assessment of docked ligand conformations.

1.4 Implementation of SBDD Techniques

Virtual screening and docking trace their roots to the early eighties[49] and, given how long drug discovery and development can take, we are only beginning to see the fruits of those labours exiting the clinic.[78] Over the last few decades papers and citations of virtual screening work has increased exponentially (figure 1.6), showing a field which is still refining its approach but beginning to yield some positive results.

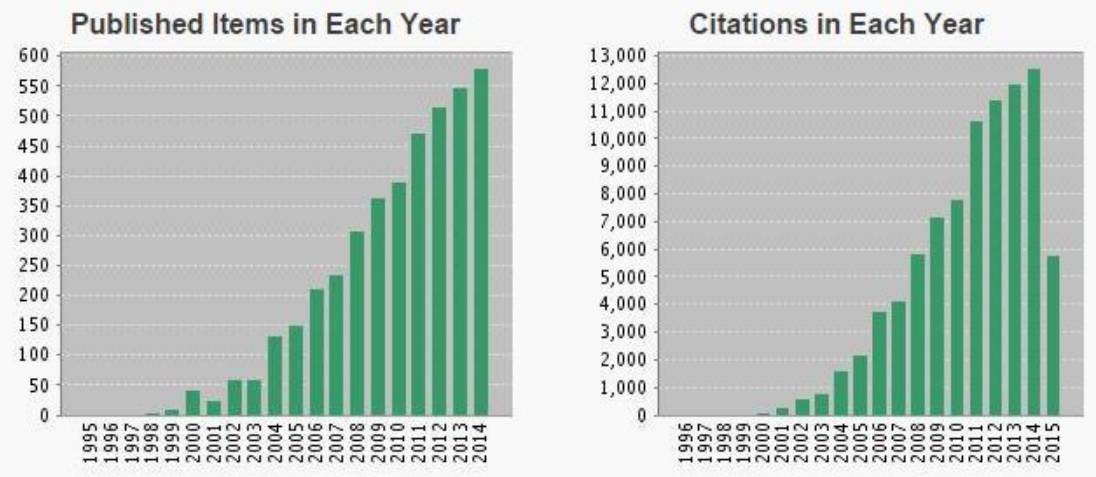


Figure 1.6: Data compiled from ISI Web of Knowledge. Graph left shows articles published using the term "virtual screening" to end of 2014 and graph right shows citations to June 2015 of those articles using the same term. Both show a dramatic increase since 1995.

There were some early successes, particularly in the antiviral field, however the drug widely accepted to be the first to be derived from structure-based drug design methods is the anti-glaucoma drug dorzolamide.[79]

Dorzolamide is an inhibitor of carbonic anhydrase, an enzyme discovered in 1932. In the mid-1980s Merck developed a topically-active carbonic anhydrase inhibitor with sub-nanomolar activity, able to reduce fluid secretion in the eye.[80] Utilising molecular modelling and X-ray crystallography structures, they added lipophilic substituents which were able to control steric and energetic problems, resulting in dorzolamide (K_i = 0.37 nM). The structure of dorzolamide is given as figure 1.7, and in its first year on the market, 1996, it was the most widely prescribed anti-glaucoma product in the US.

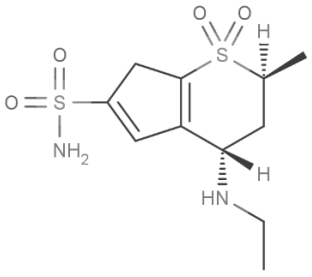


Figure 1.7: Chemical structure of dorzolamide, an inhibitor of carbonic anhydrase for glaucoma.[81]

As a topical medicine (utilised in eye-drops),[81] and one developed before Lipinski *et al.*'s paper on oral bioavailability rules, it does not conform to the "rule of five". However, what is particularly interesting about dorzolamide is the sulphoxide oxygens form a hydrogen bond to a glutamine residue (Gln92) in the binding site (figure 1.8).

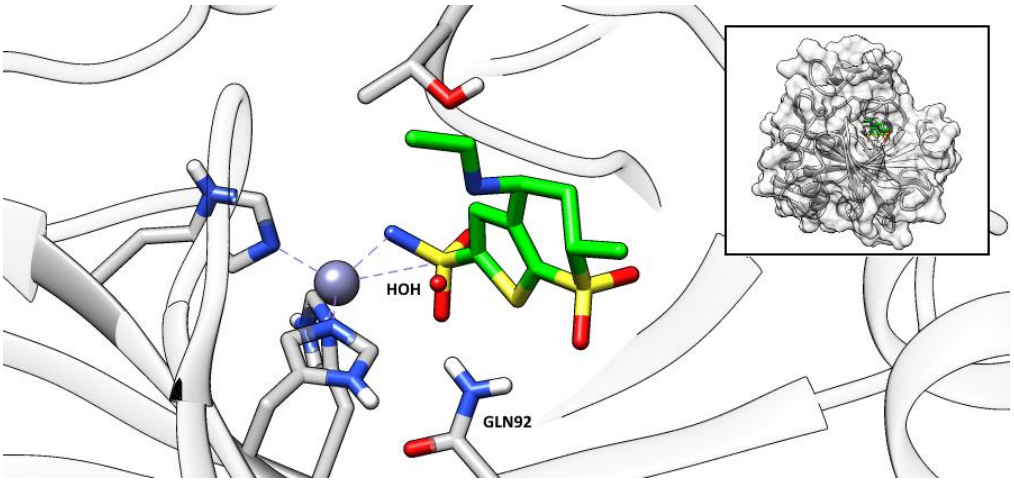


Figure 1.8: Dorzolamide bound to carbonic anhydrase II (inset), hydrogen bonding with glutamine residue and a water molecule, as well as co-ordinating with bound zinc.[82] It occupies the same position and is oriented similarly to its progenitor compound derived from docking studies.[80]

A molecule with this particular feature would, potentially, not be found or developed by design in current thinking,[13,83] as sulphoxides are poor hydrogen bond acceptors and can be metabolically problematic. That stated, dorzolamide was derived from previously-known and potent inhibitors, and is a topical rather than oral treatment, reducing the impact on the liver and so metabolic activity. This perhaps highlights the restrictive conventions which have developed over the years in drug development, and which may contribute to the overall dissatisfaction with drug development success.

Another widely known example is the development of neuraminidase inhibitors for the treatment of influenza. The first neuraminidase structure was solved to 2.9 Å in 1983 by Graeme Laver and colleagues.[84] Zanamivir[85,86] was extensively prescribed during the 2009 H5N1 'swine flu' outbreak as Relenza. Zanamivir mimics a transition state of the neuraminidase substrate N-acetylneuraminic acid and was identified using GRID, which scores the energy between a functional group probe and a protein, which is mapped onto a three-dimensional grid of a particular mesh size (1.0 Å in this case). This determined an amino group beneath an aspartate residue (Asp151) to be a key hotspot, and a transition state mimic was designed to occupy that site and form a hydrogen bond with the amino group.[86,87] Co-crystal structures of zanamivir have been published for influenza A subtypes N1,[88] N2,[89] N3,[90] N9[87,91] and influenza B.[92] Zanamivir bound to neuraminidase subtype 1 from the 2009 H1N1 is shown as figure 1.9.[89]

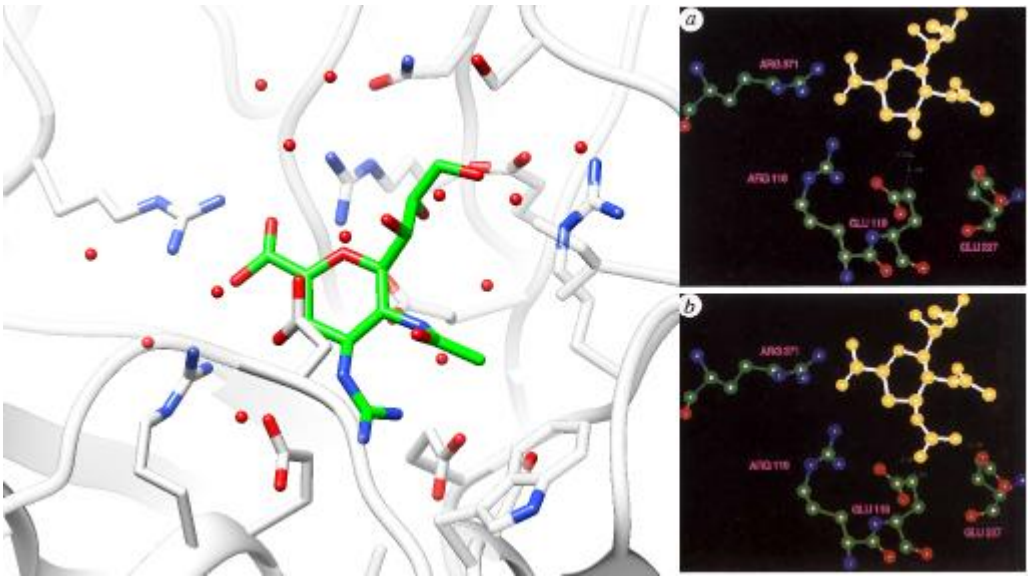


Figure 1.9: Zanamivir (green) bound to N1 from 2009 H1N1[89] alongside the original docking poses of *a*, 4-amino-Neu5Ac2en, and *b*, 4-guanidino-Neu5Ac2en, the derivatives of N-acetylneuraminic acid. Docking poses are reproduced from figure 1 of von Itzstein *et al.*[86]

Presented here are just two examples where early implementation of structure-based drug design yielded drugs for clinical use, without touching upon the plethora of HIV successes. To conclude, structure-based drug design by computer-aided methods have shown increasing viability, versatility and clinical success.

1.5 Conclusion

In comparing and contrasting various approaches to docking small molecules to proteins, it is clear these techniques can be useful tools for drug discovery. Docking studies now utilise algorithms capable of considering ligand flexibility as standard, protein flexibility to varying degrees, and essential co-factors and water molecules in the search for good scoring poses.

The techniques and approaches described in this chapter have been implemented in a drug discovery project for a qualified protein target EthR from *Mycobacterium tuberculosis*, one of the causative agents of TB. Unlike development of a new antibiotic, this project takes advantage of an antibiotic already present on the market, and is intended to act as a co-drug with the aforementioned antibiotic to improve its efficacy. This work is presented, culminating in multiple hits for optimisation.

CHAPTER II:
ETHR: A VIRTUAL SCREENING TARGET FOR TB CO-DRUGS

Tuberculosis, commonly referred to as TB, is a global health problem which claimed the lives of an estimated 1.5 million individuals in 2013[93] and, though often thought to be a relic of the pre-WWII world, it in fact remains second only to HIV as the leading cause of death by an infectious disease. TB also presents a growing drug-resistance problem worldwide, with 3.5% of the 9 million newly reported cases of TB in 2013 being classified as multi-drug resistant.[93]

The DOTS (directly-observed treatment short course) was introduced and refined through the 1990s to manage TB treatment, with an emphasis on commitment of governments, non-governmental organisations and care providers at all levels to deliver a regular drug supply, administration of which was directly observed by a health worker.[94] DOTS also included a standard daily antibiotic regimen of, for drug susceptible TB, four drugs taken over six to eight months: isoniazid, rifampicin, pyrazinamide and ethambutol.[95] The first-line drugs were all developed and put into use in the 1950s and 1960s, since which time no significant antitubercular drug has been found in replacement, however there are second-line recourses when the primary treatment fails.

Multi-drug resistant tuberculosis (MDR-TB)[96] is classified as a resistance of *Mycobacterium tuberculosis*, the causative bacterium, to isoniazid and rifampicin. According to the World Health Organisation's latest annual global report on TB, MDR-TB has highest prevalence in Eastern Europe and central Africa (figure 2.1), and the 24-month second-line treatment was implemented in 82% of those cases eligible in 2012.

Unfortunately, success rates of the second-line treatment were only 48% in 2012 (of those MDR cases detected in 2010), owing to problems of access, patient compliance, and follow-up by health services. These factors have led to the recent emergence of *extensively drug resistant* TB (XDR-TB).[97] XDR-TB was detected in 92 countries in 2012, with the highest incidence in Eastern Europe (figure 2.1). It is estimated that 9.6% of MDR-TB cases in 2012 are in fact XDR-TB, with 32% of MDR cases being resistant to either a fluoroquinolone, or a second-line injectable antibiotic (such as kanamycin) or *both*; if an MDR-TB case is resistant both second-line alternatives such as these, it qualifies as an XDR case.

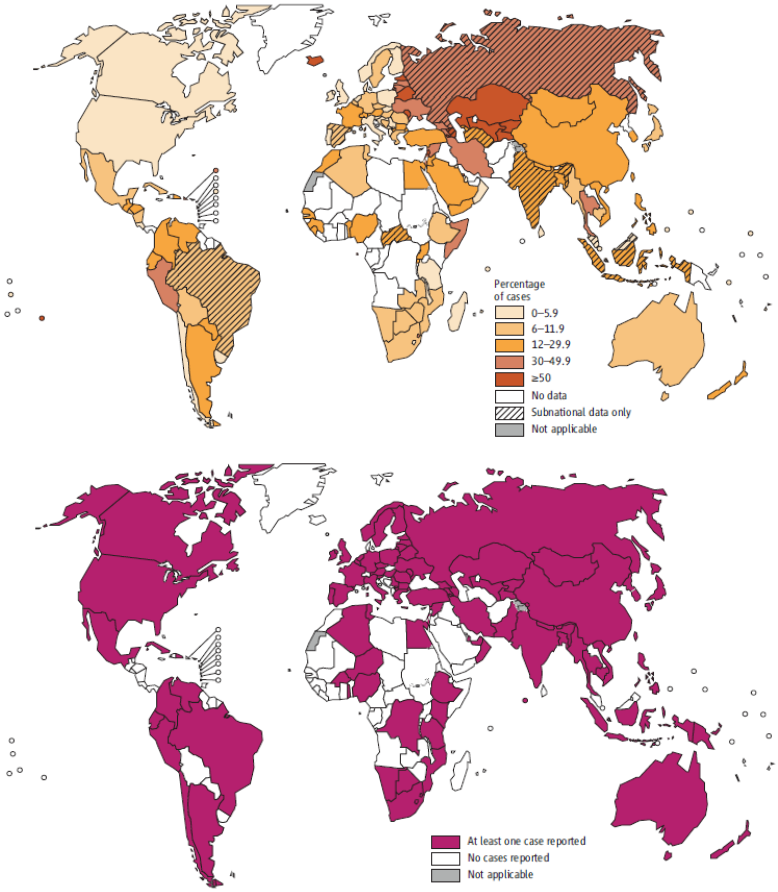


Figure 2.1: Top, percentage of TB cases which became multi-drug resistant and thus notified as MDR-TB. Data shown is for the most recent year for which data was available, which varies among countries. Bottom, countries which reported at least one case of XDR-TB in 2012 according to WHO.[98]

The role and classification of second-line antibiotics is two-fold. Firstly, they are held in reserve to prolong their efficacy, being applied only when the cheaper and more widely used antibiotics fail thus allowing our arsenal of drugs to remain useful for as long as possible.[99] Secondly, these second-line antibiotics often have poorer patient tolerance than those in the first line, and so administration is undertaken only if absolutely necessary.[100,101] As MDR-TB and XDR-TB rise, more patients will necessarily have to turn to these for treatment taking from 12 to over 24 months.[102] One such poorly tolerated second-line drug is ethionamide,[103] a structural analogue of the front-line isoniazid (figure 2.2).[104]

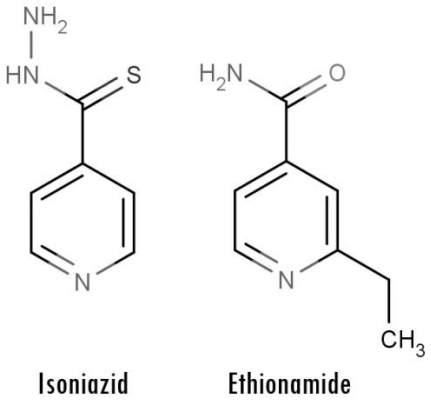


Figure 2.2: Chemical structures of isoniazid and analogue ethionamide.[104,105]

Both are type I pro-drugs and they share a common target: when activated within the mycobacterium isoniazid and ethionamide each form an inhibitory adduct with NAD against InhA,[106] a carrier protein which is part of the fatty acid elongation system (FAS-II) which produces mycolic acids for the *M. tuberculosis* cell wall.[107] Both isoniazid and ethionamide therefore disrupt cell wall formation, but they have separate and distinct activation mechanisms.[108] The bioactivation of isoniazid by the protein KatG[109] is efficient such that low therapeutic doses are sufficient in treatment; the same cannot be said of ethionamide and its activator. Ethionamide is required in very high doses (typically 10-20mg/kg daily, compared to isoniazid dose of 5mg/kg during active infection) and as such, results in a variety of negative side effects including (but not limited to) hepatotoxicity and mental disturbances.[105,110]

2.1 Target Discovery

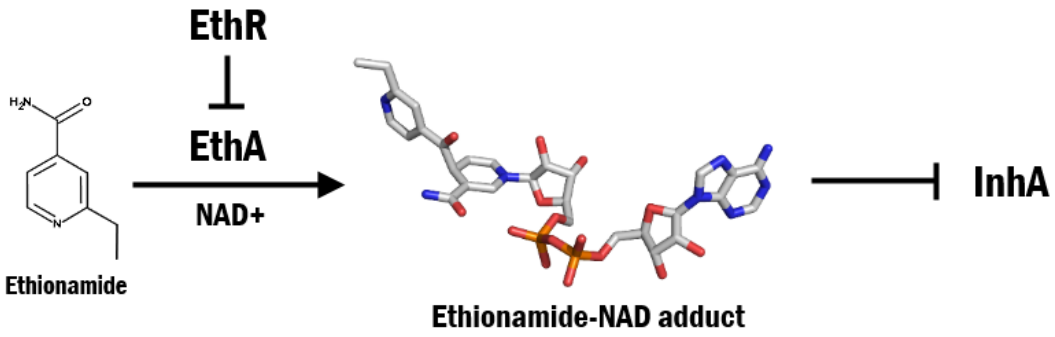


Figure 2.3: A summary of the mechanism by which ethionamide acts in M. tuberculosis – ethionamide is activated by EthA (expression of which is controlled by EthR) and forms an adduct inhibitor of InhA, preventing fatty acid synthesis. [106,107,111,112]

In 2000, Baulard *et al.*[108] published their discovery of the ethionamide activator protein which they called EthA (figure 2.3), and also noted the presence of a neighbouring open

reading frame which was homologous to known members of the TetR family of transcriptional repressors. In their paper Baulard *et al*. demonstrated that over-expression of *ethA* caused hypersensitivity of *M. tuberculosis*, *M. bovis BCG* and *M. smegmatis* to ethionamide, while over-expression of the neighbouring putative transcriptional repressor conferred resistance. Thus they proposed EthA to be the activator of ethionamide, and the neighbouring transcriptional repressor to control *ethA* expression; they named this TetR-like transcriptional repressor EthR, and made the first suggestion that EthR could be a viable target for co-operative treatment with ethionamide (ETH).[108] The first crystal structures of EthR were published in 2004, determined by two distinct research groups.[111,113] Figure 2.4 shows the repressor in its homodimeric form, and the monomer from the two independently determined crystal structures..[111,113]

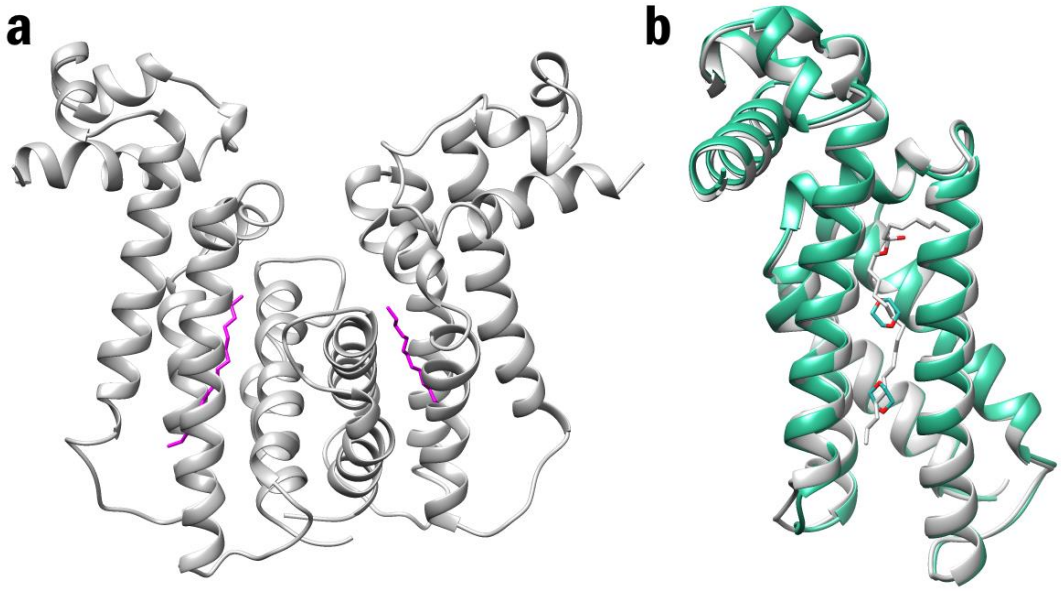


Figure 2.4: Crystal structures of EthR. Left, the homodimer (PDB: 1U9O).[111] Right, overlay of the monomeric EthR (PDB: 1U9N in grey, PDB: 1T56[113] in green.) Ligands hexadecyl octanoate (grey) and dioxane (green) are shown in **b** for comparison.

As expected, EthR showed the same canonical fold as other TetR-like repressors and high similarity within the N-terminal DNA binding region (a helix-turn-helix motif)[114] but little among C-terminal ligand-binding domain.[115] In the case of EthR, this ligand-binding domain is formed by five helices which together make a long, linear channel in the domain. In the crystal structures published by both research groups this channel was occupied. Dover *et al.*[113] determined two six-membered rings (dioxane, components of their crystallisation buffer, shown in figure 2.4) within the channel of their structure;

however, Frénois *et al*.[111] fortuitously co-crystallised hexadecyl octanoate, a long hydrophobic ester, within the ligand-binding channel.

As the structure was determined from crystals of recombinant EthR produced in *Escherichia coli*, the ligand is not a natural one (though it may resemble in some fashion the natural ligand, as EthA catalyses the conversion of ketones to their esters; the natural ligand is currently unknown).[116] This demonstrated the ability of EthR to bind a lipophilic molecule. Interestingly, EthR has to-date only been crystallised in forms which are inactive, geometrically incompatible with DNA-binding. The major grooves of B-DNA are spaced at intervals of 34 Å,[117] yet in the case of EthR the HTH motifs are situated 18 Å too far apart to contact these grooves correctly for function.[118] It has been proposed by Frénois *et al*.[111] that this is exemplary of a natural inhibitory mechanism of EthR, by which ligand binding induces a conformational change in the protein rendering EthR unable to bind DNA; this is further supported if the ligand is an ester produced by EthA, constituting a method of feedback to EthA's transcriptional regulator. This putative mechanism together with a ligand-bound crystal structure paved the way for validation of EthR's potential as a drug target for co-operative ethionamide treatment.

2.2 Target Validation & Screening

Between 2004 and 2009 work was undertaken to find a line of drug-like ethionamide boosters which culminated in the Willand *et al.* paper in Nature Medicine.[119] It was noted that the ligand-binding cavity occupied by hexadecyl octanoate has a largely symmetrical profile, with hydrophobic and aromatic residues lining the upper and lower regions with a polar "patch" in between. In the Frénois *et al* structure of EthR,[111] a water molecule forms a hydrogen bond with the nitrile of an asparagine residue; in the Dover *et al.* structure, the hydrophobic regions are populated by six-membered rings.[113] Therefore, this suggested the observed inactive conformation of EthR could be induced by small lipophilic molecules which could equally exploit the hydrophobic and polar regions of the binding channel.

Willand *et al.* developed a simple, <500 Da pharmacophore model of two hydrophobic ends connected by a hydrogen bond-capable linker of 4-6 Å. A series of 131 compounds were synthesised based on sampling the diversity of the hydrophobic ends and the nature and length of the spacer. Various chemoinformatic properties were calculated for these 131 compounds (cLogP, polar surface area, number of hydrogen bond donor/acceptors) and the best compromises according to Lipinski's rule of five were retained.[120]

An SPR (surface plasmon resonance) assay was used to determine if, and quantify how well, the compounds inhibited the ability of EthR to bind DNA.[118,121] Willand *et al.* found this screen highlighted a series able to reduce the EthR-DNA interaction by more than 50%; this was exemplified by the compound BDM14500 (IC_{50} of 38 μM), which was subsequently co-crystallised with EthR (figure 2.5).[119]

As anticipated, the EthR-BDM14500 co-crystal structure was in a conformation incompatible with DNA-binding (HTH distance 50.3 Å). BDM14500 was shown to have no bacteriotoxic effect alone, but was able to inhibit *M. tuberculosis* H37Rv growth on solid agar plates at a quantity of 20 nmol BDM14500 with 2 μg ethionamide, where double that dose of ethionamide alone was unable to do so; this demonstrated the synergistic, co-drug effect and provided a demonstrable EthR inhibitor hit for optimisation.[119]

2.3 Lead Optimisation of EthR Inhibitors

Analogues of BDM14500 were prepared by Willand *et al.*[119] and evaluated using the previously used SPR assay. Two of these, BDM31343[122,123] and BDM31381, were found to have IC_{50}s of 3.3 μM and 522 nM respectively.

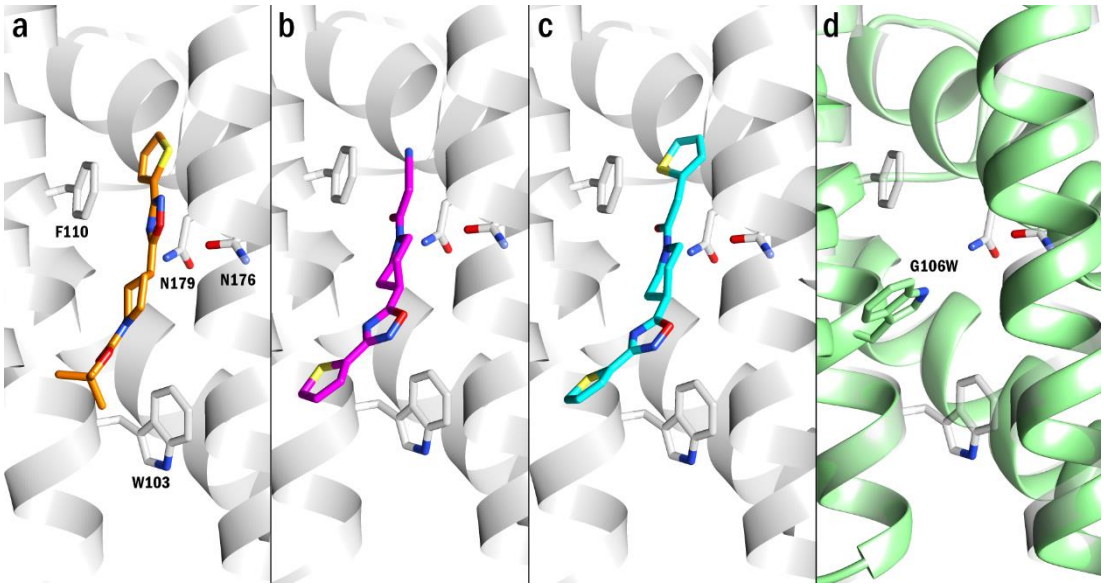


Figure 2.5: Ligands BDM14500 (**a**),[119] BDM31343 (**b**)[123] and BDM31381 (**c**)[119] bound to EthR. Panel **d** shows the location of the G106W mutant which induces an inhibitory conformation.[123]

Both inhibitors were able to boost ethionamide efficiency, BDM31343 by a factor of 10 and BDM31381 by a factor of 20. BDM31343 had the best pharmacokinetic profile and boosting activity *in vivo*, allowing a three-fold decrease in ethionamide dosage. Willand *et al*. noted that although this is already a substantial decrease which would assist in

reducing ethionamide's negative side effects, they anticipated they would be able to improve on their design.

The crystal structure of EthR bound by BDM31381 and a later structure of EthR with BDM31343 both demonstrated the previously observed DNA-binding incompatibility, and by binding in the same region of the cavity (not shown) the allosteric control theory gained more corroborating evidence. In 2011, Carette *et al.* were able to induce this inhibited conformation by mutating a single residue (G106W) in the ligand-binding channel (figure 2.5).[123]

Notably, Willand *et al.*[119] observed that in contrast to BDM14500, BDM31343[123] and BDM31381 oriented in the binding channel such that it was their carbonyl oxygen rather than the oxadiazole which formed the hydrogen bond with Asn179. This would come to be frequently observed, and further docking studies using BDM31343 and two other analogues predicted the ability of EthR to support two binding modes within the linear ligand-binding channel, centred around this important, hydrogen-bonding Asn179 residue.[122]

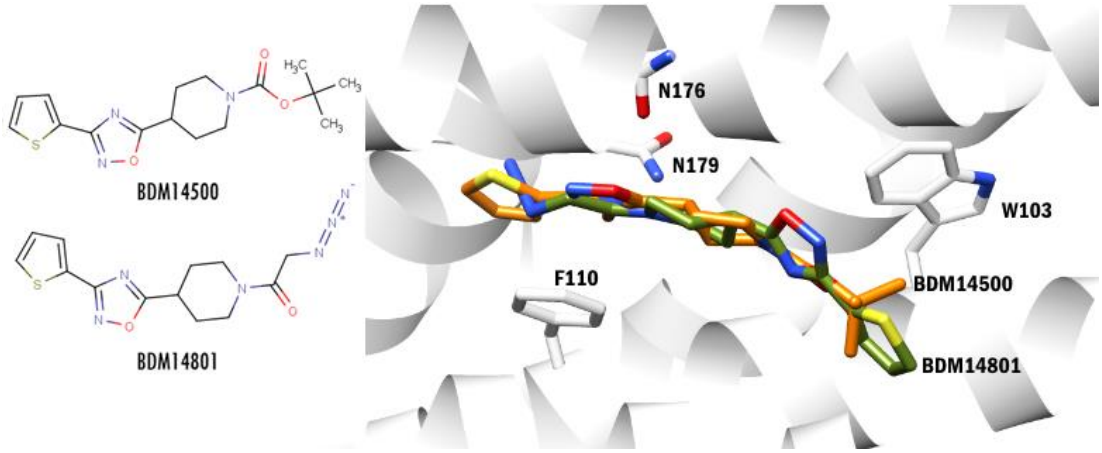


Figure 2.6: Left, chemical and co-crystal structures of the original EthR inhibitor hit BDM14500 (right, orange), and the analogue for target-guided synthesis BDM14801 (right, olive).[119,124]

In order to probe the plasticity of the ligand-binding channel, and determine if new protein conformations could be exploited in ligand design, a technique called *kinetic target-guided synthesis* was used.[124,125] In target-guided synthesis, protein is incubated with an inhibitory compound and reactive fragments to irreversibly form new products *in situ*. An analogue of BDM14500 named BDM14801 was used for the *in situ* reaction (figure 2.6) to encourage elongation of the ligand at the carbonyl tail region. The protein was incubated in six aliquots with clusters of ten aromatic and aliphatic alkynes each such that condensation reactions between substrates would yield products of notably different molecular weights. The small alkynes were selected primarily for their

potential ability to interact with the hydrophobic regions of the EthR ligand-binding channel, specifically Phe114, Phe184 and Trp138 at the top of the channel. The major kinetic product was identified and classified BDM14950 (figure 2.7) and determined by SPR to have an IC_{50} of 580 nM.

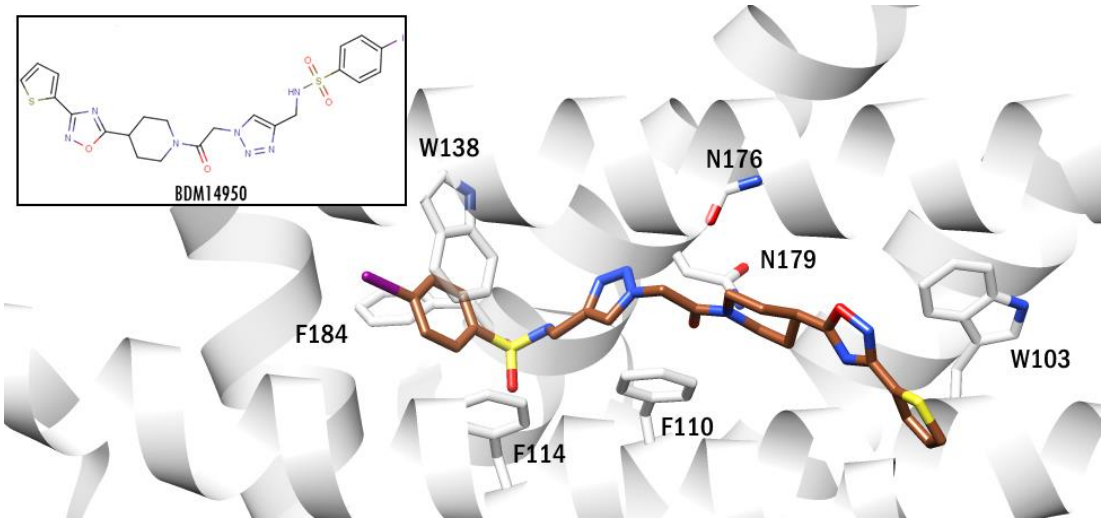


Figure 2.7: Chemical structure of the target-guided synthesis product BDM14950, and as bound in the co-crystal structure PDB:3O8H.[124]

The crystal structure of BDM14950 bound to EthR showed Phe184 and Phe114 had undergone re-arrangement (compared to previous structures with shorter ligands) to accomodate the longer BDM14950; this is despite phenylalanine being one of the amino acids with the lowest propensity towards flexibility within protein structures.[126] Willand *et al.*[124] note in their presentation of BDM14950 that due to the rearrangement of residues required to accommodate the longer ligand, traditional docking and virtual screening methods may only have been able to probe the lower region of the ligand binding channel. This will be discussed further in the next chapter.

Subsequently, structure-activity relationships were established by Flipo *et al.*,[127,128] which varied the carbonyl tail or oxadiazole head regions of this series further. Variations in the tail region showing the best improvement of potency without compromising solubility came from aliphatic chains, with emphasis on the introduction of fluorinated alkyl groups; Flipo *et al.* also investigated replacing the piperidine six-membered ring for five-membered pyrrolidines.[127] The second study from Flipo *et al.*[128] varied head groups instead and yielded the current lead compound in this series. BDM41906, which has an IC_{50} of 400 nM and occupies the lower region of the EthR ligand-binding channel, incorporates a fluorinated alkyl chain tail and a thiazolyl head ring (figure 2.8). Presently, this compound is undergoing further tests in mouse models.

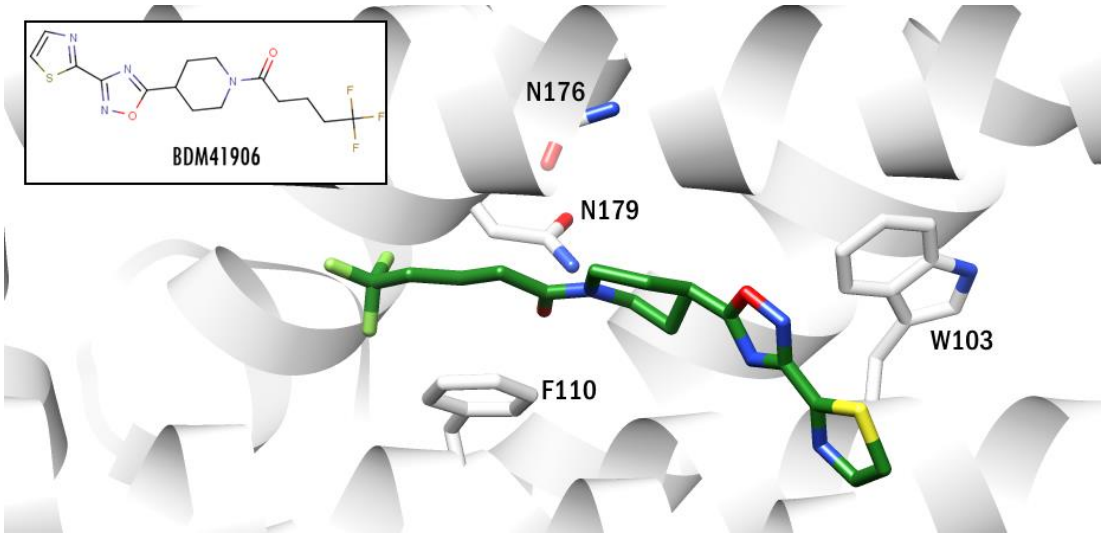


Figure 2.8: The chemical structure of BDM41906, and as bound to EthR.

2.4 Further Drug Development on EthR Inhibitors

Until 2012, published development of EthR inhibitors was limited to one series and one research group. However in 2012, Flipo *et al.*[129] presented a new line of inhibitors which were identified from a whole-cell phenotypic assay of over 14600 compounds; a 960-compound library was synthesised to optimise the hit and these were tested by a thermal shift assay. The thermal shift assay was used in the course of the research detailed in this thesis and will be discussed in chapter five.

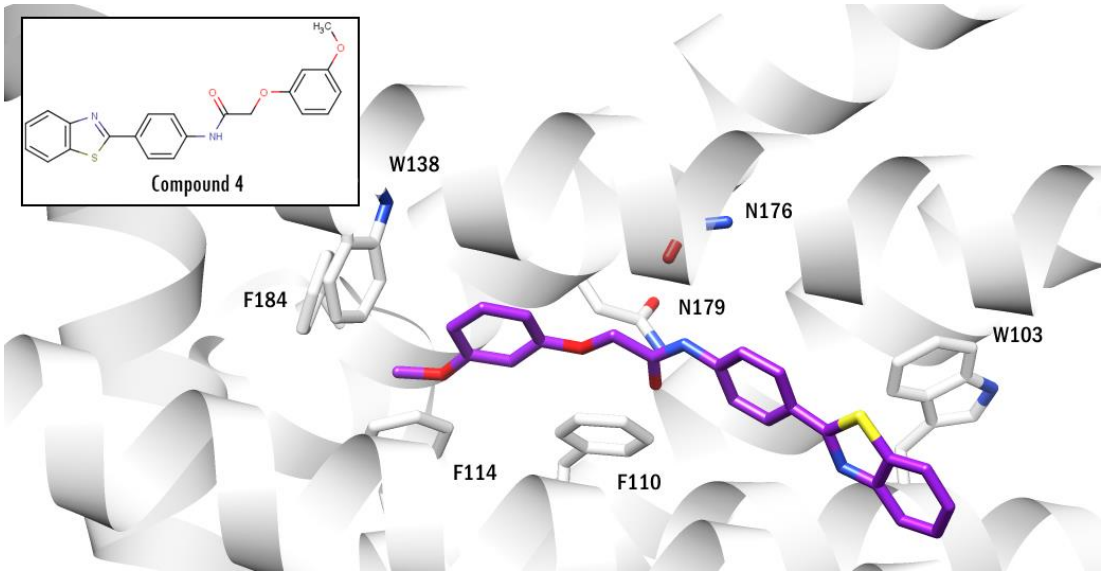


Figure 2.9: An N-phenylphenoxyacetamide class EthR inhibitor presented by Flipo *et al.*[129]

The new series (figure 2.9) exploits the Asn179 residue in the centre of the channel for hydrogen bonding while maintaining hydrophobic contacts at the top and bottom of the channel. Work is ongoing with this new family of EthR inhibitors.

In 2013 a different group with a different approach tackled EthR; fragment-based screening by Surade *et al.*[130] identified a ligand with micromolar binding affinity and a novel chemotype. A thermal shift assay similar to that used by Flipo *et al.*[129] was used to screen a fragment library of 1250 molecules, and hits were confirmed by an SPR assay based on that used by Engohang-Ndong *et al.*[121] Four of the best performing fragments (figure 2.10) were soaked into EthR crystals and resulting crystal structures (unpublished) reportedly showed clear hydrogen bonding interactions with Asn179, and favourable hydrophobic interactions. Analogous compounds were made via fragment linking; these designed molecules were docked into the fragment hit crystal structure of EthR using GOLD[39,51] and, when found to have suitable binding poses, they were tested using the SPR functional assay. This yielded a compound with a low micromolar binding affinity (figure 2.10).

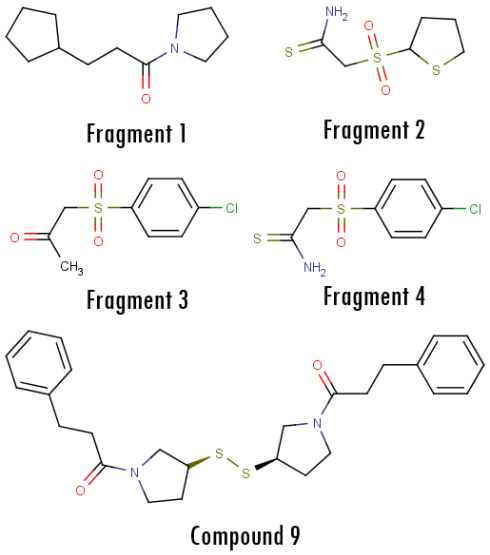


Figure 2.10: Hit fragments and the resulting micromolar inhibitor from Surade *et al.*[130]

This approach yielded a new series of EthR inhibitor leads, and Surade *et al.*[130] emphasise and credit the use of fragment-based approaches for identifying binders with low micromolar affinities which could lead to more selective lead compounds.

In 2014 Villemagne *et al.* published a study in which the fragment-based approaches of growing, linking and merging were utilised (along with crystallography and *in silico* docking) to identify a novel chemotype for EthR inhibition.[131] Application of fragment-based approaches are clearly robust for this protein and such on-going, high-impact research shows EthR to be a credible target with strong potential for real-world applications.

2.5 Conclusion

EthR is a mycobacterial transcriptional repressor which has undergone nearly a decade of target validation and inhibitor screening using ligand-based and fragment-based methods. To date, there are three distinct leads which were all derived from medium-to-high throughput biological methods.

As none of these EthR inhibitors has yet entered clinical phase trials, it is important that as many avenues of discovery are explored as possible due to the unfortunate success rates of drugs entering the process. A recent study[132] which looked at clinical phase progression, from Phase I through to FDA approval in the US (Food and Drugs Administration) between 2004 and 2010, found that only 7% of small molecule chemical drugs which entered Phase I trials were ultimately approved for the market. This compares to a 15% success rate for biologics – natural products and their derivations. These EthR inhibitors would fall into the former category rather than the latter, and therefore it is logical to cultivate multiple options in order to increase the possibility of success and eventual implementation in a TB drug cocktail for treatment.

In contrast to the previous work on EthR just described, this thesis details an *in silico* screening approach using a starter database of 6.1 million compounds to identify hits with novel scaffolds for hit optimisation.

CHAPTER III: VIRTUAL SCREENING PROTOCOL DEVELOPMENT

In silico screening has several advantages over high-throughput laboratory screening methods, including the ability to develop a selective 'assay' to identify likely binders with speed and efficiency. With an unusually rich amount of data on the EthR target, namely multiple crystal structures and established biological validation protocols, an informed protocol can be derived and honed from observations of the binding site and known protein-ligand interactions.

3.1 Protein Characterisation

Any docking or virtual screening protocol requires two components: a protein structure and a ligand structure. As multiple crystal structures of EthR have been published in the course of target discovery and validation and are available for use, two options are available: use all in an ensemble approach, or determine which would be best suited to the purpose at hand and use only one structure; the latter is preferable, as to use multiple protein structures with potentially hundreds of thousands of ligands would be extremely computationally expensive. Additionally, as these multiple structures are structurally (and statistically, RMSD across backbone and side-chains) very similar, such a large scale ensemble docking would be redundant.

Therefore multiple techniques were employed to make the most informed choice of protein model, including ensemble docking and cavity comparisons by Relibase+.[133] For the purposes of explanation and contrast of different crystal structures of the same EthR protein construct it is necessary to refer to structures individually. For clarity structures will be referred to by their four-character alphanumerical PDB code.

3.1.1 Crystal Structures of EthR

EthR, as described previously, is a bacterial DNA-binding protein of the TetR superfamily[115,121,134] which forms a dimer possessing a pair of helix-turn-helix motifs which bind DNA, and a pair of ligand-binding regions. It has been demonstrated that small molecules able to occupy these ligand-binding regions can cause a structural perturbation of the DNA-binding 'heads' thus preventing DNA-binding.[111,119,121]

Of the twelve EthR structures listed in table 3.1, all but 3Q0V were crystallised as monomers in the asymmetric unit with the space group $P4_12_12$; 3Q0V has the lower symmetry space group $P4_1$ to accommodate the dimer in the asymmetric unit.

Table 3.1: List of EthR structures in the PDB circa summer 2012 when work was undertaken. [a] denotes structures available in Relibase+ 3.1 at that time.

PDB Code	Resolution / Å	Space Group	Ligand	IC_{50} / nM
1U9N [111,a]	2.30	$P4_12_12$	hexadecyl octanoate	--
1T56 [113,a]	1.70	$P4_12_12$	diethylene dioxide	--
3G1L [119,a]	1.70	$P4_12_12$	BDM14744	Unknown
3G1M [119,a]	1.70	$P4_12_12$	BDM31381	522
3G1O [119,a]	1.85	$P4_12_12$	BDM14500	38000
3O8G [124,a]	1.90	$P4_12_12$	BDM14801	7400
3O8H [124,a]	1.90	$P4_12_12$	BDM14950	580
3SDG [128]	1.87	$P4_12_12$	BDM41425	2800
3SFI [128]	2.31	$P4_12_12$	BDM41906	400
3TP0 [123]	1.90	$P4_12_12$	BDM31343	3300
3Q0U [135]	1.70	$P4_12_12$	BDM31379	5600
3Q0V [136]	1.95	$P4_1$	BDM31369	900

Chapter two described EthR in the context of a druggable target and detailed the allosteric inhibitors developed. For developing the virtual screening (VS) protocol EthR was considered structurally. The first step was to compare the available crystal structures of EthR and catalogue their differences to determine which would be best suited as the protein model for screening.

. To achieve this objective comparison, Relibase+ was used with the first published EthR structure, 1U9N as the search model. This would compare the protein structures to 1U9N, independent of ligand-binding metrics such as K_D or IC_{50} which depend on the ligand subject.

3.1.2 Relibase+ and CavBase

Relibase+ is PDB-derived database which extends entries with cavity, secondary structure, ligand, and water characterisation.[133,137] The CavBase[138] feature in particular can be used to characterise and compare ligand binding cavities across the PDB. Notably, only protein-ligand structures are curated and there is an understandable delay between deposition in the PDB and addition to the Relibase+ release, hence why several

structures in table 3.1 were present in the PDB at the time of the work, but were not in the version of Relibase+ which was most current.

CavBase comparison searches are independent of sequence, instead matching 3D property descriptors generated using a dummy atom, which represents an area of the binding site – each of these defines a shape and chemical characteristic and is called a *pseudocentre*. The properties currently codified in CavBase are: donor, acceptor, donor/acceptor, aromatic, aliphatic, pi, and metal. Cavities are detected using a version of LIGSITE[139] and then least-squares fitting is utilised to superimpose similar binding site areas, matching properties over short pair-wise distances of less than 1 Å. A similarity score is given as a 'raw score' in the results of the CavBase search which represents the overlap of the respective surface patches – in fact the overlap of the dummy-atom pseudocentres (table 3.2).

By using CavBase on EthR, it was possible to characterise the binding sites of Relibase+-deposited structures by volume and pseudocentre, to highlight and contrast the differences. For EthR, a comparison search was performed against a total of 60,528 proteins. This set only includes EthR protein structures 3O8H, 3O8G, 3G1M, 3G1L, 3G1O and 1T56, due to the curated nature of Relibase+ (table 3.1). This does exclude five structures which are not present in Relibase+, however these were used in subsequent ensemble docking performance experiments (setion 3.1.3).

A resolution limit of 2.9 Å was set on the database search to ensure confident matches to structures of a similar or higher resolution than 1U9N, which has a resolution of 2.3 Å. The default scoring function was used but no limits were put on the minimum score to save or minimum size of permitted ligand; instead the top 100 matches were saved.

As expected, the top six results from the CavBase search were the other available EthR structures named above. Results from CavBase are tabulated with the volume of each cavity (table 3.2). A total of 97 pseudocentres within the 1U9N binding channel were identified and utilised. The 'raw score' is expressed as a percentage in the 'normalised' score, but the number of psuedocentres matched is also given, with the root mean square deviation of those pseudocentres superimposed (RMS in table 3.2).

Table 3.2: The top seven results of the CavBase comparison of the 1U9N binding channel cavity with 60,528 proteins in the Relibase+ database. A total of 97 pseudocentres were identified in the 1179 Å^3 cavity.

Cavity	Raw Score	Normalised Score	# Matched Centres	RMS	Cavity Volume / Å^3
1U9N.1	-	-	97	-	1179.38
3O8H.1	59.7	68.6	63	0.300	1018.90
3G1L.1	41.7	47.9	43	0.261	742.50
1T56.1	39.0	44.8	41	0.322	806.10
3O8G.1	34.2	39.3	37	0.485	817.60
3G1M.1	33.9	38.9	36	0.433	778.90
3G1O.1	33.8	38.9	37	0.503	872.60

The highest EthR result, 3O8H, only matched 63 pseudocentres despite the cavity similarity measured with an RMS of 0.300 Å. Several reasons for this can be suggested, such as that the Asn179 residue (an acceptor in 1U9N) is in the donor position in 3O8H and all other published EthR structures (including those not represented in the Relibase+ 3.1 release but present in the PDB) and so that region of the cavity will be, in terms of this pseudocentre mapping, very different. Also the volume of the cavity, as measured by the solvent-accessible area, is smaller within 3O8H by 160 Å^3 due to subtle side-chain positional changes (figure 3.1).

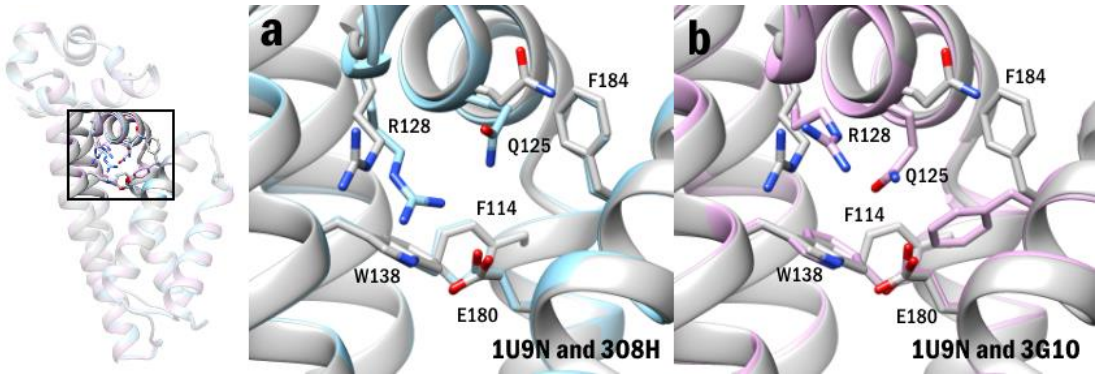


Figure 3.1: Differences in side chain positions at the top of the ligand binding channel cause differences in channel volume. Panel **a** compares the two largest cavities, 1U9N (grey) and 3O8H (cyan) which differ by 160 Å^3. Panel **b** compares 1U9N with 3G1O (pink), which differ by over 300 Å^3.

The cavity volume difference between 1U9N and 3O8H is driven by the positions of Arg128 and Gln125, which differ greatly between the two structures (figure 3.1). The difference in cavity volume between 1U9N and 3G1O exceeds 300 Å^3 and this is due to the movement of Phe184, which can be seen on the right of panel **b** in figure 3.1, above.

In 3G1O, with this phenylalanine residue down, this upper part of the binding channel is inaccessible to any ligand.

Ultimately, the Relibase+ analysis demonstrates that despite close similarity between structures by RMSD, the rearrangement of side-chains in the presence of longer ligands results in cavities differing in volume by up to 300 Å^3. For virtual screening purposes it was decided one representative EthR structure should be used to reduce the computational burden and avoid redundancy as much as feasibly possible. With notable differences in volume, ensemble docking was used to determine what effect this had on ligands binding EthR *in silico*.

3.1.3 Ensemble Docking with EthR Structures

Ensemble docking is a method by which ligands are docked into multiple protein models to account for protein flexibility in the backbone and structural variation lost in typical crystal structures.[45,46] Each ligand is docked into all presented protein structures, with the resulting poses being scored and saved depending on the user-defined termination parameters. This method has successfully been used for hit discovery with homology models,[140] molecular dynamics simulations,[66] and even NMR structures[65] in addition to crystal structures[59] within the last decade.

With multiple ligand-bound crystal structures of EthR available, a cross-docking experiment was devised to determine which of the protein structures was the most "receptive". Cross-docking is a form of ensemble docking by which the ligands of a set of ligand-bound crystal structures are extracted and then docked back into the ensemble. Therefore the most receptive structure would accept the most ligands, with a good differentiation of scores, and so be the preferred screening model.

For ensemble docking, all protein structures must be aligned so they can share the same search space; GOLD is able to perform a simple transformation to overlay the eleven EthR structures, in this case with reference to 3TP0. Figure 3.2 shows the alignment of these eleven structures in PyMol to an average RMSD of 0.56 Å over all Cα atoms. A protocol was adapted from previous work with small molecule crystal structures for the ensemble docking.[122] Therefore, all structures were aligned to the co-ordinates of the EthR structure 3TP0.

Eleven of the twelve structures listed in table 3.2 were utilised for cross-docking to establish which protein structure would perform best under virtual screening conditions. The structure 1T56 from the Dover group was omitted to form a cohort of structures from one research group, ensuring consistency among the crystal structure determination methods. In preparation for docking, all water molecules were removed

from each structure and, as GOLD uses an all-atom model, hydrogen atoms added to satisfy unfilled valences. For each of the eleven protein structures, the ligands were removed and saved as separate structure files. Ten of the extracted ligands were docked; hexadecyl octanoate, the ester ligand of 1U9N, was omitted as it was not a drug-like molecule.

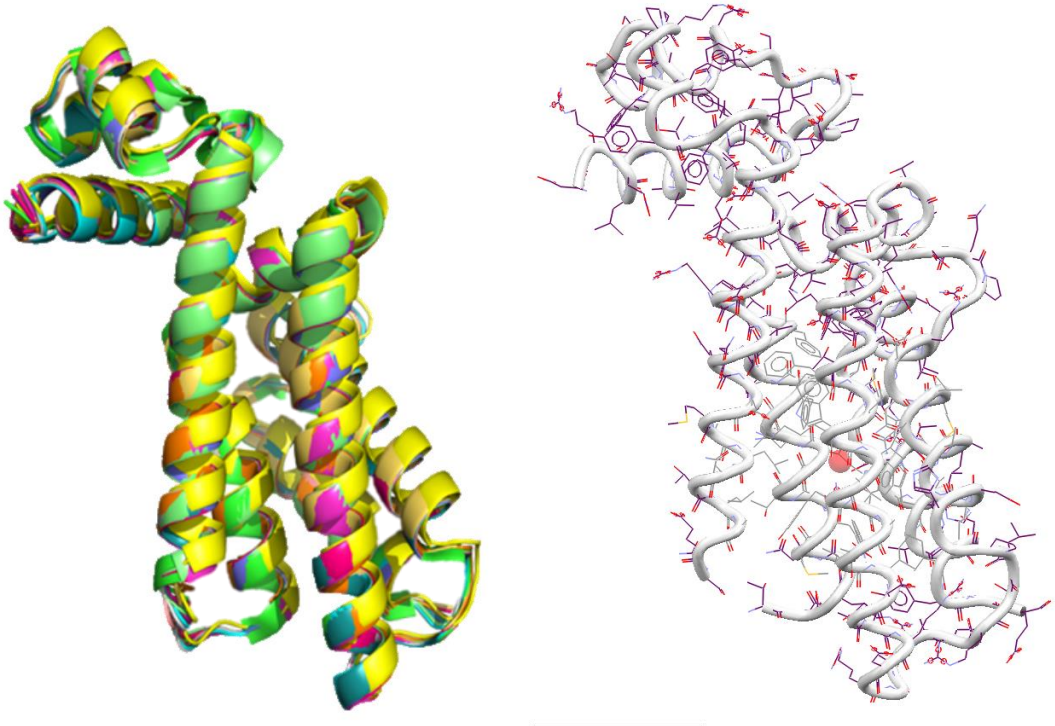


Figure 3.2: Left, alignment in PyMol of the eleven EthR crystal structures to be used for ensemble docking. Right, structure 3TP0 with red ball denoting centre of search space, with residues in grey included in 10 Å radius; purple residues not included in solvent accessible binding site (made in Hermes).

The search space was defined as a 10 Å radius around a central point in the binding site, spanning from the upper residues Phe184 and Trp138, to the lower part of the channel including Trp103 and Trp207 (figure 3.2).[122] GOLD's default settings for ensemble docking were used with a user-defined goal of 25 GA runs per ligand and the default scoring function (CHEMPLP). Maximum ligand flexibility settings were used. All 25 resulting poses were saved, with termination specified when the diverse solutions criteria (cluster size of 2 with an RMSD of 1 Å) were satisfied.

It has been previously noted that the residue Asn179 differs in structure 1U9N to the rest of the EthR structures in that the carbonyl of this residue is presented to the binding channel surface rather than the nitrile. Close inspection of the 1U9N structure showed that if the residue were to be flipped in keeping with the other structures, a supporting hydrogen bond network is formed (figure 3.3). It was supposed that in the

1U9N structure, being of a lower resolution than the others, the Asn179 residue built into the electron density in an incorrect orientation.

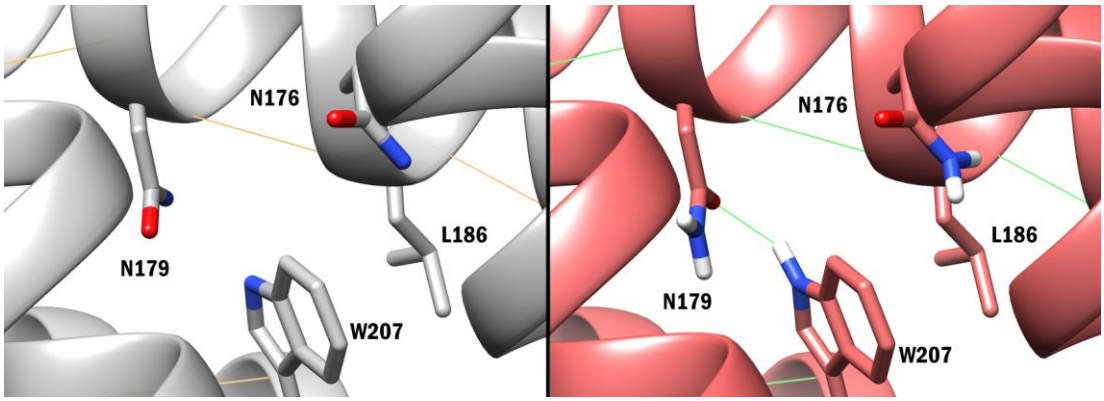


Figure 3.3: Left, the orientation of N179 in the original PDB 1U9N (which lacks hydrogen atoms). Right, the flipped orientation and resulting hydrogen bond formed with W207 upon setup for docking.

Therefore, the decision was taken to run the ensemble cross-docking twice: once with the Asn179 residue of 1U9N in place as published, and again with the residue flipped 180° in GOLD to present the nitrile in accordance with the other structures. The histogram of results for the original and flipped 1U9N are shown in figure 3.4.

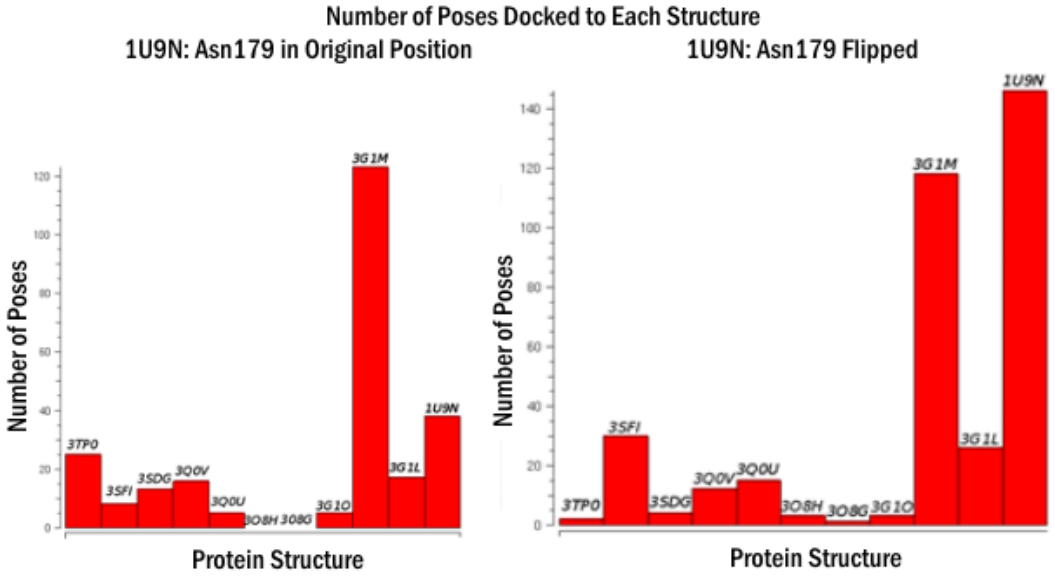


Figure 3.4: Histograms showing the number of ligand poses docked for each structure. Small variations are to be expected due to the stochastic nature of docking in GOLD.

Comparison between the original and flipped cross-docking experiments shows a similar pattern in the quantity and distribution of the docking results, with the notable exception of the edited 1U9N. Small variations are to be expected – solutions for 3TP0 and 3SFI vary by 20 poses – however 1U9N changes by 108 solutions, preferentially binding more ligands than all other structures when Asn179 is able to act as a hydrogen bond donor.

This is likely due to a bias toward hydrogen bond acceptors in the ligand structures, and the large volume of the 1U9N binding channel compared to other structures.

The promiscuity of both 3G1M and 1U9N makes each a strong choice as the potential binding protein structure for virtual screening, and indeed if they were considerably different both could be utilised and the ensemble docking adapted for a virtual screening protocol. However, given the strong structural similarity between the two (with an RMSD over all non-hydrogen atoms of 0.27 Å), it would be an unnecessary addition to an already computationally straining process.

Therefore it was decided to use the 1U9N structure of EthR for virtual screening - and in so doing maximising the area for potential ligand binding - and use the properties of known actives with the published co-crystal structures as guides when filtering the ligand structures before the screening itself.

3.2 Acquisition, Filtering and Clustering of a Ligand Database

3.2.1 The ZINC Database "DrugsNow" Set

Compound structures for virtual screening were sourced from ZINC, a free online database (zinc.docking.org).[141] The ZINC database contains over 35 million commercially-available compounds which can be downloaded in various 3D formats, ready to be used in docking. The database is organised into "ready-to-download" subsets filtered by physical properties and availability. The *Drugs Now* subset was downloaded to be used for virtual screening due to the pre-application of Lipinski's Rules and the apparent availability of the compounds therein. When downloaded, the *Drugs Now* subset consisted of approximately 6.06 million compounds.

3.2.2 Filtering in KNIME

Pipeline software KNIME (discussed in chapter one) was used to apply additional filters to the *Drugs Now* subset downloaded from the ZINC database, and to calculate some simple chemical property descriptors. Figure 3.5 shows the filtering pipeline which reduced the subset to 1.3 million compounds. Twenty descriptors were calculated, listed in table 3.3, chosen to describe the physiochemical properties of each compound in simple terms with integer values. Filters were applied in stages in order to reduce the heavy computational burden of calculating all twenty descriptors on all 6.06 million compounds. All twenty descriptors were also calculated for the ten known active EthR-binding ligands described previously.

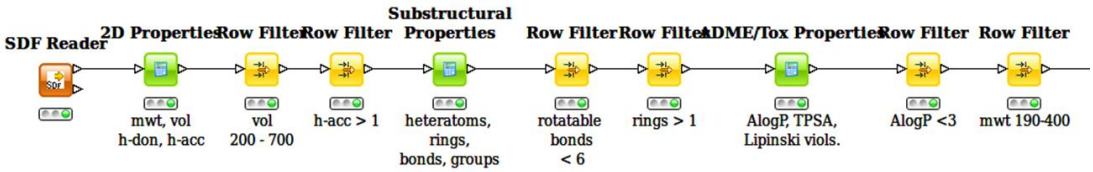


Figure 3.5: KNIME workflow for the calculation of descriptors and filtering of the *Drugs Now* subset. The orange **SDF Reader** node brings the input from the ZINC database; green nodes are calculators from the KNIME Chemistry Nodes package and the yellow filter nodes are part of KNIME's core node set.

Firstly, a filter was applied to exclude any molecule without a volume between 200 Å^3 and 700 Å^3 (the first yellow box in figure 3.5). The value of 700 Å^3 was chosen to include molecules which could occupy the main cavity of EthR (table 3.2, ranging from ~740-840 Å^3). Though the cavity of 1U9N is ~1100 Å^3, the decision was taken to exclude long, large molecules which would occupy the whole volume in order to identify small, varied scaffolds which can be further modified in downstream medicinal chemistry efforts to exploit the additional volume. This volume filter is the most stringent, reducing the cohort by 50%.

Table 3.3: Descriptors calculated in KNIME using the workflow in figure 3.4. Not all properties were used for pre- or post-screening filtering.

molecular weight	**volume**	**# hydrogen bond donors**	**# hydrogen bond acceptors**
# heteroatoms	# rings (any type)	# rotatable bonds	# halide atoms
# f atoms	# ester groups	# hydroxyl groups	# keto groups
# methyl groups	# sulphide groups	# sulphonyl groups	# thiol groups
# thioester groups	# Lipinski violations	TPSA	AlogP

The polar region of the pocket includes the residues Asn179 and Asn176, which in the crystal structures of EthR have been shown to act as hydrogen bond donors through their NH_2 side-chain group. A filter was imposed to include only those molecules with at least one hydrogen bond acceptor in order to encourage intermolecular interactions between the ligand and the protein.

The upper limit for rotatable bonds was set at six or fewer, to disfavour widely flexible molecules,[142] and a filter was imposed to ensure all molecules had at least one ring system of any kind. Together with the volume filter, the decreased flexibility and the inclusion of ring systems will exclude undesirable long, flexible molecules and retain shorter, wider, drug-like compounds. The penultimate filter excludes any molecule with an AlogP above 3, as any molecule with a value above the imposed threshold here would be extremely hydrophobic.

The final filter simply restricts the molecular weight of the included compounds to a range of 190 to 400. While molecular weight and molecular volume are correlated, this filter was chosen to exclude any large compounds which had managed to escape previous filters, and was based on the molecular weights of known actives.

The workflow used was the result of over a month of extensive literature study and parameterisation to obtain reasonable, effective filters. Due to the computational load of calculating and filtering on properties for over six million compounds, slices of approximately 131,000 compounds were introduced to the workflow separately. This resulted in 46 sets of ~131,000, each requiring an hour of clock-time to complete the workflow to give ~28,000 compounds per slice, or approximately 1.3 million compounds total.

3.2.3 Clustering the Filtered Ligand Set

While feasible to screen all 1.3 million compounds against EthR, it was not desirable; analogues and highly similar structures would populate the filtered set and so it would be more efficient to take a representative proportion to screen. A clustering pipeline was developed to group molecules by different ring chemistries (figure 3.6).

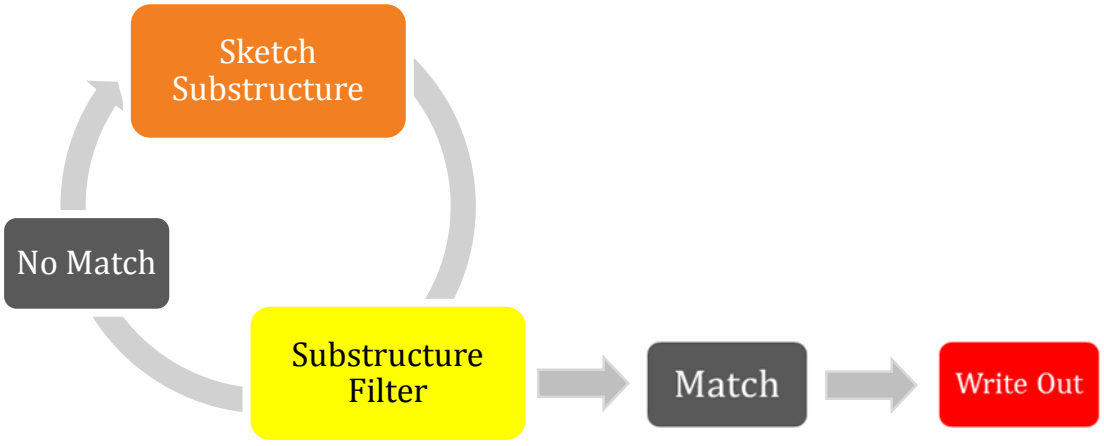


Figure 3.6: Overview of the clustering pipeline for the 1.3 million compounds filtered from the ZINC database *Drugs Now* subset. Marvin Sketch was used for defining ring structures.

All molecules were handled in SDF format at either end, which kept all information from previous KNIME functions as appended tags, and multiple structures were kept in single text files. The first decision made was to remove any molecules which the RDKit rejected

for structural anomalies (a total of 7 from 1.3 million) and a total of 5,265 rhodanines, a member of a class of molecules known as *pan-assay interference compounds* (PAINs) which provide false positives in a variety of biophysical assays to little known biological effect.[143]

The ring definitions were ordered with some respect to frequency in drug molecules. Where some rings had multiple possible arrangements (ie. oxadiazoles 1,2,4 or 1,3,4; and triazoles 1,2,3 and 1,2,4) the filtered results were concatenated into one output. Some sulphonyls caused confusion to the cluster filter on heterocycles as the pre-filtering did not distinguish between in-chain sulphonyls and sulphonyl-like arrangements in rings. Therefore, a sulphonyl filter was imposed before the final three clusters were formed. The last clusters are 'other heterocycles' either of five or six membered rings which had not been previously identified (6,091); any kind of seven-membered cycle, as defined by Murcko Scaffold[144] (10,768); and the last cluster was named 'other', to encompass structures with prolyl or eight-membered rings, and structures with only adamantanes or other similar cage groups (52,297). Table 3.4 shows the full list of clusters with their populations.

With the 1.3 million pre-filtered set clustered, subsets had to be taken from each for docking. Different sampling percentages were explored for creating the final docking set, with a target of approximately 500,000 compounds for screening. The decision was taken to make the groups as even as possible, taking different proportions from each. For small clusters of fewer than 10,000 compounds, such as thiadiazoles (8805), the cluster was taken in its entirety. For large clusters of greater than 100,000 compounds, fractions were taken so no single cluster was populated greater than 50,000 compounds (or ~10% of the final docking set of ~500,000). This results in a rather uneven sampling, with clusters of 40,000 down to clusters of only 6000 compounds, however this is more even than the initial screening, and as docking of each compound is independent, the composition of the screening set would not bias the screening itself in any way, unless the protein interactions favoured a particular chemistry (which would be an interesting result in and of itself).

Ultimately, the docking set consisted of 409,201 compounds (table 3.4), with the largest cluster, that of pyrazoles, containing 40,035 compounds (9.8% of the cohort) and the smallest, that of 6-membered ring triazoles, containing 4,150 compounds (1.0% of the cohort).

Table 3.4: Proportioning the clusters for docking, in order to take roughly a third of the filtered set in total, as equally represented as possible. The '5', '6' and '7' prefixes refers to the ring size.

Cluster Name	Cluster Size	Screening Cluster Size	Cluster Name	Cluster Size	Screening Cluster Size
5imidaz	121369	**36411**	**6tria**	4150	**4150**
5pyraz	133450	**40035**	**5thiaz**	36150	**10845**
5pyrrol	240174	**24017**	**5tetraz**	9372	**9372**
5triaz	49055	**14717**	**6pip**	182898	**18290**
5oxaz	39025	**11708**	**6oxane**	29035	**14518**
5oxadiaz	11439	**11439**	**6diox**	11814	**11814**
5thidiaz	8805	**8805**	**6morph**	38142	**11443**
6pyrim	107658	**32297**	**5thio**	22934	**11467**
6pyraz	91590	**27477**	**5oxodiox**	28635	**28635**
6pyrida	16227	**16227**	**Sulphonyls**	32986	**32986**
Hetother	6091	**6091**	**Other**	52297	**15689**
7cycle	10768	**10768**	**Total**	1284064	**409201**

3.3 Preliminary Docking Tests: Training Sets and Parameter Optimisation

3.3.1 "Decoy" Set for Probing GOLD Parameters

Small training sets seeded with compounds known to be active inhibitors of EthR served two key purposes; firstly, a decoy set was formed by which to test the developed docking protocol for the ability to prioritise active compounds; and secondly, the set gave a dataset of manageable size for fast docking in order to probe the functions and limitations of GOLD in the protocol development process.

For decoy sets, it would be ideal to use known inactives to form a set seeded with known actives, to test a protocol for its ability to prioritise those with biological activity. Without such information available for EthR, a set was derived from a similarity search. Using the lead compound from the work published by Flipo *et al.* in 2012,[128] the structure of BDM41906 was used to design a query of the ZINC database based on simple physiochemical properties, chosen to be similar to the actives to avoid artificial enrichment and to provide a challenge to the docking protocol. Molecules were selected from the ZINC database at random given a set of physiochemical properties as filters. The properties in question were based on guidance from Lipinski's rules,[120] with seven or fewer rotatable bonds; five or fewer hydrogen bond donors; up to ten hydrogen bond acceptors; a molecular weight between 300-500 g mol^{-1}; and an *x*logP lower than 3. A total of 350 of these were chosen to satisfy a recommended 36:1 ratio of unknowns to actives (of which there were nine used), as suggested by Wallach *et al*. in 2011.[145]

Given the virtual screening set was derived from the same source with similar filters imposed through KNIME, cross-over was to be expected (though, of the decoy set of 350 compounds, only 4 are present in the virtual screening set and none in the biophysical set for screening); however the key aim of the decoy set would be fulfilled if the leads were prioritised in the top 10% of the results. The docking protocol described below was evolved from the ensemble docking described previously.

3.3.2 Optimising GOLD Parameters for EthR

There were five key protocols tested and carefully analysed for active enrichment with the decoy set, focussing on both 1U9N (which has a more extended binding site owing to the Phe184 twisting up to expose the auxiliary pocket) and 3G1M (which performed well in the ensemble docking but has a binding site of smaller volume). In each case, the 350 molecule decoy set of compounds were used in addition to nine active EthR inhibitors from the published crystal structures (BDM14500,[119] BDM14801,[124] BDM14950,[124] BDM31343,[123] BDM31369 from 3Q0U, BDM31379 from 3Q0V, BDM31381,[119] BDM41425[128] and BDM41906[128]). Similarly, a search space radius of 10 Å about a central point in the binding site was defined, the co-ordinates for which differ based on the slight structural variation between proteins used, though an effort was made to make this centre positioned in the same structural environment to that used for the ensemble docking described previously. For 1U9N, the centre point was defined as [x = 32.5127, y = 73.6746, z = 11.1142]; for 3G1M, the centre point was defined as [x = 31.987, y = 74.0719, z = 10.2291]. Autoscale of 100% (1.0) was employed, with automatic genetic algorithm settings utilised for 10 GA runs per ligand, of which the best five were kept. Finally, CHEMPLP was used as the scoring function, and the Diverse Solutions definition for end-point were defined as clusters of two molecules at least 0.5 Å in difference by RMSD.

Two of the five decoy docking protocol experiments were with 1U9N and 3G1M respectively, unedited (but for the Asn179 residue of the former) and no additional protein flexibility. The third changed the parameters for 3G1M to include five library-defined rotamers of the phenylalanine residue F184 to ascertain if opening up the option of the longer binding site would affect enrichment. The fourth and fifth experiments were cross-docking ensemble experiments combining 3G1M and 1U9N (aligned to the latter) under non-flexible (four) and flexible (five) phenylalanine conditions.

For each docking experiment, all proved able to prioritise actives in the top 10% of results by fitness score, and therefore early enrichment was not helpful in differentiating protocols. Inclusion of the flexible phenylalanine F184 in 3G1M gave no

noticeable advantage to using that protein structure over 1U9N, and the ensemble results again showed 1U9N able to dock more ligands (data not shown); therefore, 1U9N was retained as the protein model for virtual screening. Focus instead shifted to prioritising a protocol which was fast, would potentially be feasible to run on a desktop PC with a limited core capacity, and would still be able to prioritise active ligands.

To increase the speed of the docking, autoscale search efficiency was dropped to the GOLD default for virtual screening of 30% and then to 20%, and the 1U9N solo protocol was rerun to give good early enrichment from training to test set in each case. The diverse solutions setting was altered to define cluster sizes of one pose only with a 1 Å RMSD, to increase the sample of conformational flexibility in the results; this proved successful. Various search space radii were tested but smaller radii did not best utilise multiple residues lower in the pocket. Similarly a radius of 12 Å was able to include the upper auxiliary pocket but the Asn179 interaction was deemed most important, and as GOLD is known to artificially enrich hydrophobic ligands due to the scoring function (see chapter 1), the decision was taken not to encourage the hydrophobicity to the potential detriment of crucial hydrogen bonding lower in the channel. Finally, to determine the disk space and real-time running time on the desktop PC being used, ten thousand of the intended virtual screening ligands were run using this protocol to extrapolate to an estimate for the virtual screening output.

The final protocol can be found as the CONF file used in the appendix. To summarise, a 20% autoscale was used, with a search radius of 10 Å. GOLD's automatic function was used for the genetic algorithm, with five GA runs conducted per ligand and all poses saved. Diverse clusters of one pose per cluster with an RMSD of 1 Å between clusters were defined.

3.3.3 Running the Virtual Screening Protocol

Using previous runs and the 10,000 ligand test run, it was possible to estimate how long the protocol would take to complete on a PC, and how much file-space for output would be required. Though file-space using the SDF output format was not a problem given the option of external hard-disk drives, it became clear the running time would be problematic; unfortunately, for the protocol and the 409,201 ligands, it was determined a run-time run of over 52 days would be likely on the desktop PC running the Ubuntu 12.04 LTS (64-bit) operating system, using four Intel® Xeon® CPU W3530 processors at 2.80 GHz and 12.5 GB memory. It was undesirable to reduce the number of ligands for screening any further, and so the decision was taken to turn to a supercomputer.

As such, the CONF file, edited protein structure (1U9N with hydrogens, no water molecules or ligands, and a flipped Asn179 residue in SYBYL MOL2 format) and ligands for docking (409,201 in 23 SDF files) were sent via FTP to the CCDC server. The virtual screening protocol was set up there on my behalf and run (taking approximately 3-4 hours), utilising the University of Cambridge High Performance Computing facilities, specifically the Darwin Supercomputer.

With five poses for each of the 409,201 ligands screened, over two million poses resulted.

CHAPTER IV: POST-SCREENING FILTERING PROTOCOL

The virtual screening consisted of five poses for each of the 409,201 ligands screened. A total of over 2 million ligand poses had therefore been generated and required analysis and filtering to derive a cohort of compounds suitable for experimental testing against EthR. To achieve this the GoldMine software tool, as part of the GOLD Suite (CCDC) which had been used for screening, was utilised for the post-screening data processing.

GoldMine allows for analysis of docking results with a variety of customisable descriptors to characterise binding interactions. This can include simple physical descriptors (such as molecular weight and the number of rotatable bonds) but extends to characterising the protein-ligand interactions of docking poses, such as hydrogen bonds or whether particular regions are occupied by a ligand. Via descriptors it is also possible in GoldMine to characterise unexploited interaction potential in occluded polar and hydrogen bond-capable atoms.

This chapter details the rational process by which unfavourable ligands were excluded from the pool of potential binders in order to derive a small set of promising compounds for biophysical evaluation.

4.1 Retrieval of Data

In preparing for the virtual screening, various test sets were used to optimise and fine-tune protocols; one such set used was a 10,000 molecule set referred to as '*minidock*' (see chapter three) which was subject to the full virtual screening protocol. This set was used in the familiarisation with the GoldMine interface. This initial work was undertaken in the CCDC.

The virtual screening output consisted of five sets of co-ordinates (poses) for each of the ligands. Docking failed to identify poses for six compounds from the 409,201 screened resulting in 409,195 successfully docked ligands.

4.2 Filtering

As five poses were generated for each ligand, a total of 2,045,975 poses resulted. To create a GoldMine database of all poses would be an exceptional computational burden – to put this in context, it took the best part of eight hours (wall-clock time) to set up the GoldMine for ~40,000 compounds so it would have taken around 16 days to set up over

2.05 million, assuming the software would even allow it and remain usable and functional. Therefore the decision was taken to take only the top-scoring pose of each ligand -making the assumption that the best scoring pose of that ligand would make the best interactions and have the best conformation - giving the highest scoring 10% of ligands, a total of 40,919.

4.2.1 Filter by Fitness

To filter by fitness and retrieve the top 10% of top-scoring poses, a Python script was used. This script, which accessed each directory sequentially, read the associated text file which lists the poses in order of rank; if the highest scoring pose listed there scored higher than any in the current buffered list of top 40,919 ligands, it replaced the lowest scoring pose, otherwise the script moved to the next ligand directory of poses. This Python script can be seen in appendix A.

To determine the proportions of each cluster retained in these results, the top ten percent were re-clustered through the pipeline detailed in section 3.2.3 and shown in appendix A; these data are shown as a graph in figure 4.1.

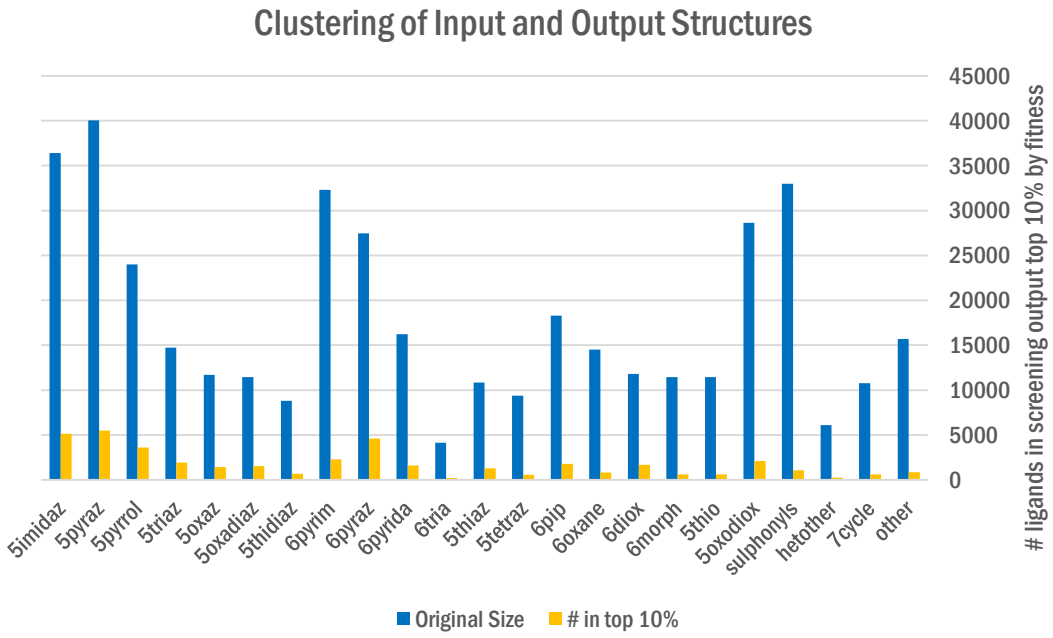


Figure 4.1: Graph of re-clustered virtual screening output, according to KNIME pipeline in section 3.2.3. Blue bars shows the distribution in the virtual screening input set of 409,200 compounds; orange bars show the number of molecules from those clusters present in the top 10% of virtual screening output. Cluster names given on the *y*-axis refer to the ring chemistry used, detailed in chapter three.

The re-clustering shows that the virtual screening output contains a range of chemical diversity and the output is roughly proportional to the distribution of input structures.

Therefore the virtual screen is not biased toward a certain type of chemistry. This is particularly encouraging as the docking process is stochastic and so each ligand and pose is unique and unbiased. As this virtual screening approach was taken to identify novel scaffolds for EthR inhibitors, the top ten percent of the output is clearly very diverse.

4.2.2 Hydrogen Bond Geometry Filter

All descriptors calculated in KNIME for the filtering of the ZINC database subset were carried forward through screening and into the GoldMine, allowing filters to be used beyond simply the fitness score and its components. GoldMine and Hermes also have functionality to calculate novel descriptors based on docking poses.

With this in mind, the first descriptor calculated within GoldMine was used to define a minimum hydrogen bond contact angle. The hydrogen bond tolerance in GOLD, as previously discussed in section 1.3.1.3, has a minimum angle of 100 degrees. The descriptor was intended to filter out ligands which do not form hydrogen bonds to the protein, and ligands which do not form hydrogen bonds at optimal geometries; this was implemented first as it was clear from previously developed inhibitors for EthR that hydrogen bonding in the polar region of the EthR ligand binding site is especially important.

The definition of a hydrogen bond can be edited with respect to actual distance or Van der Waals distance, and the required hydrogen bond angle can also be user-defined. To understand the effect of these descriptors on the filtering of molecules prior to their application to the virtual screening results, the *minidock* set was again used. An initial filter on the fitness, retaining the top 10% of the scoring ligands, resulted in 1846 poses for 907 ligands. Initially, the default hydrogen bond definition parameters were used, with a minimum angle of 90° and a maximum distance being the sum of the Van der Waals radii of the donor and acceptor (which was kept constant). This default-parameters descriptor reduced the set to 1182 poses for 693 ligands, a reduction which demonstrates 24% of the ligands were not forming a registerable hydrogen bond to the protein.

By increasing the minimum angle through 100°, 110° and 120°, the number of ligands was reduced incrementally to 669, 653 and 629 ligands respectively. The final parameter, with a minimum angle of 120° and a maximum distance being the sum of the Van der Waals radii, retained 69% of the ligand population. Whether this parameter would be used as a final filter was not decided, however it could be considered an additional property for consideration.

4.2.3 Calculation of Additional Descriptors

To prioritise likely binders, filters upon different characteristics were added or edited in an additional, rational process until a set of non-stringent but decisive filters were combined. Ideally, it was hoped these filters would exclude those ligands which were less likely to bind EthR or were unfavourable due to physiochemical characteristics, and prioritise those which were more promising. To cultivate multiple options for filter parameters, GoldMine's descriptor calculators were used.

Additional descriptors can be calculated on ligands in GoldMine, including the number of buried donatable hydrogens, buried acceptors, and buried polar atoms, as well as the number of ligand atoms forming hydrogen bonds and the number of atoms causing protein-ligand clashes. All of these descriptors were calculated for the full set of 40,919 ligands in order to maximise the amount of information available on the docked structures.

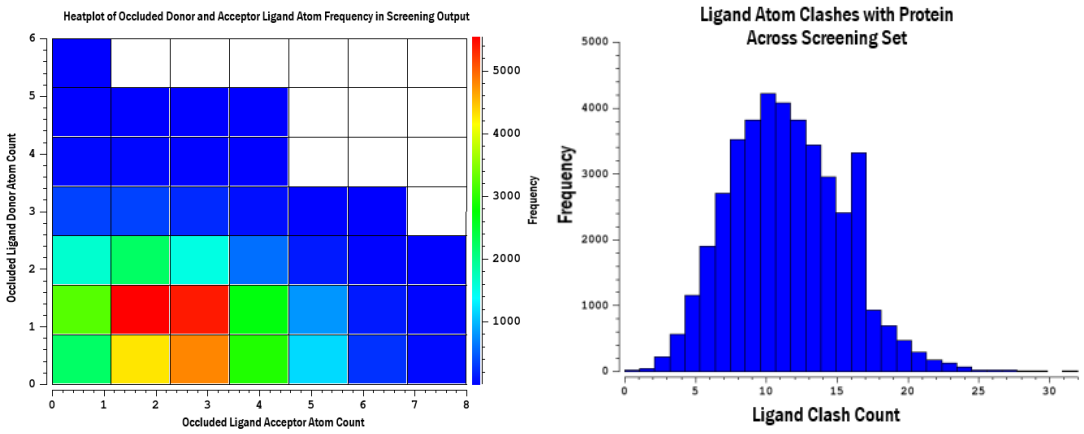


Figure 4.2: Left, heatplot output from GoldMine showing frequency of the occluded ligand donor and acceptor atom counts, with colour-scale shown on right of red denoting denser populated bins. Right, Histogram showing distribution of calculated ligand clash counts, distinct from the score contribution shown in figure 4.4.

Of the 40,919 ligands, 23,261 possess at least one occluded (obstructed or unfulfilled) polar atom; 25,164 possess at least one occluded donatable hydrogen and 39,850 ligands have at least one occluded hydrogen bond acceptor atom (figure 4.2, left); and distinct to the previous angle-dependent hydrogen bond filter, a total of 25,941 make at least one hydrogen bond (as defined at 90°). Finally, a ligand clash count was calculated, which ranged from zero to 32 (figure 4.2, right).

In addition to calculated descriptors, both those carried from the pre-screening pipeline and those defined and calculated via GoldMine, components of the scoring

function were looked at as useful indicators of favourable *in silico* binders. Though components of the scoring fitness function contribute to an overall discernment of predicted binders, some terms (such as for CH---O interactions) when taken in isolation did not help in identifying likely poor binders for exclusion. However, three terms from the CHEMPLP fitness function were found to be of particular relevance for assisting in the exclusion of unfavourable poses: the ligand torsion term, the ligand clash term – distinct from the clash count calculated - (figure 4.3), and the internal correction term (figure 4.4).

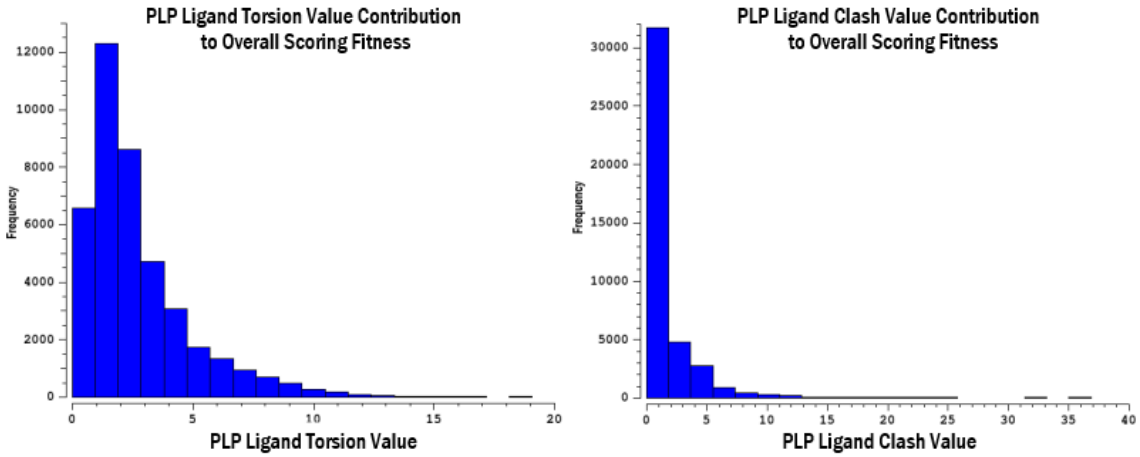


Figure 4.3: Histograms of scoring fitness contributions for the ligand torsion (left), and ligand clashes (right) Equations for these terms can be found in chapter one.

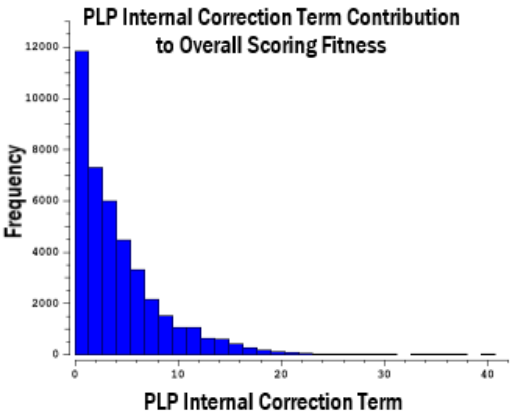


Figure 4.4: Histogram of the PLP ligand correction term from the CHEMPLP scoring function. Equation for this term can be found in chapter one.

4.2.4 Iterative Filter Design

Firstly, the hydrogen bond contact descriptor was applied (detailed above), compounded with a filter on the hydrogen bonding scoring term from Chemscore with a minimum

value of 0.1. In principle this would reduce the screening result set to only those which formed a 'good' hydrogen bond; this filter returned a total of 17,249 ligands (42.15%).

In chapter one, physiochemical rules-of-thumb for drug molecules were discussed, including a maximum polar surface area of 120 Å^3 as suggested by Veber.[142] This was not implemented in the pre-screening pipeline, and so was added here. Figure 4.6 shows the distribution of the total polar surface area (TPSA) of molecules in the screening output set.

From figure 4.5 (overleaf, left), it is apparent that the majority of molecules fall below the 120 Å^3 threshold, and indeed implementation of this parameter shows a total of 39,704 ligands remain. When added to the hydrogen bond parameters (with a score increased to 0.5), the overall set is reduced to 13,786 ligands (33.6%). This filter iteration was considered to be *filter1*.

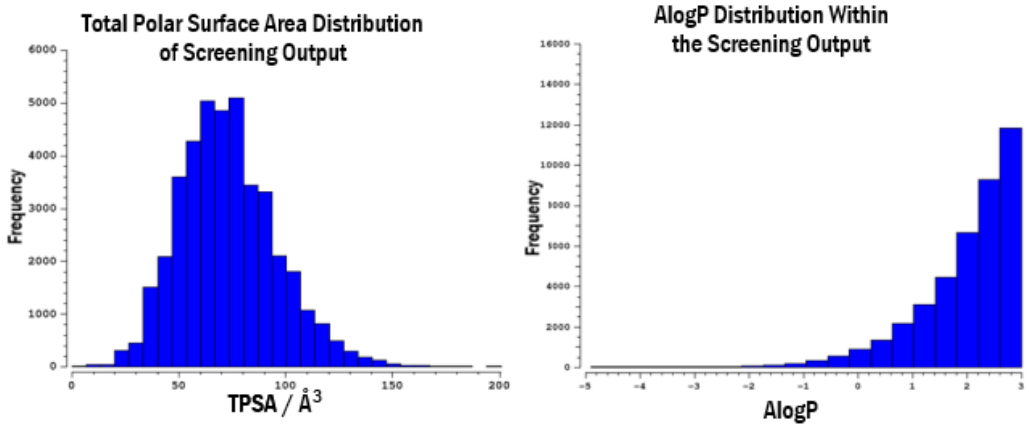


Figure 4.5: Left, histogram output from GoldMine, showing the distribution of the total polar surface area of molecules within the screening output. A minority of molecules possess extremes of this property. Right, the distribution of calculated AlogP of molecules within the output. In pre-processing, a filter restricted the set to AlogP ≤ 3.

A parameter which had been utilised in the pre-screening process was AlogP, and values calculated for the more potent active compounds using the same node gave a range from -3 to +1. Figure 4.6 shows a histogram for this property in the screening set. With the active-derived range of -3 to +1 implemented, a total of only 5,453 ligands remain – 13% of the screening output. When implemented alongside the other parameters, the initial set of 40,919 is reduced to only 2,363 ligands (5.77%).

Although quite stringent, this filter is based on the properties of ligands known to bind EthR, and allows for compounds that must be quite lipophilic to interact with the binding site as well as pass through the mycobacterial cell membrane. The combination of *filter1* with the addition of the AlogP filter was named *filter2*.

Several parameters were considered as additions to *filter2*, specifically from the atom descriptors calculated in GoldMine (figure 4.2). These were implemented in *filter3*: the extremes of occluded ligand donor and acceptor atoms were excluded with a filter of 0-2 atoms (39,772 ligands) and 0-4 atoms (38,266 ligands) respectively. On figure 4.2, this excludes the low-populated blue areas on the heat plot. The occluded polar atom count was similarly restricted to 0-2 atoms, including 39,466 ligands, and the ligand clash count was restricted to a maximum of 16 (figure 4.3, right). *Filter3* reduced, through small exclusions, the screening output from 40,919 ligands to 1,761 (4.30%).

There were three iterations of *filter4*, all of which only slightly altered already-implemented parameters in the filter. This included putting a lower-band limit on TPSA of first 70Å^3 (*filter4*), then revised to 75Å^3 (*filter 4a*) to coincide with Veber's recommendations. The ligand clash count was reduced by one point to 15 (*filter4*), and for *filter4b*, the number of occluded ligand acceptor atoms was reduced to a range of 0-3. Initially *filter4* included 1,303 ligands, but small changes through *filter4a* (1,097) and *filter4b* reduced this to only 889 ligands (2.17%). Though a small percentage of the total virtual screening output, this remained too high a collection of ligands for comfortable, valuable visual inspection, and *filter4c* reduced the clash count further to 14, and the TPSA range to 75-115Å^3, reducing the set only to 777 ligands.

Filters 5 and 6 introduced components of the scoring function as parameters to the filter, namely the PLP ligand torsion term, the PLP ligand clash term, and the Chemscore internal correction term (these are discussed fully in section 1.3.1.3, histograms as figure 4.4). Specifically, *filter5* set upper-bound limits of 4 and 2 respectively for the ligand torsion (including 32,844 ligands) and ligand clash (31,950 ligands) terms. Overall, this gave a set of 506 ligands. The addition of an upper-bound limit of 4 for the internal correction term (24,776 ligands) in *filter6* reduced this further to 447 ligands.

These filters in combination were sufficient to exclude extreme examples of any one descriptor, but in combination provided a small set for visual examination and final selection for laboratory testing. Therefore, *filter7* was a careful change to the ranges acceptable in ligand selection, to provide non-stringent over-biasing descriptor filters and a small ligand cohort. Table 4.1 summarises *filter7*.

It was possible to remove the filter on occluded acceptor atoms and, with small changes to the number of occluded polar and donor atom filters as well as the scoring function terms used, bring the total number of ligands included after filtering to only 284 (0.69%). As can be seen in table 4.1, only the TPSA and AlogP filters exclude more than half the ligands, and they were decisive in returning a set of manageable size for visual

inspection. Also included in table 4.1 is the overall CHEMPLP fitness, though the selection of the top 10% of ligands occurred during file retrieval not in GoldMine; nevertheless, it is a filter upon which ligands were selected and it must be considered as such explicitly. Figure 4.6 shows the result of the filtering process on the number of ligands included for consideration, which follows an exponential decline through each iteration.

Table 4.1: Parameters of final filter, *filter7*.

Parameter	Range	Ligands Included
CHEMPLP Fitness	Top 10%	40919
Hydrogen Bond Angle ≥ 120°	1 : 6	21949
TPSA	75 : 110	15033
AlogP	-3 : +1	5390
Ligand Hydrogen Bond Count	1 : 7	25941
Occluded Donor Atoms	0 : 1	33603
Occluded Polar Atoms	0 : 1	33231
Ligand Clash Count	0 – 14	32399
PLP Ligand Torsion	0.027 : 3.000	28248
PLP Ligand Clash	0 : 2.000	31950
Chemscore Internal Correction	0 : 3.000	20529

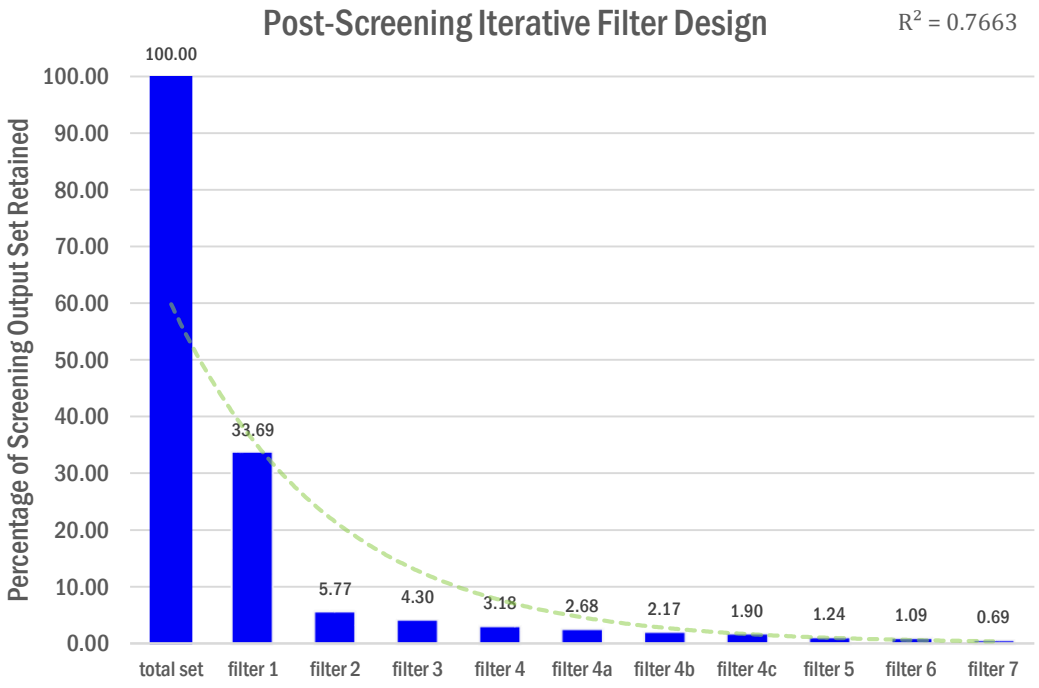


Figure 4.6: Bar chart displaying the effect of each iterative filter on the overall set. In this case "total set" is the 40919 ligands derived by retrieving the top 10% of ligands by fitness score. The trend-line in green indicates the exponential decrease through each iteration of filter, with the R-squared given as 0.77.

Through GoldMine, the virtual screening output of over 400,000 compounds had been reduced to a small set of potential binders which possess the ideal physiochemical properties and interaction characteristics with the protein. The filters implemented have resulted in exponential reduction to a small set of less than 300 ligands for manual inspection and selection.

4.2.5 Visual Inspection Aided by Mogul

Mogul utilises structural data in the CSD (currently containing over 700,000 small molecule crystal structures), to evaluate the poses of the selected ligands in conjunction with a general visual inspection to make an informed decision on the quality of the docking result. By comparing the torsions and ring structures of the ligand docking into EthR during virtual screening to the Cambridge Structural Database, those with unusual torsion angles could be evaluated for taking forward.

Similarly, general chemical knowledge and medicinal chemistry understanding was used in the evaluation of likely good hit candidates. For example, thiols are known typically for their use in chelating metals, known to cause adverse reactions and are generally highly reactive. For such compounds these potential disadvantages were considered against any favourable protein-ligand interactions, as troublesome chemical groups could be replaced by bioisosteres in lead optimisation stages if active against EthR; additionally, they were not excluded prior to screening for similar reasons.

Through this process of manual inspection and evaluation, the 284 ligands were reduced to 133 ligands, largely due to poor geometries in ligand orientation, as identified via Mogul.

The next step was to remove any duplicate scaffolds, in order for the cohort of compounds for biophysical screening to avoid treading the same ground; after all, for any active compounds identified, output at any stage of this screening pipeline would be probed for similar compounds for testing, and the duplicates which made it through to this part of the filtering would be the first choice for further investigation.

The 133 ligands were imported into KNIME[74] and the RDKit nodes used to calculate the Murcko scaffold[144] of each structure, which was subsequently converted into a SMILES string. By counting the number of unique SMILES strings duplicates could be quickly identified. For 124 ligands the scaffold occurred only once, which is ideal for the diversity of testing set desired. However, three scaffolds occurred twice each, and a fourth scaffold was present for three ligands.

For the ligands ZINC08980600, ZINC15789306 and ZINC06952722, the structures differ by the nature of their terminal group: a *t*-Butyl ester group, a dimethyl

group, or an ethyl ester respectively (figure 4.7). The scores were also similarly clustered between 86.55 and 84.94.

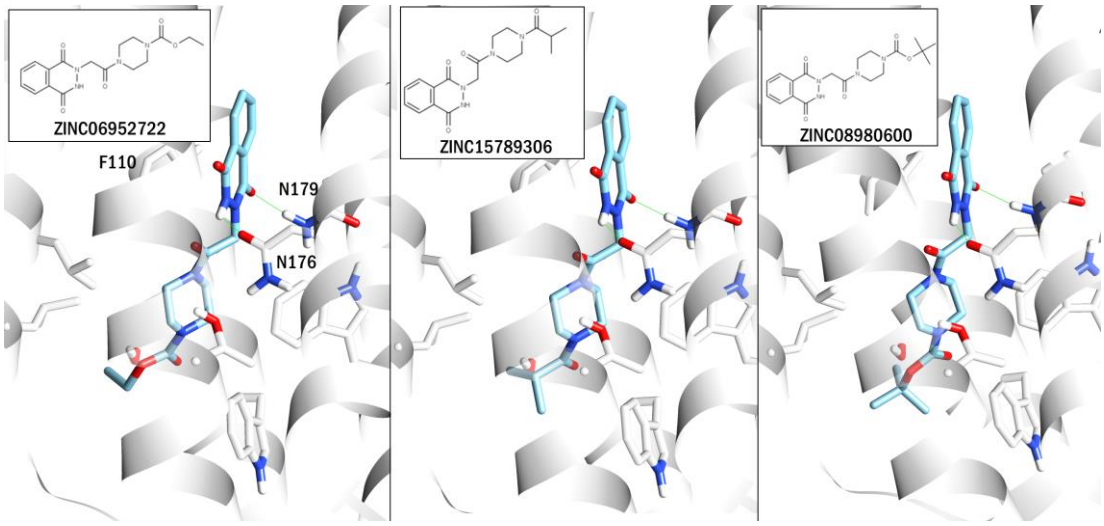


Figure 4.7: Ligands ZINC06952722, ZINC15789306 and ZINC08980600 which share a Murcko scaffold. All three form hydrogen bonds to Asn179 and Asn176, and pi-stack against Phe110.

The docking poses for each were examined and the differences were minimal; however, ZINC08980600 made better hydrophobic contacts, and the geometry of the hydrogen bonds formed was improved over the two scaffold duplicates, and so ZINC089080600 was taken forward while ZINC15789306 and ZINC06952722 were excluded.

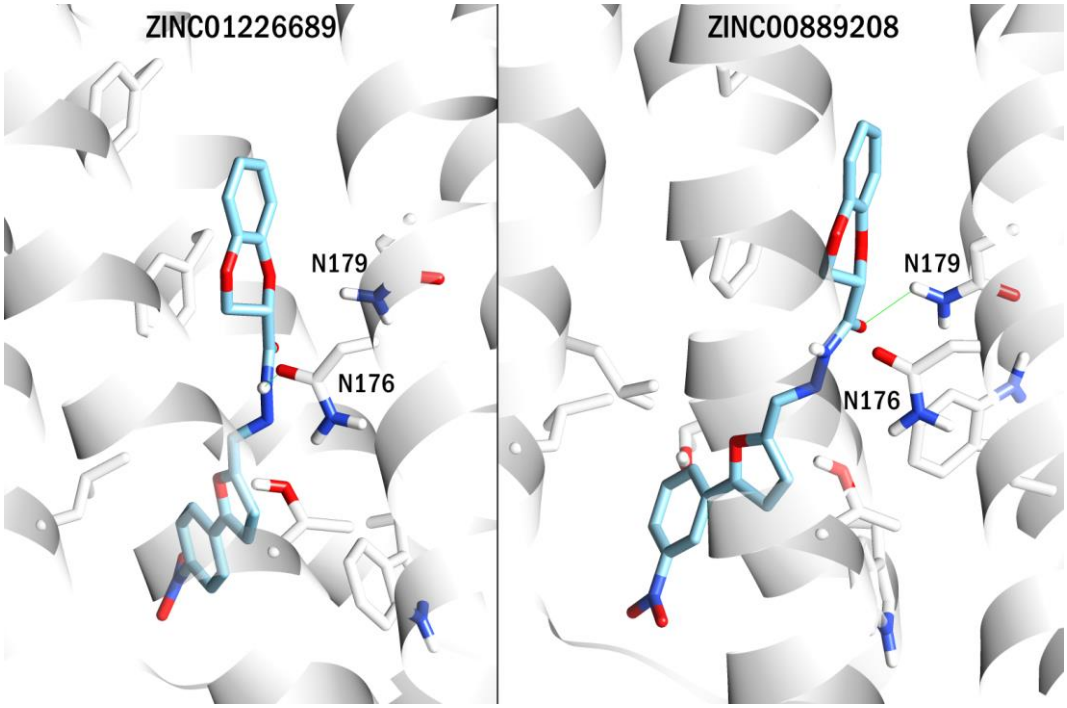


Figure 4.8: ZINC01226689 differed from ZINC00889208 in that the latter molecule possessed a nitro group in the meta group as opposed to the para position.

The decision between ZINC01226689 and ZINC00889208 was more straightforward, as they differed only in the position of the nitro group on the terminal phenyl ring (figure 4.8); the former made better hydrogen bond contacts and as such, scored far higher and so ZINC01226689 was included for biophysical assay against EthR.

Similarly, the ligands ZINC00073180 and ZINC00072059 (figure 4.9) differ in score by only four points and in structure by only a methyl group. While the score was higher for the methylated form, this group made no observable important contacts and so this small increase in score could be an artefact of an increased hydrophobic interaction surface. Moreover, the torsion of the amide function within the methylated form was shown to be more unusual via Mogul, and so it was decided to take the unmethylated form ZINC00072059.

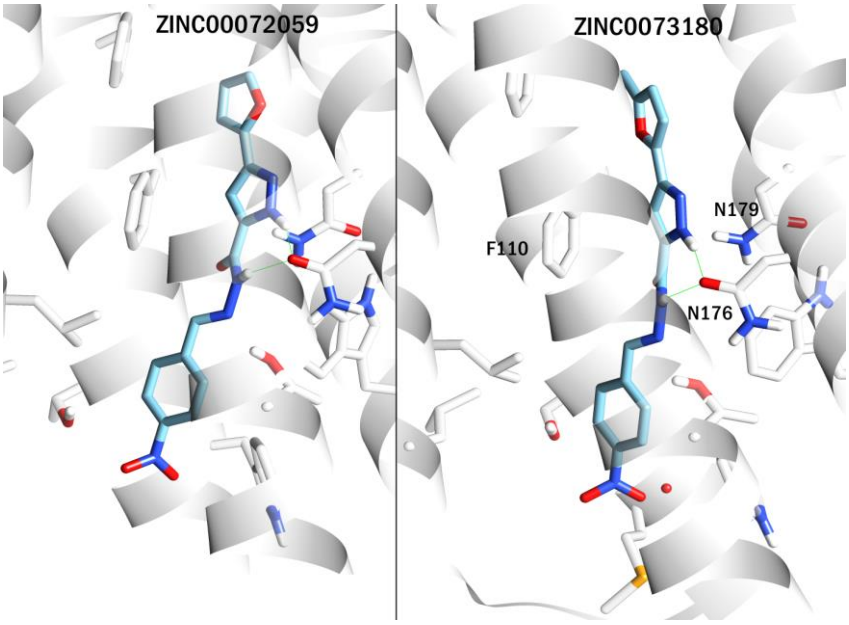


Figure 4.9: ZINC0072059 was included but ZINC0073180 was not, due to the former possessing more acceptable torsion around the amide function.

Finally, based upon the favourable contacts made between the far-higher scoring ZINC10777920 with Asn179 and Tyr148 in the protein, this compound was chosen over closely-related ZINC15754437.

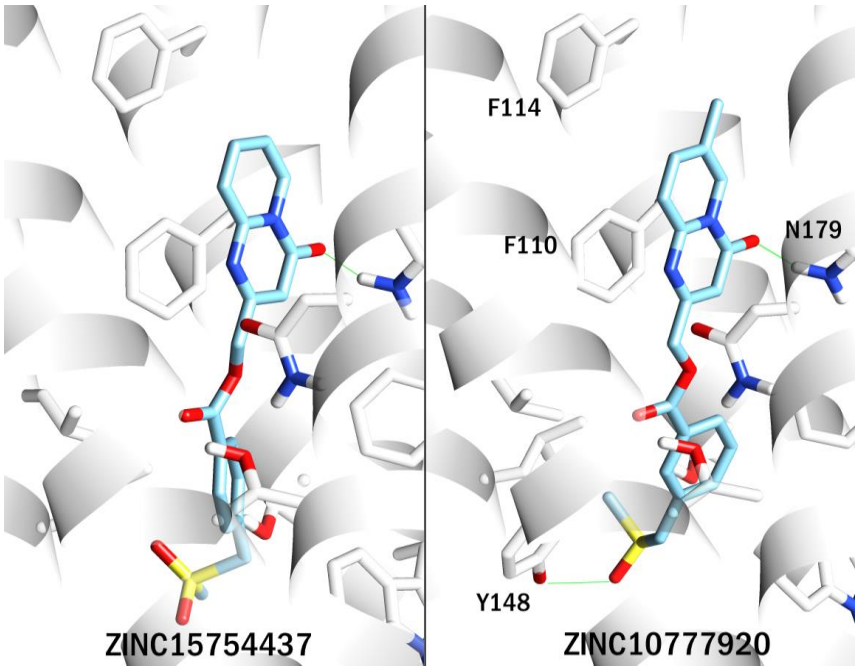


Figure 4.10: The protein-ligand contacts made by ZINC1077920 were far more favourable than the closely related structure ZINC15754437, as methyl sulphone groups are typically poor hydrogen bond acceptors.

A final total of 128 ligands were selected, however many compounds were no longer commercially available and therefore the final cohort of molecules for biophysical testing against EthR, as listed in appendix B, consists of 85 compounds.

CHAPTER V: THERMAL SHIFT ASSAYS ON POTENTIAL ETHR INHIBITORS

Extrinsic fluorescent dyes, which rely on covalent or non-covalent probes rather than the inherent protein fluorescence from residues such as tryptophan or tyrosine, can be used for a variety of purposes in measuring the protein solution state. Covalent probes are typically used for measurements such as FRET, whereas non-covalent probes are utilised for a variety of other purposes: for example, non-covalent fluorescence probes have been used to monitor surface hydrophobicity;[146] as active site probes;[147] to monitor chemical degradation;[148] to characterise protein aggregation and fibrillation events;[149,150] to characterise proteins and differentiate between closely-related protein family members;[151] and to monitor folding and unfolding processes.[152] Dyes typically work by charge transfer from an electron donating group to a cyclic system[153,154] when the dye moves from a polar (ie. aqueous) environment to a more hydrophobic, apolar environment (ie. when hydrophobic protein core residues are exposed during unfolding or conformational change).

This chapter details the use of such non-covalent fluorescent molecular probes in identifying compounds which bind a target protein. This is possible using a thermal shift assay (TSA), also known as Differential Scanning Fluorimetry (DSF)[151] or Thermofluor.[155] Figure 5.1 shows a typical protein melting curve, represented by heating EthR in the presence of fluorescent probe SYPRO Orange (Invitrogen).

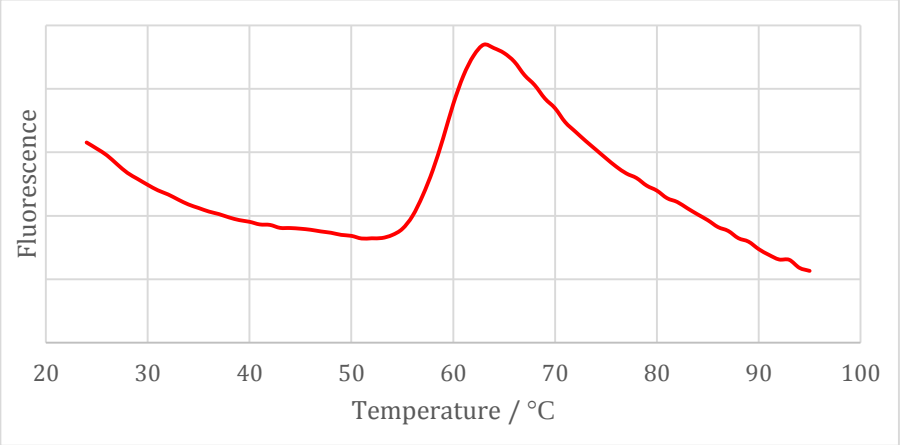


Figure 5.1: Protein melting curve (native EthR). The protein remains folded and the probe largely quenched until the T_m of the protein is reached and hydrophobic core residues are exposed.

The protein melting curve is a standard denaturation curve, monitored by the fluorescent probe. Theoretically, as the temperature increases, the protein gains more kinetic energy but does not unfold. When the melting temperature (T_m) of the protein is approached, the protein begins to unfold and denature, exposing the usually-buried hydrophobic residues of the protein core to the aqueous environment. This causes the probe to be unquenched and an increase in fluorescence is directly proportional to the unfolding of the protein, reaching a maximum when all the protein is unfolded. A decrease in fluorescence is then observed at high temperatures due to aggregation of the degraded protein products which serves to once again sequester hydrophobic residues and quench the probe in solution.

This method has been used in the past to determine optimum stabilising solutions for crystallisation,[156,157] evidenced as a positive shift in the curve and resulting T_m. In recent years, this has been adapted for high-throughput screening of potential binding compounds in drug discovery programs, due to its scalability, simplicity of execution, and low protein and compound consumption.[157] Rather than, in crystallisation screening Thermofluor assays, changing the buffer and metal conditions, a ligand screening assay keeps the buffer conditions constant and varies only the compound to be screened.

5.1 Thermal Shift Assays on EthR

The first use of a thermal shift assay with EthR was reported in 2011,[123] to compare the wild-type functional protein with a site mutant G106W which was unable to bind DNA, and each with ligands BDM31343, BDM31381 and BDM14801 to demonstrate their effectiveness in binding the wild-type but not the G106W mutant (figure 5.2).

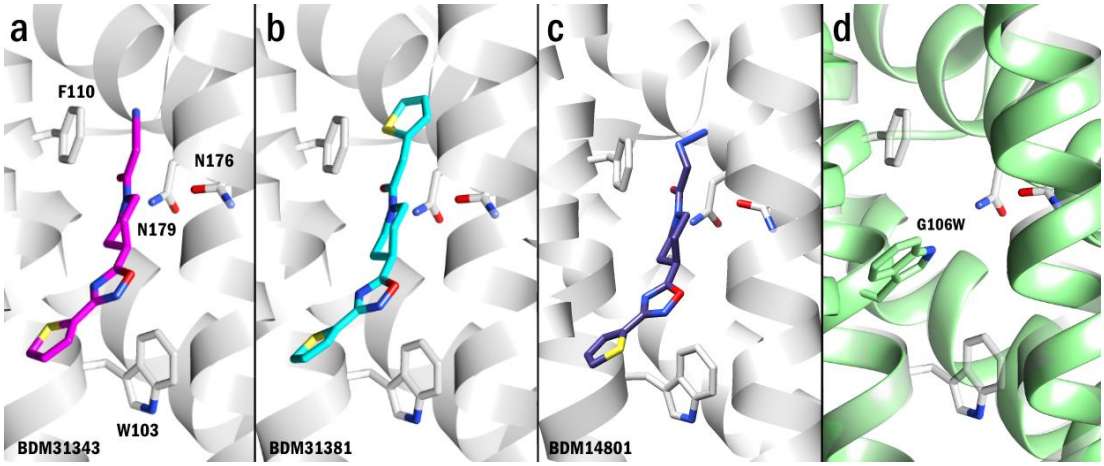


Figure 5.2: Ligands BDM31343,[123] BDM31381[119] and BDM14801[124] bound to EthR. The G106W mutant precludes ligand binding by blocking the binding site, and induces the inhibitory conformation of EthR.[123]

Carette *et al*.[123] utilised proprietary dye SYPRO Orange (Invitrogen) to monitor the protein unfolding and a Lightcycler 480 (Roche), a real-time PCR apparatus for microtitre plates. Samples were heated from 37 to 85°C, at a rate of 0.04°C/s. Fluorescence was measured at excitation/emission wavelengths of 465/510 nm. The protein and buffer conditions are detailed below in section 5.3. In this case, the thermal shift assay was used for corroborating the biophysical behaviour of a mutant EthR; in 2012 and 2014 the assay was adapted for the discovery of inhibitors with novel scaffolds, and fragment inhibitors respectively.

For the identification of novel N-phenylphenoxyacetamide derivative inhibitors[129] the authors used a thermal shift assay to screen 22 compounds, identified through phenotypic screening, against a reporter-modified strain of *Mycobacterium smegmatis*. All of these compounds demonstrated, in dose-response experiments, IC_{50} values below 10 μM. Six compounds, of the 22 screened by thermal shift, showed a significant change of greater than 4°C in the melting temperature. Among these, three were members of this novel N-phenylphenoxyacetamide family, with the best activity being an IC_{50} of 2.9 μM and ΔT_m of 6.4°C.[129]

In the development of fragment-derived inhibitors for EthR, Villemagne *et al*.[131] identified a small fragment used in previous click-chemistry approaches[124] as having a weak binding affinity for EthR. This fragment showed a very small shift of 0.1°C to the T_m, a weak but notable difference given the sensitivity and accuracy afforded due to a 0.04°C/s temperature change rate. Orthogonal testing with SPR determined the fragment in question to have a 160 μM IC_{50} for the inhibition of DNA binding by EthR. Subsequent fragment growing and linking approaches yielded eleven small compounds with a range of ΔT_m from 0.56 to 6.1°C, however only three of the initial eleven demonstrated an ability to boost ethionamide activity in *M. tuberculosis*-infected macrophages at an EC_{50} below 10 μM. Subsequently one of these compounds was taken forward for development and of the five resulting compounds which demonstrated the ability to shift the T_m by 2.3 to 11.2°C, only three were able to effectively boost ethionamide.[131]

This aptly demonstrates the supposed propensity of thermal shift assays toward false-positive inhibitor identifications.[158] Thermal shift assays are often misunderstood as a method of inhibitor identification but this is not the case: thermal shift assays identify molecules which alter the melting temperature of the target protein. Ideally, molecules which do this are those which bind the protein and would have an inhibitory effect; however, it is also possible that these molecules form transient interactions or otherwise alter the aqueous environment to stabilise or destabilise the protein, without

true binding. It is therefore important when conducting thermal shift assays to be mindful of the limitations of the method, and to utilise orthogonal biophysical methods to confirm 'hits'.

5.2 Expression of EthR for Biophysical Assays

The thermal shift assays were conducted over multiple batches of protein and in triplicate. Initially, EthR used was obtained from the Baulard group (Institut Pasteur de Lille), however later batches were expressed in-house using the pET-15b-*ethR* plasmid provided by the Baulard group and prepared as previously described.[121] The plasmid encodes EthR with an N-terminal histidine tag with the sequence MGSSHHHHHHSSGLVPRGSHM.

Protein was expressed using *Escherichia coli* BL21(DE3) cells, to which pET-15b-*ethR* was transformed. *E. coli* was grown in LB broth inoculated with ampicillin to an OD600nm of 0.6-0.7, at which point isopropylthiogalactosidase (IPTG) was added to a final concentration of 1 mM, and grown for an additional three hours. Cells were harvested by centrifugation at 12000 g at 4°C and subsequently re-suspended in lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, pH 7.5, 10 mM imidazole) and lysed by two passages through a French Pressure cell at 6.2 x 10^6 Pa, followed by centrifugation at 20000 g for 25 min at 4°C. The supernatant was recovered and histidine-tagged EthR was separated from the whole-cell lysate using Ni-NTA agarose chromatography (Qiagen). His6-EthR was eluted from the column with 100 mM imidazole in lysis buffer and dialysed overnight against 50 mM NaH_2PO_4, 300 mM NaCl at 4°C for subsequent separation on a Superdex 200 gel filtration column (GE Healthcare) to ensure maximum purity. SDS-PAGE gels were used to confirm good purity and confirm protein presence at around 24 kDa. Identity of the protein was confirmed by in-gel trypsin digest and subsequent MALDI-TOF mass spectrometry.

Approximately 1 mg ml^{-1} of purified protein in a total volume of approximately 50 mL was obtained for four litres of initial growth, as determined by Nanodrop (ThermoScientific). Protein was dialysed against 10 mM TRIS HCl pH 7.5, 200 mM NaCl and stored at 4°C until use, concentrated as needed by spin concentration, operated according to the manufacturer's recommended parameters (Vivaspin, Generon).

5.3 Protocol Implementation: SYPRO Orange Excited at 505-535 nm

Using a protocol for thermal shift assays developed in-house on both lysozyme and glucose isomerase,[155] initial tests with DMSO-only and positive controls were conducted

to determine if previous experiments by Flipo *et al.*[129] could be replicated in Durham, and that melting curves for EthR could be obtained.

Initially, this combined the protein concentrations and preparations from Flipo *et al.*[129] with the operational protocol of the Applied Biosystems Fast 7500 qPCR for differential scanning fluorimetry.[155] Experiments were conducted in 96-well PCR plates (ThermoScientific) and sealed with polyolefin film (ThermoScientific). Protein and SYPRO Orange were mixed to an initial concentration of 20 μM and 5X respectively, with this solution and the 40 μM compound solution being added to the well in a 1:1 ratio. Final sample concentrations for negative controls were 10 μM (0.24 mg ml^{-1}) EthR, 2.5X SYPRO Orange, 1% DMSO, and for positive controls included 20 μM ligand. Samples were then heated from 24 to 95°C with a heating rate of 1°C/min. The fluorescence was measured at Ex/Em = 510/510 nm, and data interpreted by NAMI[155] to give the melting temperature, T_m.

This protocol adapted from Flipo *et al.*[129] resulted in high background fluorescence, and poor melting curves for unbound EthR (figure 5.3).

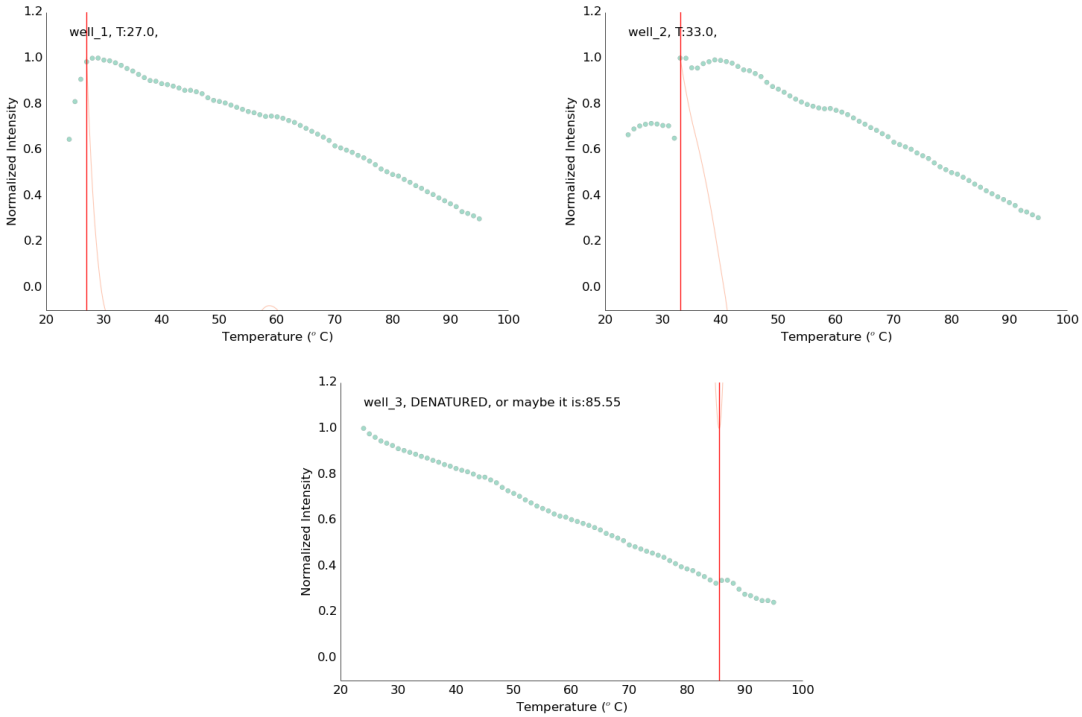


Figure 5.3: Melting curves (green) for unbound EthR as provided by NAMI.[155] High background fluorescence obscures any signal. The green points show the raw fluorescence at each degree, the pale pink line is the result of interpolation to calculate the rate of change between points, and the solid red line indicates suggested melting temperature.

Additionally, very weak transitions were observed for known low-affinity binders (such as BDM31343) although strong transitions were identified for known high-affinity

binders to EthR (such as BDM41906). Examples of these curves as analysed by NAMI can be seen in figure 5.4. However without a protein-only control, no shift could be determined from these curves.

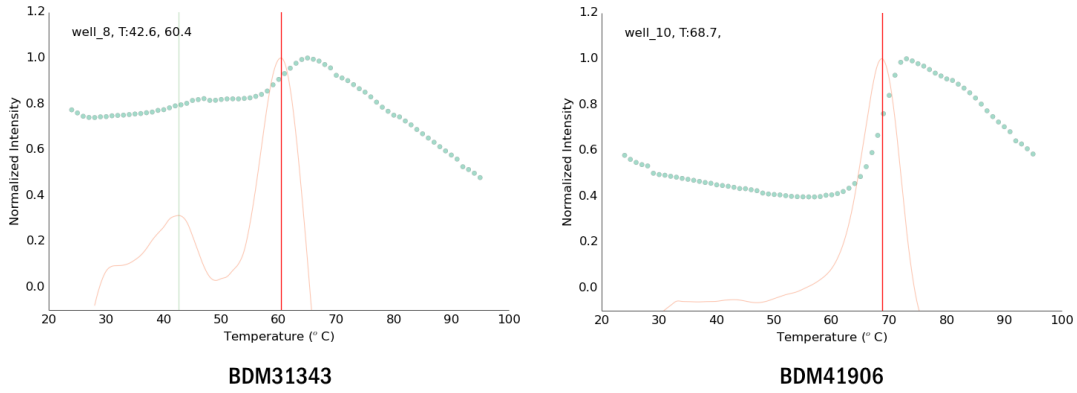


Figure 5.4: Melting curves (green) for EthR with BDM31343, left, and BDM41906, right. Despite weak transition for BDM3143, NAMI[155] was able to interpolate a melting temperature (red line).

Initially it was supposed that EthR, possessing a hydrophobic binding site, could be sequestering the SYPRO Orange such that weaker-binding compounds were unable to displace it; the high background fluorescence observed could therefore be a result of unquenched SYPRO Orange in the EthR binding site. However, adaptation of the protocol to allow the protein and target compound to incubate for 30 minutes prior to addition of SYPRO Orange showed no improvement, and where the protein had been incubated without compound for the same amount of time, only a weak transition was identified around 59°C. This would not be sufficient to identify low affinity ligands, nor is it a reliable negative control. Therefore, alternative strategies were investigated.

5.4 Protocol Optimisation: SYPRO Orange Excited at 455-485 nm

At an excitation wavelength of 510 nm, it was not possible to derive a melting curve for EthR, either in the native state without a ligand or bound to weak ligands. However, this protocol had worked for other model proteins and so alternative dyes were considered. Firstly, SYPRO Red (Invitrogen) and later Bis-ANS (Sigma Aldrich).

SYPRO Red is a molecular probe typically used, like SYPRO Orange, for protein gel staining; they are highly similar, but SYPRO Red is reported to have lower background fluorescence which was deemed desirable for use with EthR. Another key difference between the two SYPRO dyes are their excitation and emission spectra – SYRPO Red excites at a higher wavelength (maxima ~ 550 nm compared to ~ 480 nm for SYPRO Orange) and therefore the protocol described above in section 5.3 would not have to be

adapted. However, this too proved unsuitable, as comparison with a lysozyme control demonstrated weak transitions and high instrinsic fluorescence for EthR similar to that observed with SYPRO Orange (data not shown).

Therefore, the decision was taken to reduce the excitation wavelength to this range (Ex/Em: 485/510 nm) and test SYPRO Orange with 10 μM EthR under varying concentrations of dye. This experiment was only run from 51-95°C as the EthR melting temperature was known to be in the region of 56-59°C from previously published data.[159] Cconcentrations of 20X SYPRO Orange down to 2.5X SYPRO Orange were able to give good melting curves for EthR with even a concentration of 1.25X sufficient for NAMI to derive a melting temperature for EthR (56.3 ± 0.2°C).

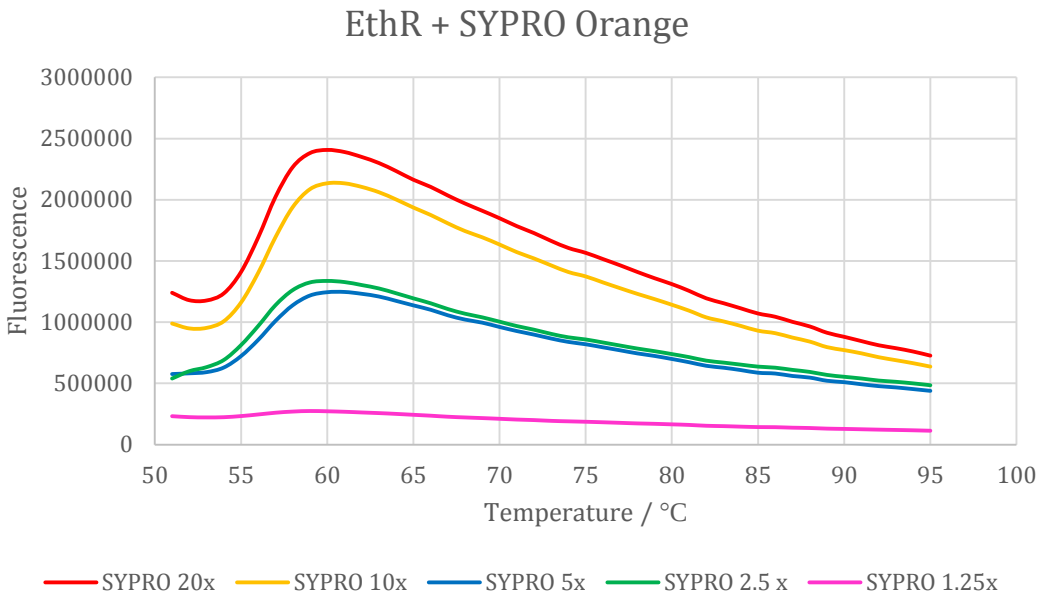


Figure 5.5: 10 μM EthR with varying concentrations of SYPRO Orange dye at Ex/Em: 485/510 nm.

To determine the optimum EthR and SYPRO Orange concentrations, a grid screen-style thermal shift was used. Concentrations of EthR from 1.8 to 0.056 mg ml^{-1} were tested with decreasing concentrations of SYPRO Orange from 20X to 0.625X. Figure 5.6 shows the results as the table output from NAMI,[155] annotated from one experiment.

From figure 5.6, it is clear to see that at low protein concentrations, high SYPRO Orange concentrations provide too high a background fluorescence to accurately differentiate the protein melting curve signal. Conversely, at low SYPRO Orange concentrations, the dye content is not sufficient to give enough signal upon protein unfolding.

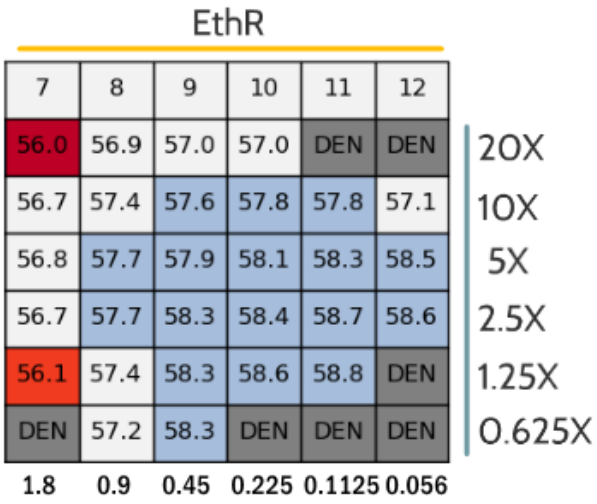


Figure 5.6: Partial table output from NAMI for EthR (1.8 – 0.056 mg ml^{-1}) vs. SYPRO Orange (20X – 0.625X) screen. "DEN" denotes denatured samples or those curves for which NAMI could not assign a T_m. An estimated reference temperature of 57.5°C and a range of 0.5° was defined to colour code stabilising (blue) and destabilising (red) conditions, and highlight the ideal region of protein:dye concentrations.

The ideal range for SYPRO Orange concentration therefore falls between 10X and 2.5X. The data for 5X across all concentrations of EthR gave the most promising melting curves, shown as figure 5.7, below.

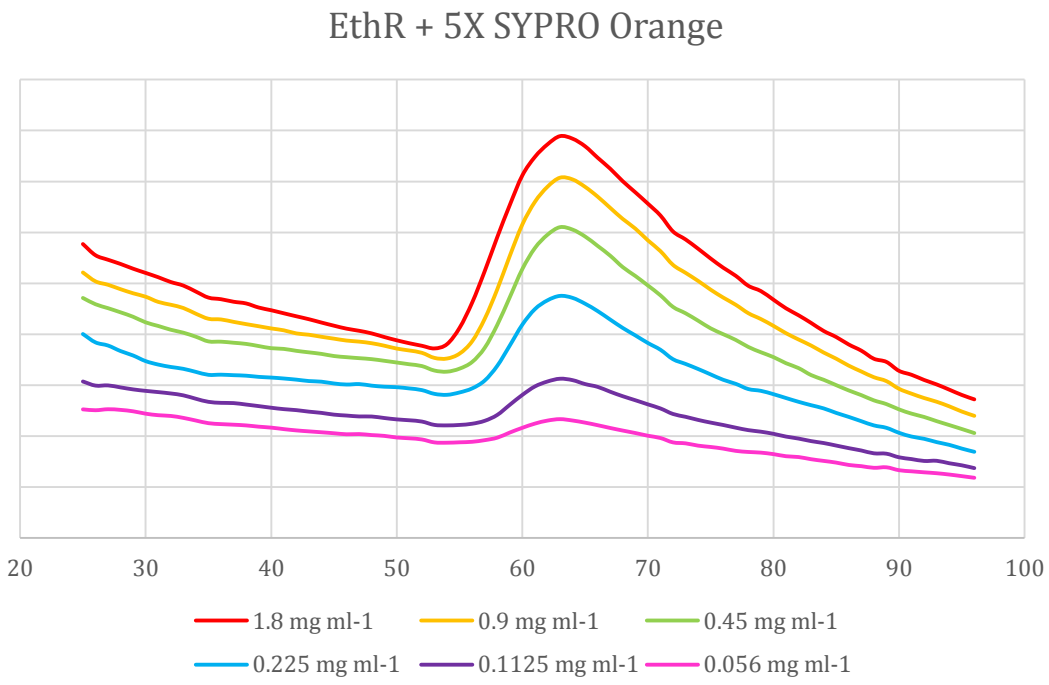


Figure 5.7: Fluorescence melting curves for varying concentrations of EthR with 5X SYPRO Orange.

The best compromise between SYPRO Orange concentration and EthR concentration, which gave a deep and clear transition under melting, was chosen as 0.9 mg ml^{-1} with 5X

SYPRO Orange (orange curve in figure 5.7, second from the top). For future experiments, the concentration of EthR was rounded to 1 mg ml^{-1}.

For thermal shift assay, protein was concentrated to 2.5 mg ml^{-1} by spin concentration (Vivaspin, Generon). Each well contained 8 µL of the 2.5 mg ml^{-1} EthR stock, 2 µL SYPRO Orange (diluted to 50X from 5000X stock, Invitrogen) and 10 µL of ligand sample. This gave final concentrations of 1 mg ml^{-1} EthR, 5X SYPRO Orange and ligand concentrations at either 400 µM, 320 µM, 160 µM or 40 µM.

Each 96-well plate was used to assay up to 23 compounds plus one negative control, arranged such that A1, B1, C1, and D1 for example all contained the same compound in decreasing concentrations. Samples were then heated from 24 to 95°C with a heating rate of 1°C/min. The fluorescence was measured at Ex/Em = 485/510 nm, and data interpreted by NAMI[155] to give the melting temperature, T_m.

5.5 Results: Biophysical Identification of Potential EthR Binders

For each of the 85 compounds tested and at each concentration (400, 320, 160 and 80 µM), thermal shift assays were conducted in triplicate. Full tabulated results for each compound in triplicate can be found in Appendix B.

By using multiple concentrations of compound it was possible to assess which shifted the melting temperature of EthR in a concentration-dependent manner; this would indicate compounds which were truly affecting the protein melting temperature. A total of 42 negative control melting temperatures were taken for an 'average' EthR melting temperature of 59.3 ± 1.2°C in the absence of compound. Similarly, triplicate melting temperatures for each compound at each concentration were averaged, and the melting temperature of EthR subtracted to give the shift (ie. ΔT_m). A total of twenty compounds from the initial screening cohort of 85 were selected as demonstrating a concentration-dependent shift in the melting temperature of EthR (figure 5.8).

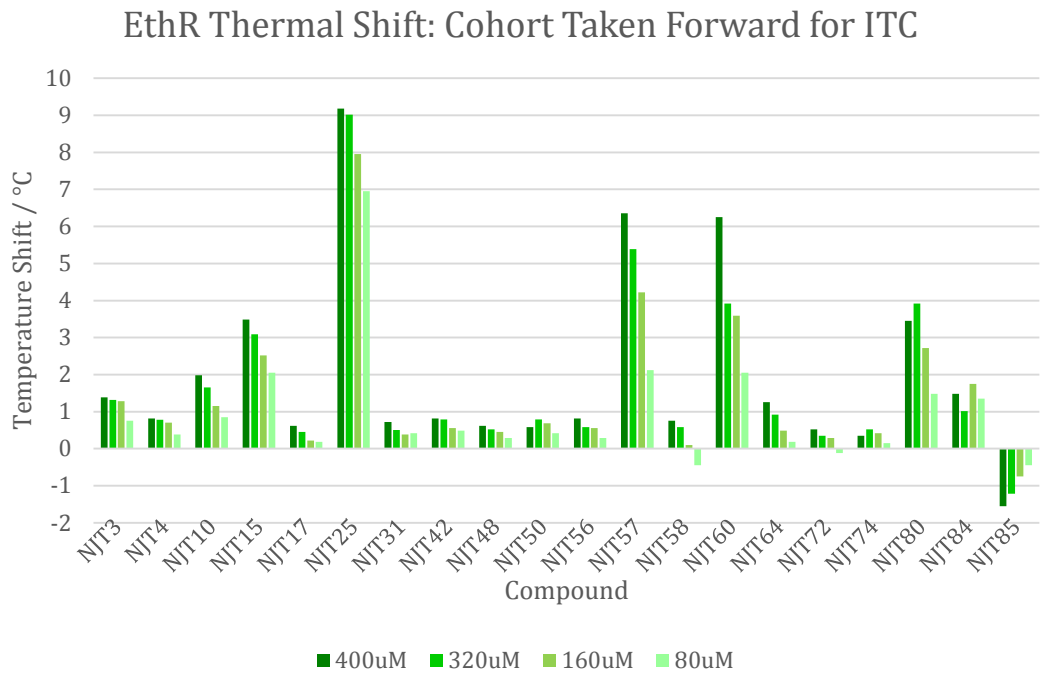


Figure 5.8: Cohort of 'hits' represented as a bar chart of compound concentration and ΔT_m.

Selection criteria was lenient with dependent-like shifts in order to include a range of promising hits and potential binding affinities. Compounds were screened using a 'working code' in order to prevent confusion between long alpha-numerical ZINC codes. Table 5.1 correlates the working code used during experimentation (NJT#) with the ZINC code of the compound.

Table 5.1: Working codes of thermal shift 'hit' compounds with the original ZINC code.

Code	ZINC	Code	ZINC
NJT03	ZINC67692217	**NJT56**	ZINC08987568
NJT04	ZINC65406277	**NJT57**	ZINC09689958
NJT10	ZINC67974892	**NJT58**	ZINC71795591
NJT15	ZINC00237604	**NJT60**	ZINC14848503
NJT17	ZINC32576321	**NJT64**	ZINC69416783
NJT25	ZINC06726755	**NJT72**	ZINC03420970
NJT31	ZINC24615603	**NJT74**	ZINC08980600
NJT42	ZINC53275627	**NJT80**	ZINC01502258
NJT48	ZINC69489717	**NJT84**	ZINC09087663
NJT50	ZINC12794882	**NJT85**	ZINC19234873

5.5.1 Temperature-Dependent Thermal Shifts

The twenty selected 'hits' demonstrate the ability to shift the melting temperature of EthR in a concentration-dependent manner. Figure 5.9 shows the top five compounds which shift the melting temperature of EthR by greater than 3°C.

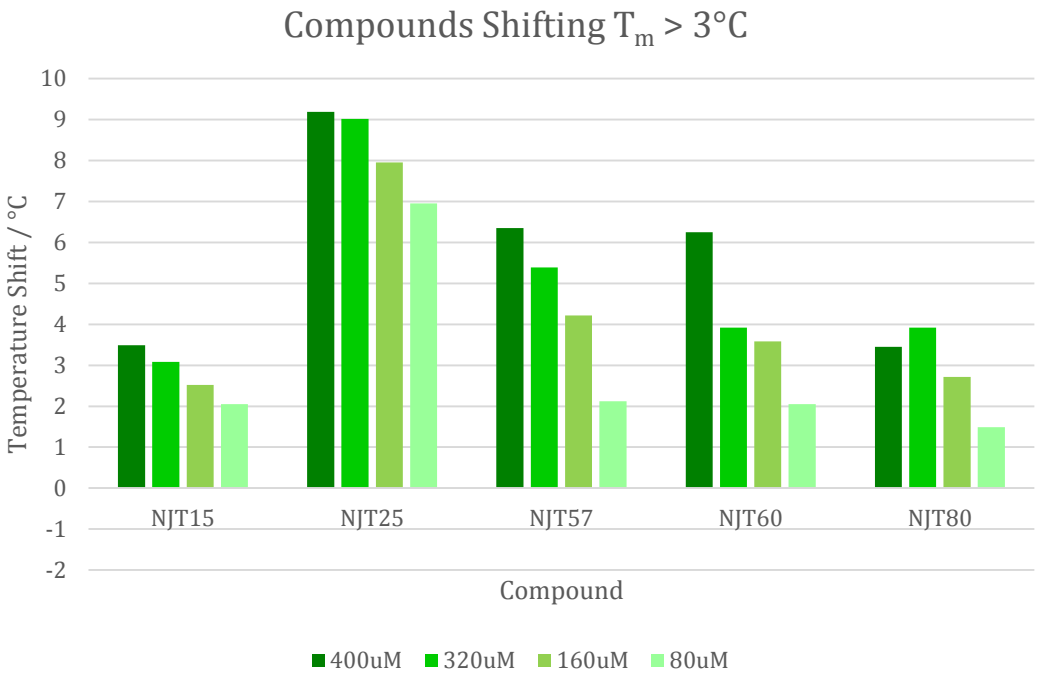


Figure 5.9: Cohort of 'hits' which showed a ΔT_m of greater than 3°C at 400 µM.

Compound NJT80, though 400 µM does not follow the concentration-dependent pattern, is included in the top five compounds as the shift is still greater than 3°C. The second cohort of five compounds are able to shift the melting temperature of EthR by greater than 1°C (figure 5.10). Interestingly, compound NJT85 is the only compound which causes a concentration-dependent *negative* shift in the melting temperature of EthR, suggesting a destabilising rather than stabilising interaction effect on the protein.

A proportion of 50% of the chosen 'hit' compounds demonstrate an ability to shift the melting temperature of EthR, in a concentration-dependent manner, to a measure of 1°C or greater.

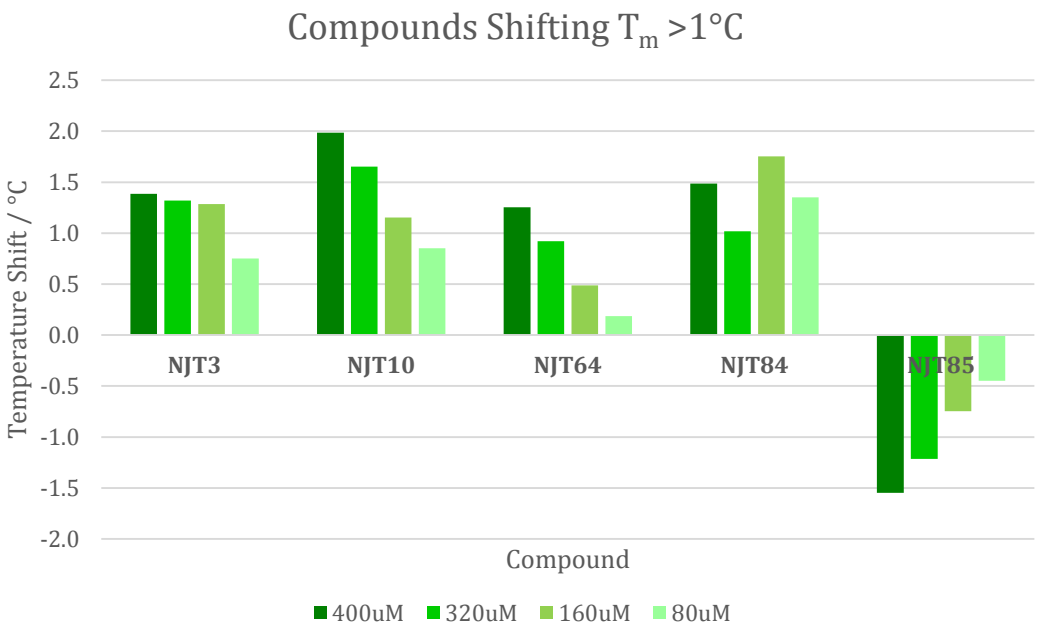


Figure 5.10: Cohort of 'hits' which showed a ΔT_m of greater than 1°C at 400 µM. NJT85 is notable as the only compound which showed a concentration-dependent *destabilisation* (or negative-shift) in EthR T_m.

5.5.2 Small-Shift Compounds

A total of ten compounds produced shifts smaller than 1°C and so within the error of the EthR melting temperature (59.3 ± 1.2°C), but were included as thermal shift 'hits' due to a concentration-dependent pattern to the shift in melting temperature caused by the compound. In this preliminary stage of 'hit identification' the decision was taken to be as generous and inclusive as possible in order to maximise the chances of identifying a range of promising EthR-binding compounds.

5.6 Summary

Of the 85 compounds tested from the virtual screening results, a total of twenty are able to shift the EthR melting temperature (in a positive, stabilising direction with one exception) with a range of success. The structures of these compounds can be found in appendix B, with the associated binding data as determined by SPR (chapter seven). Five compounds in particular show great promise, shifting the EthR melting temperature by greater than 3°C: NJT15, NJT25, NJT57, NJT60 and NJT80. Additionally, compound NJT85 demonstrated a negative shift in protein melting temperature, suggesting a destabilising effect upon binding. All twenty, with keen emphasis on these six most promising and interesting hits, were taken forward into co-crystallisation trails and for preliminary assay by additional biophysical methods; namely surface plasmon resonance (SPR).

CHAPTER VI:
SURFACE PLASMON RESONANCE STUDIES

Surface plasmon resonance (SPR) is a chip-based method of binding and kinetic assay for biophysical interactions, adaptable for a variety of interactions.

SPR measures biomolecular interactions in real-time, and technically label-free.[160] SPR can be used to determine interaction affinities, association and dissociation rates, and determine stoichiometry. This is achieved in practice due to the reflection of incident light from a coated metal-filmed chip. Chips for SPR have a variety of functionalised surfaces on which a binding partner can be immobilised, and subsequent binding of an analyte introduced at a constant flow rate over the chip surface causes changes in the refractive index of the incident light; the sensorgram of this interaction is provided in real-time.

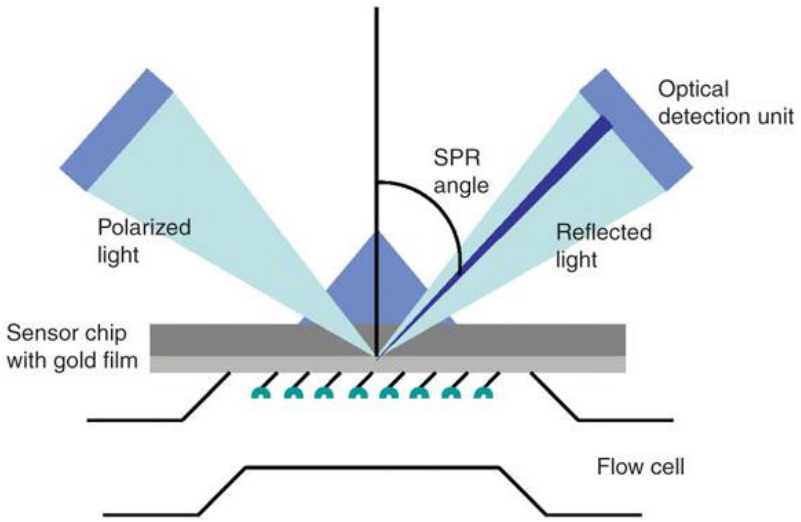


Figure 6.1: Surface plasmon resonance utilises functionalised chip surfaces to which ligands are immobilised. Analytes can then be injected over the flow cell, and the change in SPR angle of the reflected light is directly proportional to molecular mass change.[161]

The change in magnitude of the signal is directly proportional to the mass of the molecule binding, and the rate of change in the sensorgram can be interpreted for association, dissociation and equilibrium rate constants.

The BIACore 3000 system (www.biacore.com, GE Healthcare) uses chips consisting of a glass slide with a thin gold film, with four channels known as "flow cells". The chips are functionalised for immobilisation, for example the CM5 chip for covalent amine coupling; the Ni-NTA chip for polyhistidine-capture; and the SA chip functionalised with streptavidin, for biotin-tag capture.

This chapter details the use of SPR to perform preliminary assessment of the ability of potential EthR inhibitors (identified via virtual screening and subsequent thermal shift assays) to bind EthR.

6.1 Confirmation and Quantification of EthR-Binding by SPR

SPR has been used as a functional assay to determine IC_{50} values for EthR inhibitors,[119] as well as to kinetically characterise their binding to EthR.[159] Notably, Crauste *et al*. demonstrated the conformational change EthR undergoes upon binding to inhibitors is such that changes in the response unit were dose-dependent, and therefore EthR can be immobilised on a CM5 chip and binding can be determined by direct capture.[159] Additionally, they observed unexpected negative signals with the characteristic shape which they analysed successfully. Here, the method developed by Crauste *et al*. was implemented to test the novel binding compounds identified by virtual screening. However, without the EthR mutant G106W to use as a negative control, the reference cell used is simply unmodified surface lacking immobilised EthR.

6.2 Method: Immobilisation of EthR

EthR (20 μg/ml in 10 mM sodium acetate, pH 4.5) was immobilised on a CM5 chip in HBS-EP buffer supplemented with 1% DMSO by amine coupling according to recommended protocols to a target of 2000 RU. It is recommended kinetic experiments are performed at the lowest manageable immobilisation level, and an immobilisation level of 2000 RU would give a theoretical response of 28 RU for a 350 Da analyte. This corresponds to approximately 2 ng/mm^2 of EthR on the chip surface.

The Biacore surface preparation wizard was utilised for amine coupling immobilisation on flow cell 4, utilising flow cell 3 as a control. Briefly, 70 μL of 0.2M EDC and 0.05 M NHS was injected at a flow rate of 10 μL/min to activate the surface, followed by several short injections of EthR to achieve the desired immobilisation level (figure 6.2). Once reached, the surface was deactivated with a 70 μL injection of 1 M ethanolamine pH 8.5. EthR was immobilised to a final level of 2245 RU.

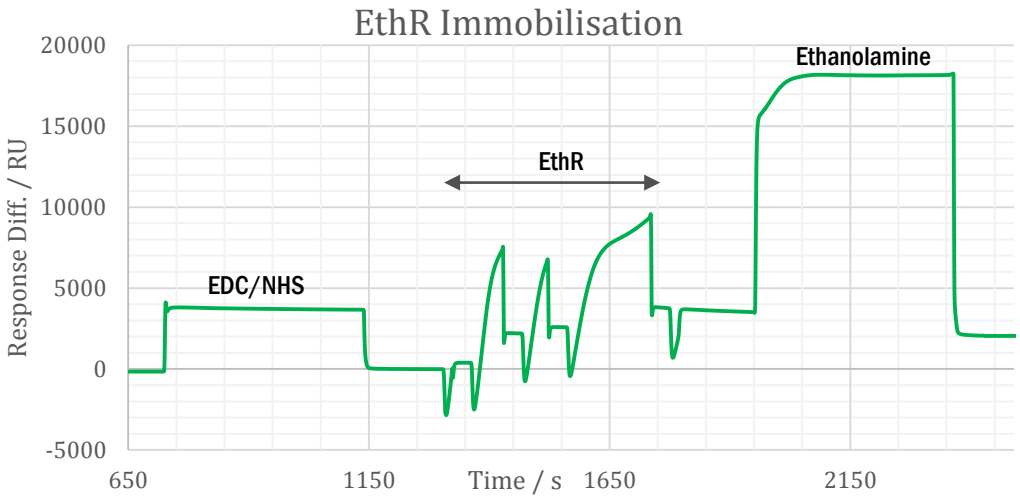


Figure 6.2: SPR Sensorgram of EthR immobilisation on CM5 chip by amine coupling utilising EDC/NHS for activation, and ethanolamine for deactivation (labelled).

6.2.2 Method: Determination and Quantification of Compound Binding

Per the method described by Crauste *et al.*,[159] ligands were injected for five minutes at a flow rate of 20 µL/min, a sufficient time to observed equilibrium for at least two concentrations. Figure 6.3 shows the type of sensorgram observed by Crauste *et al.*[162]

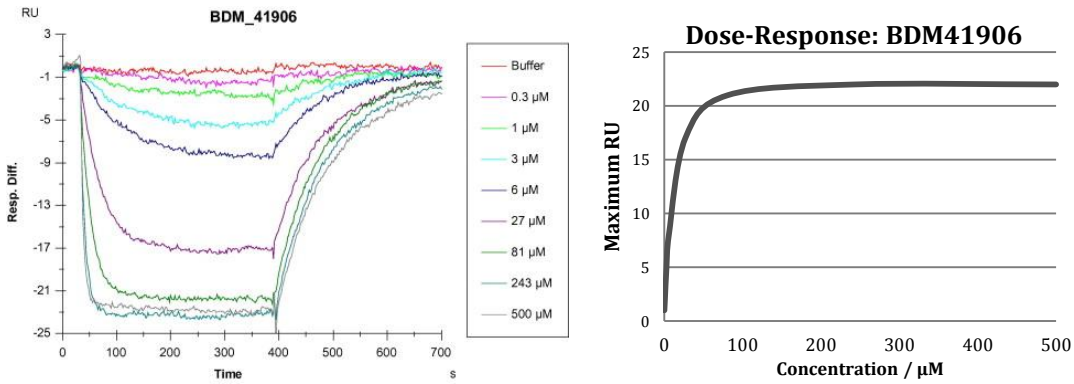


Figure 6.3: Example sensorgram from Crauste *et al.*[162] and a corresponding dose response curve (values multiplied by -1, as treated by authors in data processing).

Due to limited compound stock, it was not possible to assay a wide range of concentrations, nor was it possible to always perform experiments more than twice. Concentrations used therefore vary between 250 and 5 µM, and experiments were performed at least in duplicate, and in triplicate where possible.

Compounds were injected over the chip surface at a flow rate of 20 µL/min. Ligands were prepared in HBS-EP buffer with a final DMSO concentration of 1% to match running buffer. Regeneration of the surface was observed during the course of the

running buffer, therefore a period of 20 minutes (15 minutes dissociation, 5 minutes regeneration time) was used to recover initial baseline. Figure 6.4 shows an example sensorgram, from injection start to the end of the dissociation phase.

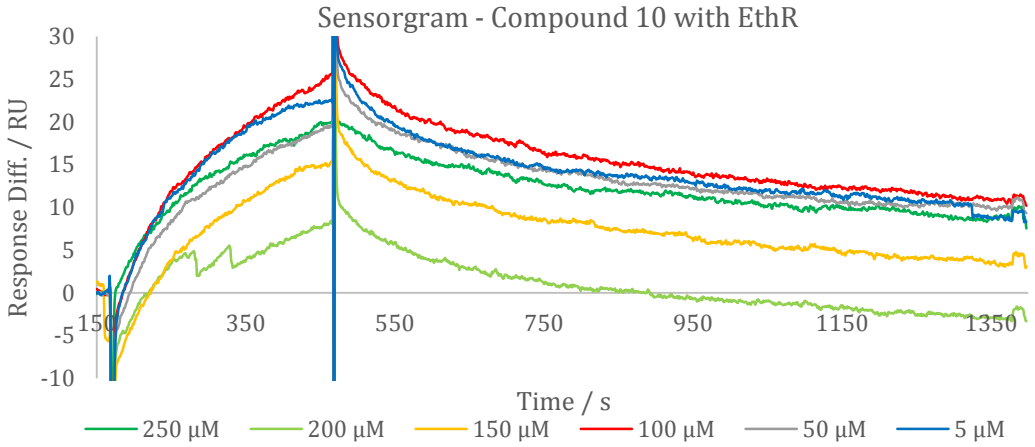


Figure 6.4: SPR sensorgram of compound 10 binding to EthR. Large difference injection start/stop peaks, arising from the small time delay between the two channels, are omitted for clarity. Programmed dissociation phase ends at 1400 s, 15 mins from injection end.

6.3 Results

Full raw spectra for each compound can be seen in Appendix C. Initially, separate fitting of ka and kd was performed in order to obtain estimates of these values for global fitting; subsequently global fitting was performed to the 1:1 Langmuir binding model with baseline drift. In this manner, estimations of the binding constants could be made, and are given in table 6.1.

However it can be observed from figure 6.4, exemplary of many sensorgrams obtained, that the magnitude of binding does not correlate with concentration as would have been expected. Maximum response unit values were extracted for each curve at the end of injection and were plotted against ligand concentration (figure 6.5).

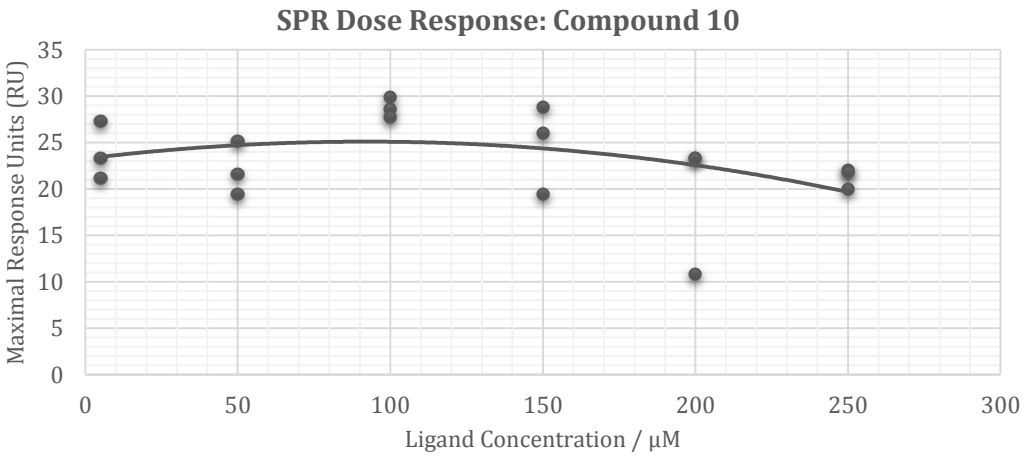


Figure 6.5: Dose-response curve for compound 10, corresponding to figure 6.4.

The corresponding dose-response curve for compound 10, used for figure 6.4, can be seen as figure 6.5. Dose-response curves were created in this manner for all SPR experiments with all compounds, and can also be found in appendix C. As can be clearly observed in figure 6.5, there is some considerable variation between maximal RU values for the same concentration (ie. 200 μM), and overall there is no observable concentration-dependent relationship.

Though not conclusive, some dose-response curves show promising trends. For compound 85 there is a notable upward trend, and for compound 25 a notable negative trend (figure 6.6).

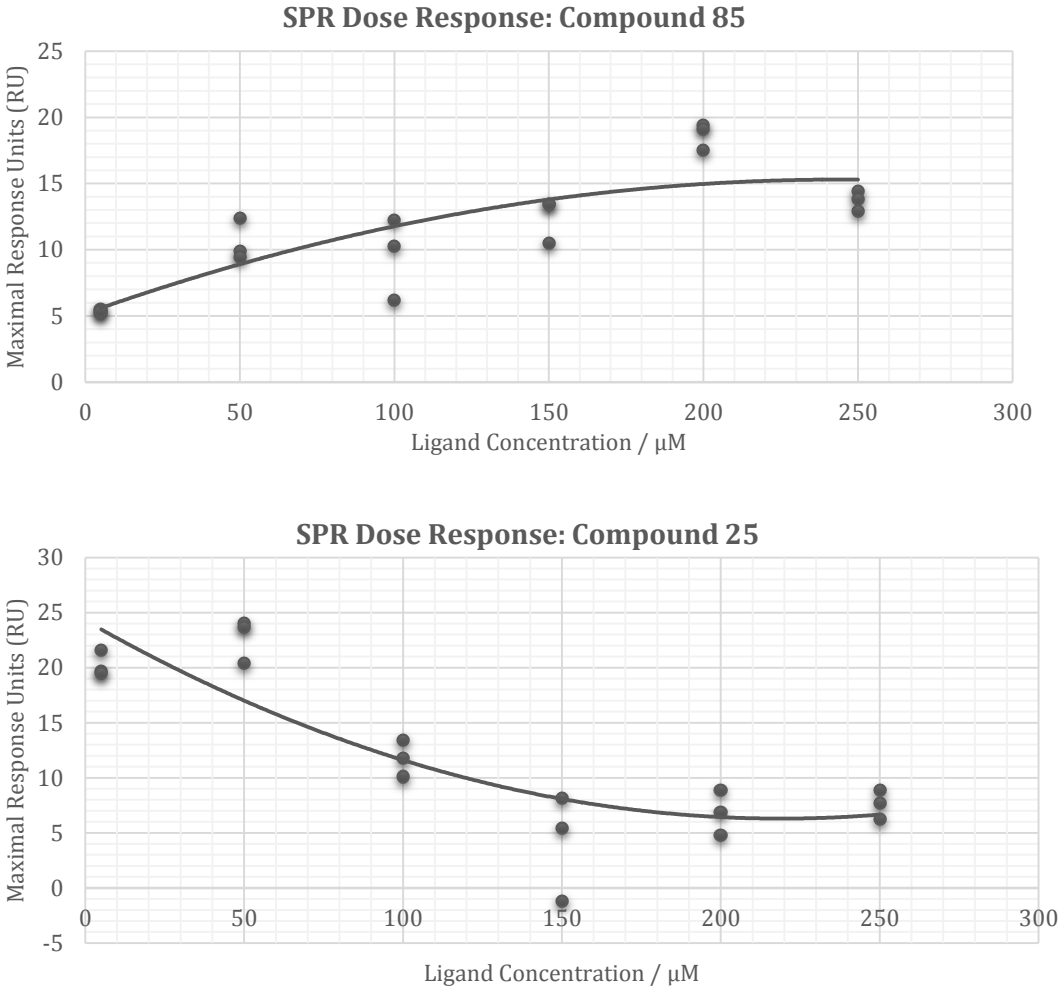


Figure 6.6: Dose-response curves for compounds 85 and 25.

The slight increase in the dose-reponse relationship for compound 85 indicates binding, and potentially covers the beginnings of saturation where the curve-fit approaches a horizontal; a fuller concentration range will determine if this is the case. Compound 25 shows a very different dose-response relationship, as the response units – directly correlated to mass on the surface of the chip – *decreases* with increasing concentration.

This profile of curve indicates compound 25 binds at low concentrations, but at concentrations exceeding 50 μM, this compound begins to dissociate from the protein, or causes some other loss of mass from the chip surface. Again, a fuller dose-response relationship will be elucidated from a further screen of wider concentrations.

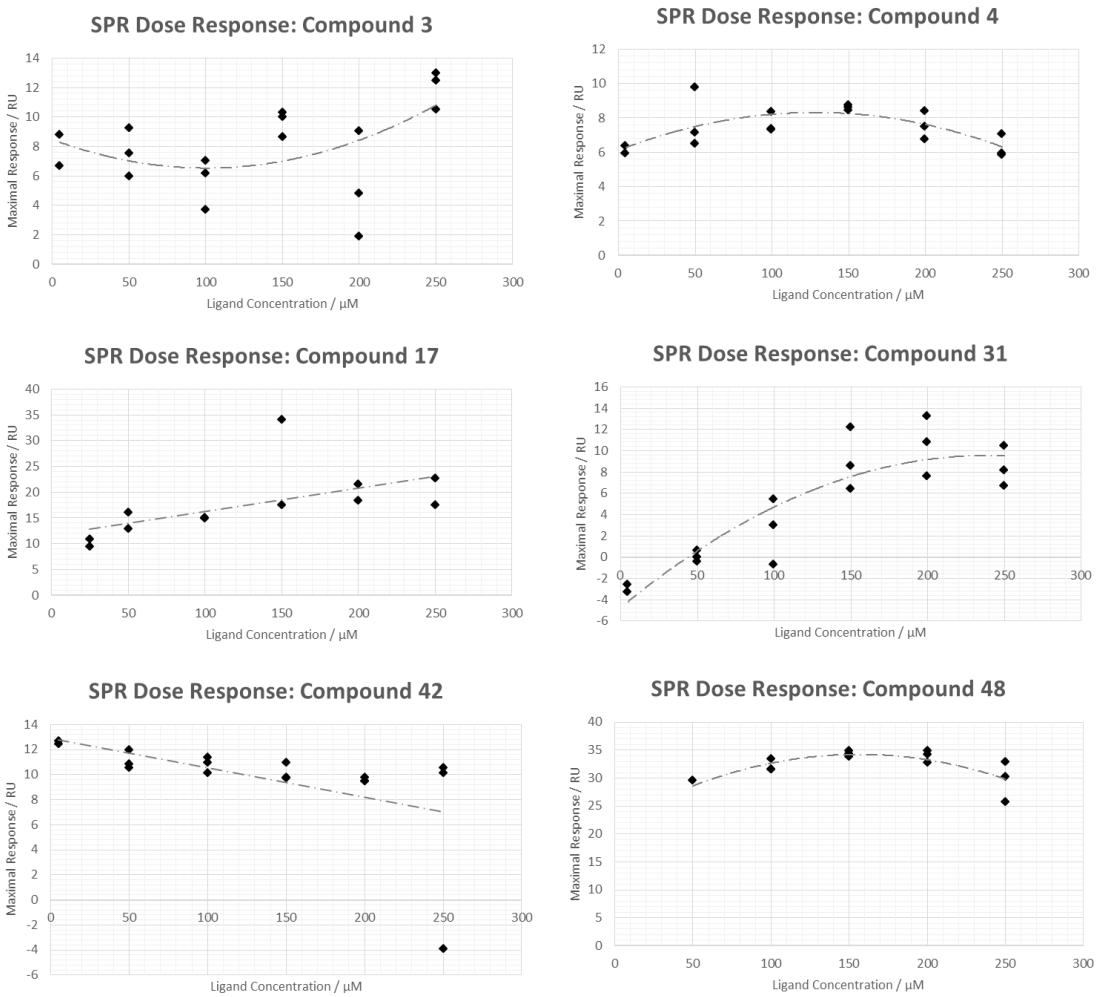


Figure 6.7: Dose-response curves for compounds 3, 4 17, 31, 42 and 48.

Of the dose-response curves shown above in figure 6.7, several compounds show little variation (ie. compound 48, figure 6.7, bottom right). At a maximal response of 30-35 RU, this indicates complete binding but is insufficient to derive a binding affinity. Contrastingly, compound 31 shows an upward trend in dose-response.

The dose-response curve for compound 50 (figure 6.8, top left) shows no binding whatsoever under the experimental conditions. This implies either an inability to bind EthR or a very weak affinity in excess of 250 μM. The dose-response curve for compound 57 (figure 6.8, top right) shows little variation in binding at the concentrations used, which could indicate saturation has been reached. Data for compound 80 (figure 6.8, bottom right) is especially noisy, with a wide range within each concentration.

Ultimately, the curves obtained from these preliminary studies can only be used for tentative conclusions about binding capability, and not for binding affinity.

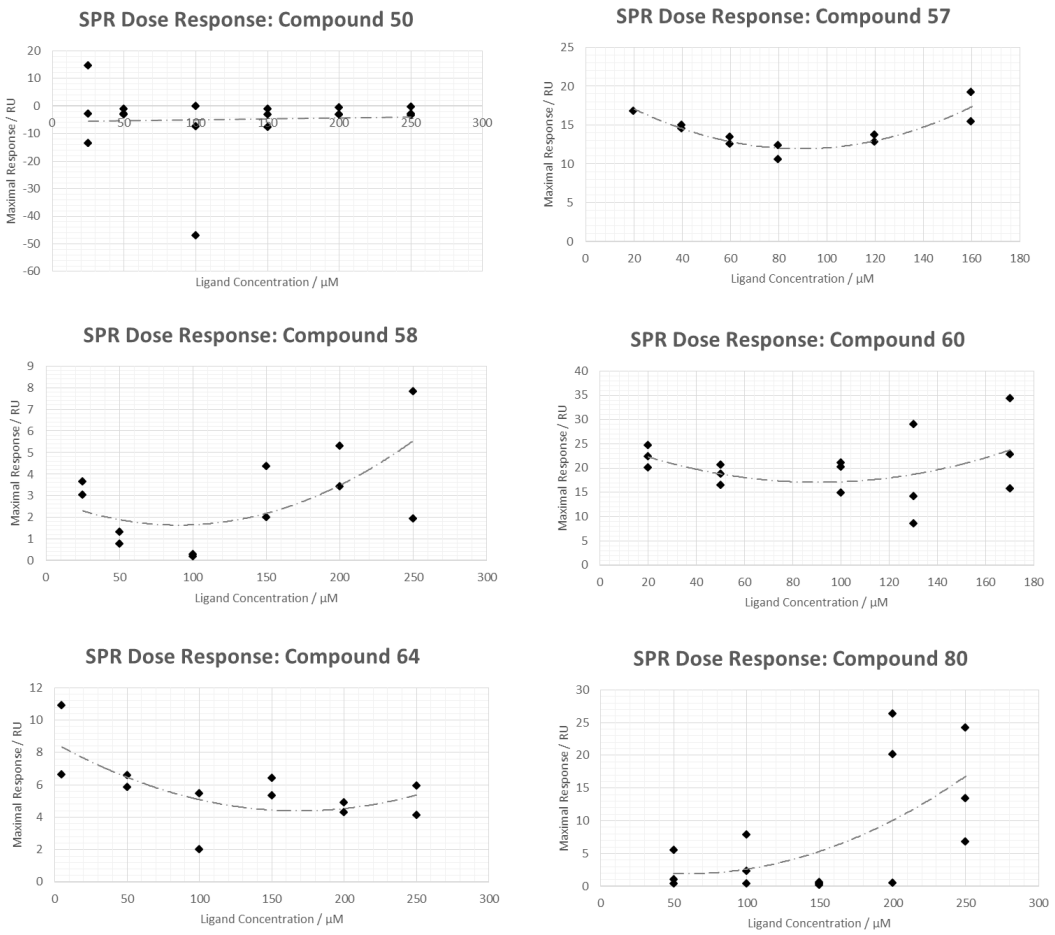


Figure 6.8: Dose-response curves for compounds 50, 57, 58, 60, 64, and 80.

In order to determine if these curves, particularly those which are more linear, show pre-binding or saturation of the protein, these experiments will have to be repeated with a wider range of concentrations. This range must include sub-micromolar concentrations as well as those in excess of 250 µM in order to accurately determine binding affinity and acquire publication quality data.

The BIAEvaluation software (v.3.1) was able to fit a 1:1 Languir binding model with baseline drift to all SPR binding data obtained. Despite presenting low Chi-squared values and a generally good appearance of fit, it is clear from the dose-response curves that the calculated binding affinities are estimates at best. They are presented here in table 6.1 as binding affinity estimates, but should be treated with due caution. Separate fitting of k_a and k_d was performed in order to obtain estimates of these values for global fitting; subsequently global fitting was performed to the 1:1 Langmuir binding model with baseline drift.

Table 6.1: Summarised binding affinity data for each compound. Chi-squared values are all within ideal values – 5% of a 25 Rmax corresponds to a χ2 of 1.25. Compounds 15, 60, 72, 74 and 84 were untested due to low compound stock. Compound 50 showed no binding and so did not undergo fitting.

Compound	K_D / μM	χ2
03	106.0	0.725
04	465.0	0.270
10	71.7	0.708
17	34.0	0.962
25	14.4	0.303
31	430.0	0.374
42	284.0	0.099
48	73.8	0.858
56	12.9	0.588
57	5.0	0.412
58	465.0	0.270
64	3.8	0.356
80	178.0	1.14
85	20.9	0.475

6.4 Summary

Pilot experiments undertaken with the remainder of each compound stock after crystallisation trials indicate many compounds capable of binding EthR, some with potentially low-micromolar affinity. For full binding affinity characterisation to occur, a wider range of concentrations will be required.

Only one compound from this cohort has been observed not to bind EthR at all, namely compound 50. Of the 14 compounds tested by SPR, this gives a hit rate of 93%. Once binding affinities have been accurately determined, the kinetic data can be used in conjunction with structural data (presented in chapter seven) to identify good hits to use as optimisation starting points.

CHAPTER VII:
CO-CRYSTAL STRUCTURES OF EthR AND POTENTIAL INHIBITORS

In conjunction with biophysical studies previously described, co-crystallisation experiments were undertaken to determine crystal structures of EthR bound to identified potential hit compounds. This chapter details the methods and results of crystallisation and co-crystallisation endeavours; as well as analysis of the resulting crystal structures with notable comparisons to the *in silico* predicted binding modes.

7.1 Introduction

Well-established, known conditions (referenced below for table 7.1) were used to first crystallise EthR in the absence of added compound. Once this could be achieved, co-crystallisation would be undertaken under the same conditions using published co-crystallisation methods for EthR to acquire liganded structures.[1]

7.2 Native EthR Crystallisation Methods

EthR, freshly purified and dialysed into a buffer containing 10 mM TRIS HCl pH 7.5 and 200 mM NaCl, was concentrated to 9 mg ml^{-1}. A 24-condition crystallisation screen was prepared covering previously published crystallisation conditions,[111,119,124,127–129] detailed in table 7.1. All solutions were prepared and sterile filtered in 10 mL stocks with ammonium sulphate (VWR), MES (2-(N-morpholino)ethanesulfonic acid, Melford Laboratories Ltd.) and glycerol (Melford Laboratories Ltd).

7.3 Native EthR

Six hanging-drop crystallisation trays of EthR at 9 mg ml^{-1} (drops consisting of 1 µL protein solution : 1 µL reservoir solution) were set up according to the conditions in table 7.1. Protein crystals grew at 20°C within two days, in 41-85% of conditions; crystallisation was least successful at the highest ammonium sulphate concentration, and at the highest concentration and pH of MES (ie. row D, 0.15M MES pH 6.5).

Table 7.1: Crystallisation screen for EthR, utilised for native and ligand co-crystallisation.

	1	2	3	4	5	6
A	1.40 M $(NH_4)_2SO_4$; 0.1M MES pH 6.0; 10% glycerol	1.45 M $(NH_4)_2SO_4$; 0.1M MES pH 6.0; 10% glycerol	1.50 M $(NH_4)_2SO_4$; 0.1M MES pH 6.0; 10% glycerol	1.55 M $(NH_4)_2SO_4$; 0.1M MES pH 6.0; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.1M MES pH 6.0; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.1M MES pH 6.0; 10% glycerol
B	1.40 M $(NH_4)_2SO_4$; 0.1M MES pH 6.5; 10% glycerol	1.45 M $(NH_4)_2SO_4$; 0.1M MES pH 6.5; 10% glycerol	1.50 M $(NH_4)_2SO_4$; 0.1M MES pH 6.5; 10% glycerol	1.55 M $(NH_4)_2SO_4$; 0.1M MES pH 6.5; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.1M MES pH 6.5; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.1M MES pH 6.5; 10% glycerol
C	1.40 M $(NH_4)_2SO_4$; 0.15M MES pH 6.0; 10% glycerol	1.45 M $(NH_4)_2SO_4$; 0.15M MES pH 6.0; 10% glycerol	1.50 M $(NH_4)_2SO_4$; 0.15M MES pH 6.0; 10% glycerol	1.55 M $(NH_4)_2SO_4$; 0.15M MES pH 6.0; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.15M MES pH 6.0; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.15M MES pH 6.0; 10% glycerol
D	1.40 M $(NH_4)_2SO_4$; 0.15M MES pH 6.5; 10% glycerol	1.45 M $(NH_4)_2SO_4$; 0.15M MES pH 6.5; 10% glycerol	1.50 M $(NH_4)_2SO_4$; 0.15M MES pH 6.5; 10% glycerol	1.55 M $(NH_4)_2SO_4$; 0.15M MES pH 6.5; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.15M MES pH 6.5; 10% glycerol	1.60 M $(NH_4)_2SO_4$; 0.15M MES pH 6.5; 10% glycerol

Notably, crystallisation was most successful with freshly purified protein (ie. ≤ 2 days since size exclusion chromatography). Figure 8.1 features examples of the crystals obtained, all either needle or column-shaped in morphology.

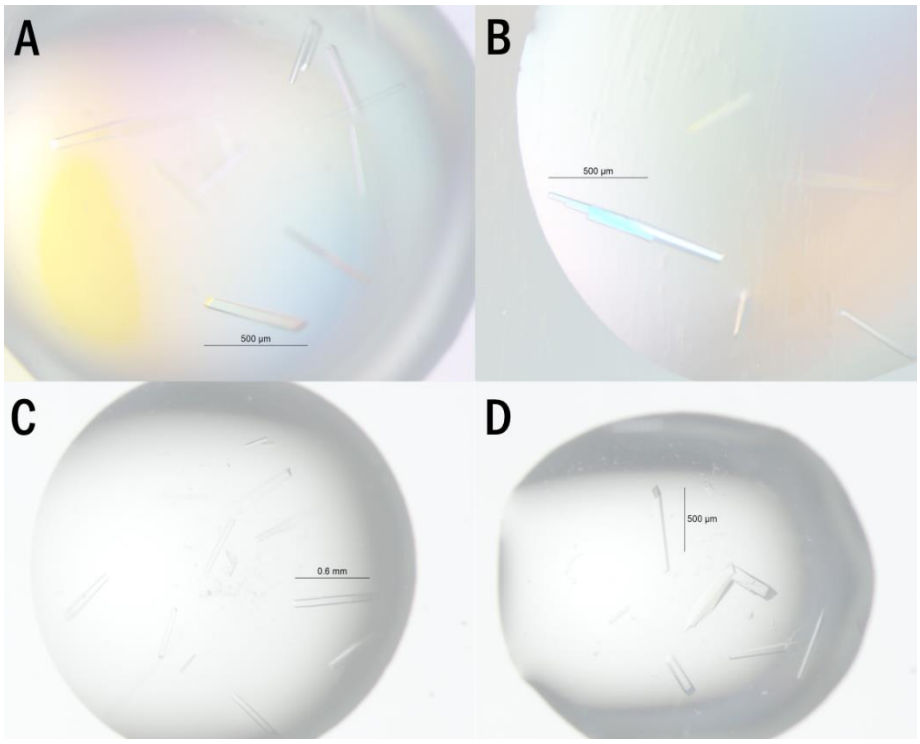


Figure 7.1: Crystal growth from conditions in row A, B, C and D. Scale bars A, B, D at 500 µM; C at 600 µM.

Crystals were mounted on an in-house diffractometer (Bruker Microstar with Cu-Kα rotating anode) to ascertain their identity as the desired protein. No ice rings or ice formation was observed, and the crystals were well frozen, indicating 10% glycerol sufficient as cryo-protectant as part of the crystallisation cocktail. Diffraction was observed to approximately 3.5 Å and was sufficient to determine the unit cell as being tetragonal P, with $a = b = 122$ Å, and a c length of 33 Å. These matched unit cells of published EthR crystal structures.

7.4 Co-Crystallisation Methods

Using the crystallisation screen detailed in table 8.1, crystal trays of EthR with all twenty hit compounds were established. A solution of 9 mg ml^{-1} EthR was allowed to equilibrate with a compound at room temperature for approximately 24 hours prior to setting up crystallisation experiments, by mixing 9 µL protein with 1 µL of compound (at a concentration of 33 mM in 100% DMSO); this gave a final compound concentration of 3.3 mM, and a final DMSO content of 10%. Initially, a half-hour incubation period was used for protein-ligand equilibration, however trays established with the remaining protein-ligand cocktail, which had been left at room temperature overnight, proved far more successful.

Of the successful trays, crystals were most abundant in trays of compound 85 - somewhat unexpected given the thermal shift assay had demonstrated this compound caused a negative effect on the melting temperature of the protein, indicative of a interaction which destabilised protein. A crystal from a tray of EthR co-crystallised with compound 85 was mounted in-house to ensure the unit cell resembled that of EthR. Diffraction was observed to around 3.5 Å, as with the native EthR crystals. The unit cell obtained was slightly different from those published, in that the crystal system best fitted by the matrix was found to be Orthogonal P, with cell lengths of 33 Å, 118 Å and 120 Å for a, b and c respectively. However, these match closely enough to previously obtained unit cells for EthR that there was cause for reasonable confidence these crystals were at least of the protein. For high-resolution data, crystals of EthR (native and co-crystallised with compounds) were flash-cooled[162] in liquid nitrogen and transported to Diamond Light Source for data collection.

7.5 Synchrotron Data Collection and Assessment of Ligand Density

Crystals obtained from co-crystal trays for compounds 25, 57, 60, 80, and 85, and a further set of crystals representing compounds 3, 10, 15, 42, 50, and 74 were tested on beamlines I04-1 and I03 respectively at Diamond Light Source by remote access. For

each of the crystals, four images were taken (Δ = 45°) at 20% transmission of beam at wavelength (λ) 0.91741 Å, with 0.2s exposure time, and a resolution limit of 2.3 Å (detector distance = 310 mm). The aperture of the beam used was 70 μm by default, or reduced to 30 μm if deemed necessary for smaller crystal sizes. Strategy calculations are carried out automatically by EDNA Mxv1[163] and Mosflm.[164] Typically, the starting position suggested by Mosflm was utilised. The *xia2* pipeline was used to initially determine data quality and (by rigid body refinement) the presence of ligand prior to manual processing.

Of the first cohort of potential co-crystals, ligand density was observed for data corresponding to experiments with 25, 57, 60, 80 and 85; for cohort two, ligand density was observed for data corresponding to experiments with 3, 10, 15, 42 and 74. Therefore, of the crystals tested, only crystals from conditions including compound 50 failed to yield data with ligand density present; compound 50 was previously shown not to bind EthR in pilot SPR experiments (chapter six).

Upon confirmation of the presence of ligand density, the data were carefully reprocessed in XDS.[165] Initially, all collected data were indexed and scaled, and the output examined to determine an appropriate place to cut low-quality, highest resolution data from the data set. This resolution limit was then used to re-index and re-scale the data. To convert the output of HKL co-ordinates and intensities from XDS into structure factors in the MTZ format for electron density refinement, Aimless[166] was used as part of the CCP4i (v6.5) package.[167] FreeR flags were generated for 5% of the data.

Each set of data per successful co-crystallisation experiment will now be discussed in further detail. In all but two cases, rigid body refinement in Refmac5[168] was used to place the monomeric protein model with water and ligands removed. In the cases of EthR with compounds 3 and 10, this procedure failed and was followed by molecular replacement in Phaser.[169,170] Refinement was carried by Refmac5 with additional modelling of side-chains, waters and ligands in Coot[171]. Ligand dictionaries were generated using PRODRG[172] based on co-ordinates from virtual screening poses.

7.5.1 Compound 3

Data for a crystal of EthR co-crystallised with compound 3 were obtained to a resolution of 1.8 Å. Full scaling/integration statistics are given for each structure in Appendix D.

Attempts to solve the structure of EthR with compound 3 via simple rigid body refinement were unsuccessful, resulting in a poor model fit and R_{work}/R_{free} factors of 45% and 48%, respectively. Molecular replacement in Phaser was employed to produce a better initial solution and starting protein model for subsequent refinement. As for the

rigid body refinement, the 3TP0 structure of EthR was utilised. Unit cell contents analysis by Phaser determined 52.6% solvent content based on a protein molecular weight of 24 kDa and 1 molecule in the asymmetric unit. The output solution was given in space group $P4_12_12$ (identical to the published space group), with an RFZ of 4.0, a TFX of 8.7 and an LLG of 106. Rigid body refinement was then repeated with the resulting model with initial R_{work}/R_{free} factors of 31.2/30.9%. Subsequent restrained refinements were performed with water molecules and side-chain rotamers fitted, at which point (R_{work}/R_{free}: 18.4%/22.7%) the ligand was fitted to the unbiased density (figure 7.2).

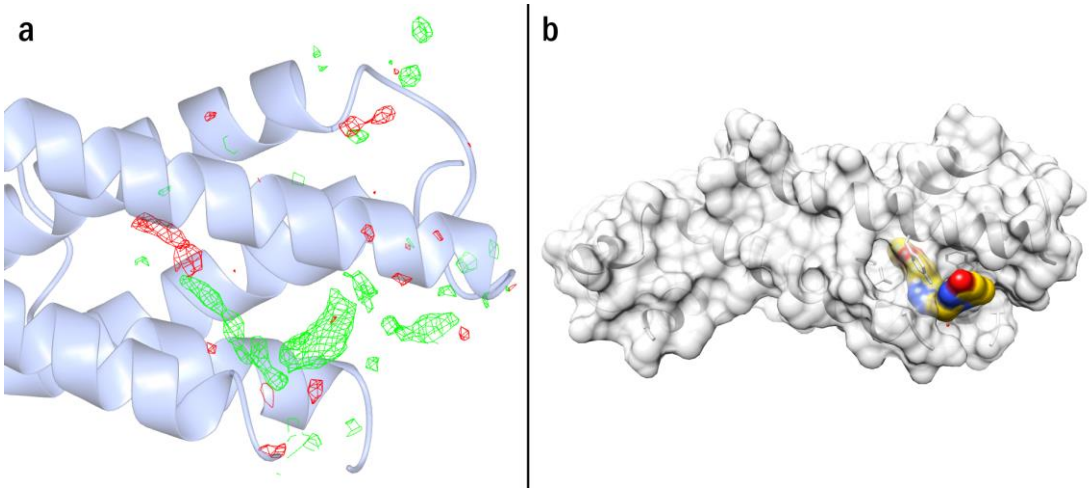


Figure 7.2: (a) Unbiased Fo-Fc difference density in ligand binding region to which compound 3 was fitted, contoured at 3 σ. Image made in CCP4MG.[173] **(b)** Space fill (molecular surface) model of EthR (grey), compound 3 in yellow. Image made in Chimera.[174]

Unlike previously observed ligands, compound 3 occupies a position much lower in the binding pocket and is, ultimately, solvent exposed at the methyl-pyrimidinone group. Consequently, many commonly observed interactions between EthR and ligand binding partners are not observed here (ie. asparagine hydrogen bonds, phenylalanine stacking).

Final R factors for restrained refinement are 18.3/22.6% respectively for R_{work} and R_{free}, with an RMS values for bond length and angle of 0.02 Å and 1.81° respectively. Figure 7.3 shows compound 3 with surrounding residues. Two hydrogen bonds are made, though that formed between tyrosine residue 148 and the pyrimidinone nitrogen is of better distance and geometry than that formed between Thr149 and the dioxane oxygen. No stacking interactions between the various ring systems of the ligand and the protein can be observed, with the ligand out of range of Phe110, Trp207 and Trp103.

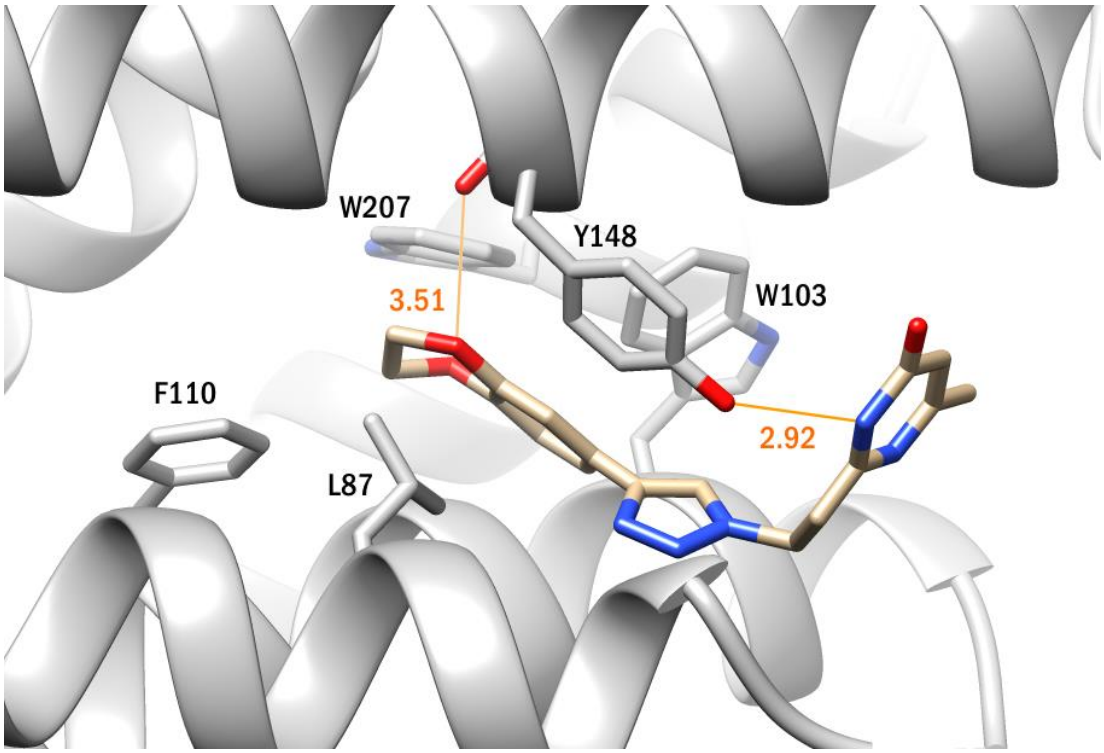


Figure 7.3: Compound 3 bound to EthR. Hydrogen bonds labelled with key residues shown.

Virtual screening had suggested two potential binding modes for compound 3. Figure 7.4 shows the five poses generated. The actual positions differ greatly, thus emphasising the necessity of experimental data over only modelled ligand poses.

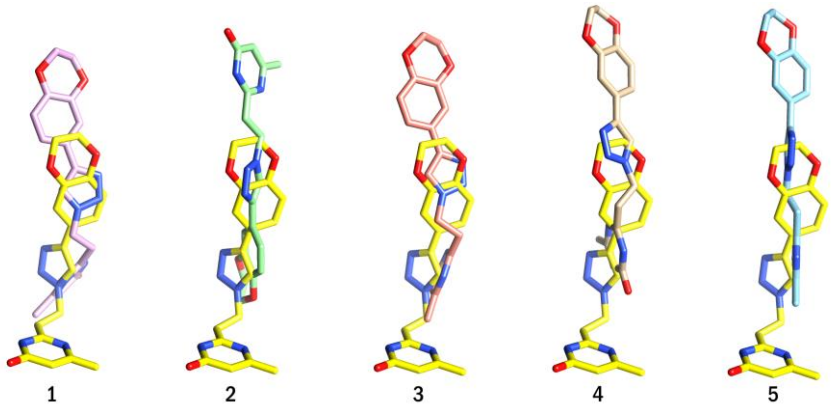


Figure 7.4: The experimentally derived ligand conformation (yellow) compared to the ranked poses from virtual screening. Four out of the five poses are oriented with the phenyldioxane group at the 'top'; pose two exemplifies the 'alternate' conformation caused by a 180° flip in orientation.

Despite the unconventional binding mode, the ligand is successfully modelled at full occupancy, and makes observable hydrogen bonds. However, estimated binding affinities from SPR suggest this compound binds weakly, which could be attributed to

this unexpected, highly solvent exposed binding mode. Overall, the co-crystal structure of compound 3 with EthR shows that the binding ability was predicted whereas the binding mode itself was not. This demonstrates the vitality for experimental validation of virtually modelled ligand poses.

7.5.2 Compound 10

Diffraction data for co-crystallisation of EthR and compound 10 were obtained and after ascertaining ligand presence, the data were re-integrated and re-scaled to a resolution of 1.5 Å. As with compound 3, initial attempts to provide a protein model for phasing via rigid body refinement failed, with initial R-factors exceeding 50% and poor model-density fit. Molecular replacement was used instead, and Phaser[169,170] produced a model in the expected space group of $P4_12_12$, with one molecule in the asymmetric unit and a predicted solvent content of 53.5%. After rigid body refinement and restrained refinements, ligand density in the binding site was sufficient to begin modelling.

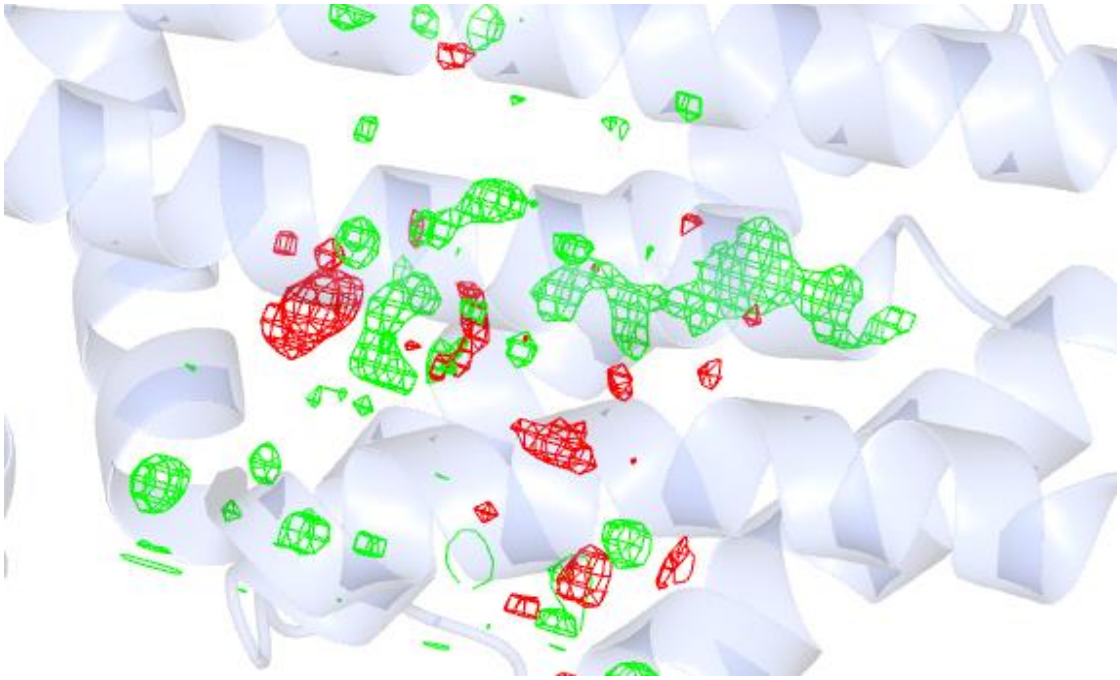


Figure 7.5: Fo-Fc density map shown at 3 σ around the binding site.

The ligand was placed but unsatisfied Fo-Fc difference density in and around occupied positions suggested insufficient occupancy even at 100% for this ligand orientation (figure 7.6).

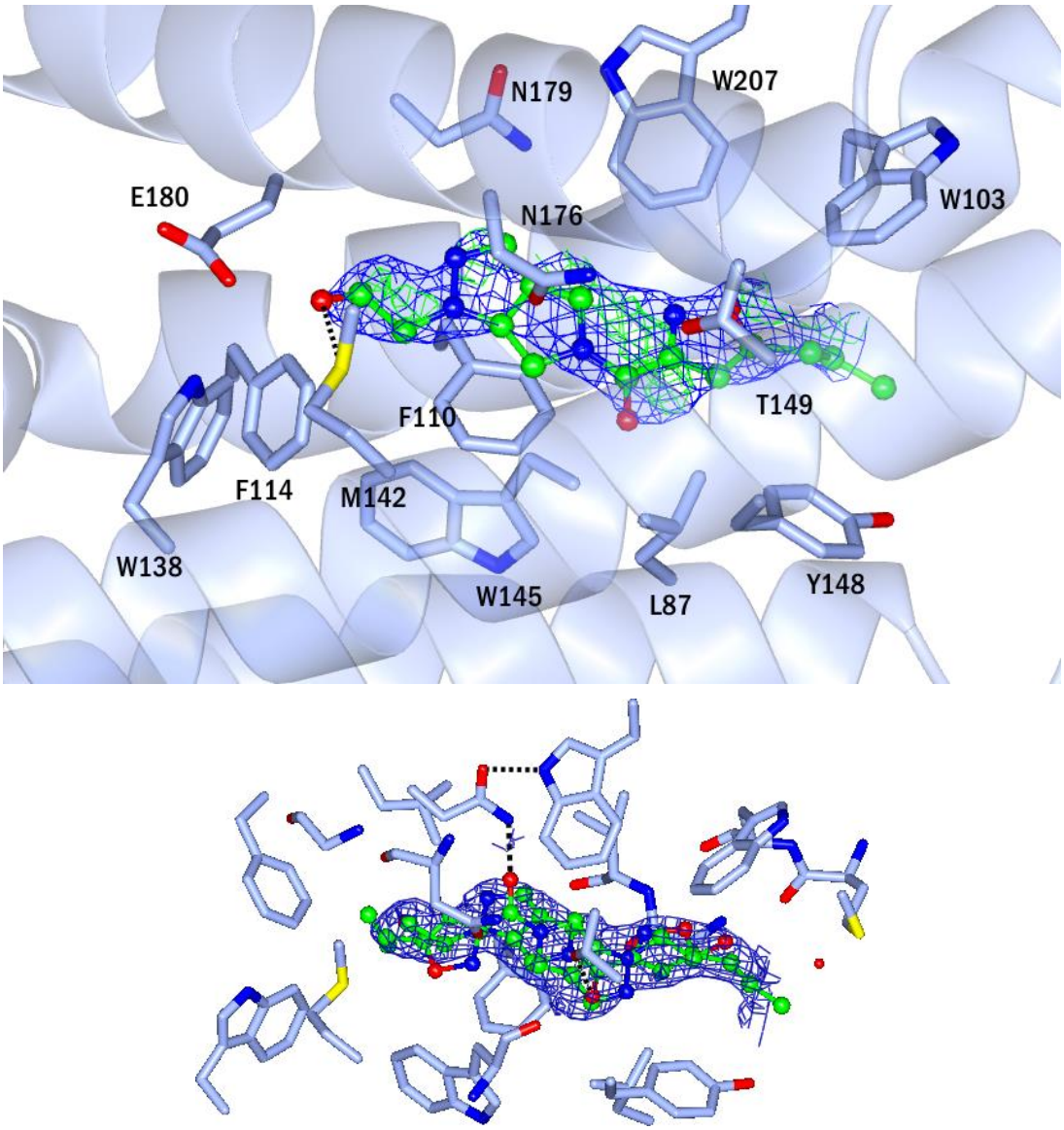


Figure 7.6: Positive difference density (Fo-Fc) observed at 3 σ, with 2Fo-Fc (blue) contoured at 1 σ. Top, with ligand occupancy at 100%, this suggested the presence of a second molecule, each at 50% occupancy. Bottom, both ligands modelled with same density parameters.

The only satisfactory modelling, therefore, was with both potential ligand orientations, A and B, each with 50% occupancy. This resolved the difference density observed and gave improved refinement statistics (table 7.2).

Table 7.2: Refinement statistics for EthR with compound 10, before ligand placement, after single molecule placement, and the final refinement statistics with both orientations modelled.

	Prior to placement	Single conformation	Final Structure
R_{work}/R_{free} (%)	22.8 / 24.8	19.2 / 22.3	19.1 / 22.0

Only one of the predicted poses from GOLD is in a similar conformation as A (figure 7.7, yellow), with the alcohol tail toward the aperture of the channel (figure 8.7, blue). This pose was ranked second; all other poses share the same conformation as B.

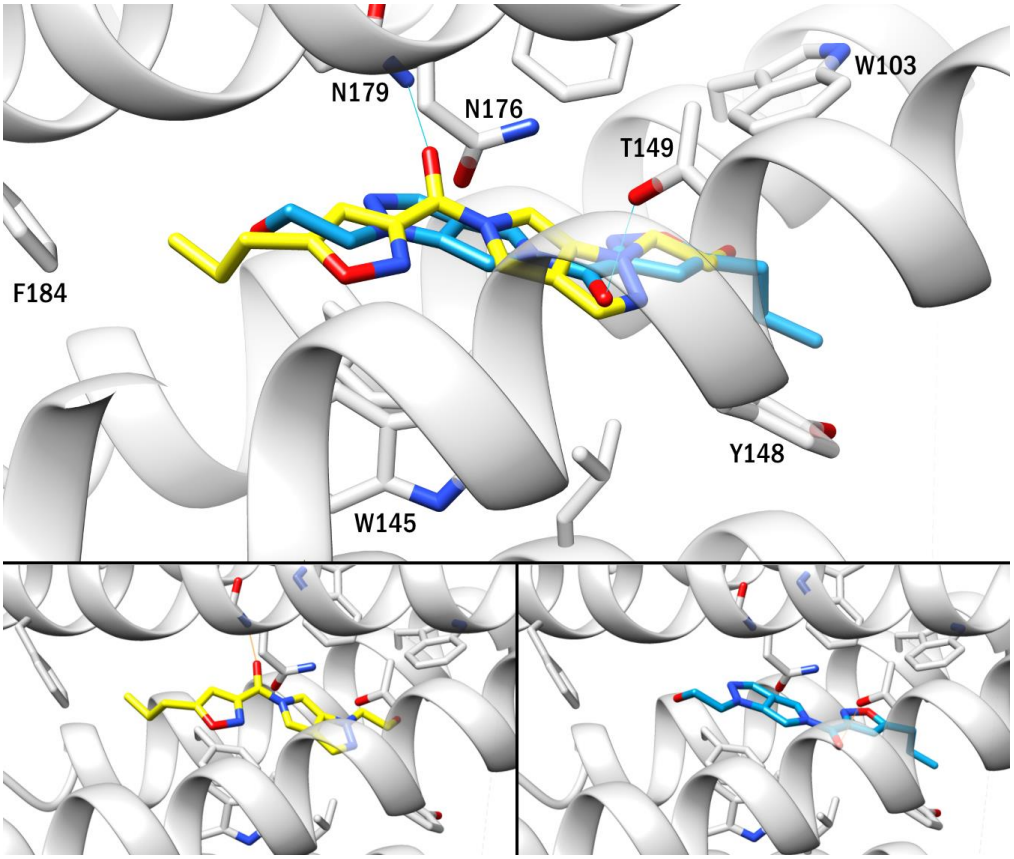


Figure 7.7: Labelled image of compound 10 bound to EthR in two orientations. Top panel shows both orientations, bottom left shows orientation A, bottom right shows orientation B.

Conformation A, shown in figure 7.7 in the bottom left panel, forms the frequently-observed hydrogen bound with Asn179 via the carbonyl at a short distance of 2.3 Å. Conformation B, shown in figure 7.7 in the bottom right panel, forms a hydrogen bond through the same carbonyl with Thr149, at a distance of 3.1 Å. No other atom in the ligand, in either conformation, appears to be involved in hydrogen bonding despite the presence of donors and acceptors.

Ultimately, this is the first crystallographic evidence for EthR supporting multiple binding conformations of the same ligand, though previous docking studies[122] and the virtual screening results had suggested it to be possible. Such modelling was made possible by very high resolution data. Further computational exploration, beyond the scope of this thesis, is currently underway to probe this interesting feature of the EthR binding site and determine if multiple binding modes can be predicted through linear interaction energy calculations.[175,176]

7.5.3 Compound 15

Initial experiments to crystallise EthR with compound 15 were unsuccessful due to the formation of a coloured film over the drops, presumably a result of precipitation of the hydrophobic ligand. To migate this effect, the concentration of compound was reduced from a 10:1 molar ratio to a 4:1 molar ratio of compound to protein, which yielded crystals of the same morphology as other crystallisation experiments with EthR.

Upon initial rigid body refinement prior to manual data processing, difference density was observed indicating a ligand presence. However, after manual data processing (to 1.9 Å), rigid body refinement and subsequent restrained refinement, it became clear the density did not appear to match the expected ligand and was insufficient to confidently model a bound molecule (figure 7.8).

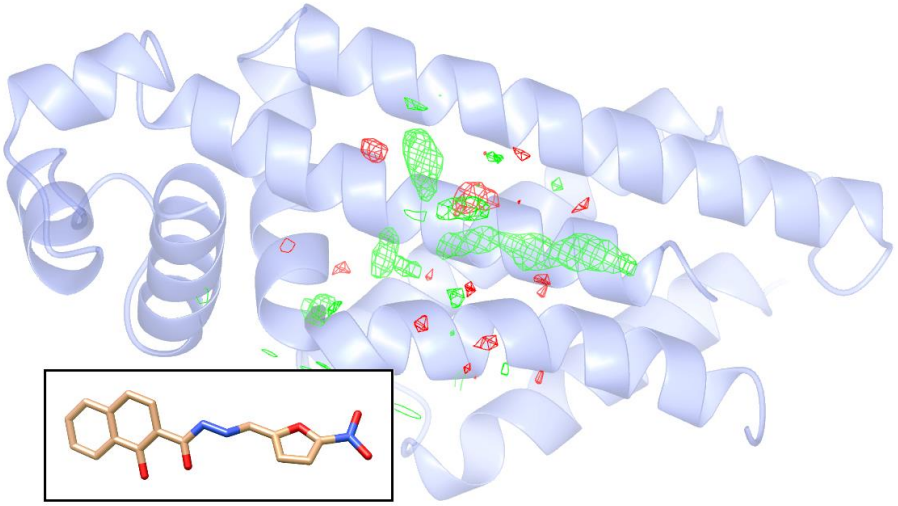


Figure 7.8: Compound 15, inset, did not appear to match the ligand density (Fo-Fc) and could not be modelled in good confidence. Difference density is shown at 3 σ.

The structure was refined to final R_{work}/R_{free} values of 19.3/21.3%, but the unbiased ligand density did not improve enough to allow sufficiently confident modelling.

Compound 15 had showed a moderate (3°C @ 400 μM) positive shift in thermal stability in the thermal shift assays (chapter five), and crystal structures have been obtained with ligands which caused a weaker stabilising shift (ie. compound 85, compound 42). Therefore we can speculate that the hydrophobicity of this ligand (highlighted by the poor solubility in the first co-crystallisation trial) is a barrier to good occupancy despite the lipophilicity of the EthR binding site. Alternatively, compound 15 possesses a furan ring, a moiety which has been shown to be prone to degradation in DMSO stocks.[177] It could be that such an effect has occurred here, with multiple divergent

species formed; averaging in the crystal structure then causes a loss of definition between these molecules.

Ultimately, with other structures to focus efforts on and no confident way forward for ligand modelling, the crystal structure for EthR with this ligand was not pursued any further.

7.5.4 Compound 25

Upon crystallisation of compound 25 it was revealed this compound did not match the compound expected; upon investigation of compound stocks this was discovered to be because the compound delivered was not, in fact, the one ordered. Compound 25 was therefore fortuitously crystallised with EthR, but cannot be compared to a docked ligand from the virtual screening campaign.

To obtain the ligand structure for modelling, it was possible to trace the compound code back to a ZINC record. This compound would have been present in the original 6.1 million compound cohort obtained from ZINC but the pre-screen filtering would have excluded it from the screening set based on size alone (193 g mol^{-1}).

The crystal structure has the distinction of being the most high resolution structure of EthR to date, at 1.4 Å (R_{work}/R_{free} of 19.9/22.1%; full crystallographic refinement statistics given in appendix D). Figure 7.9 shows compound 25 and its position in the EthR binding site. Similar to published ligands crystallised with EthR, compound 25 consists of a piperidine ring with an amide group.

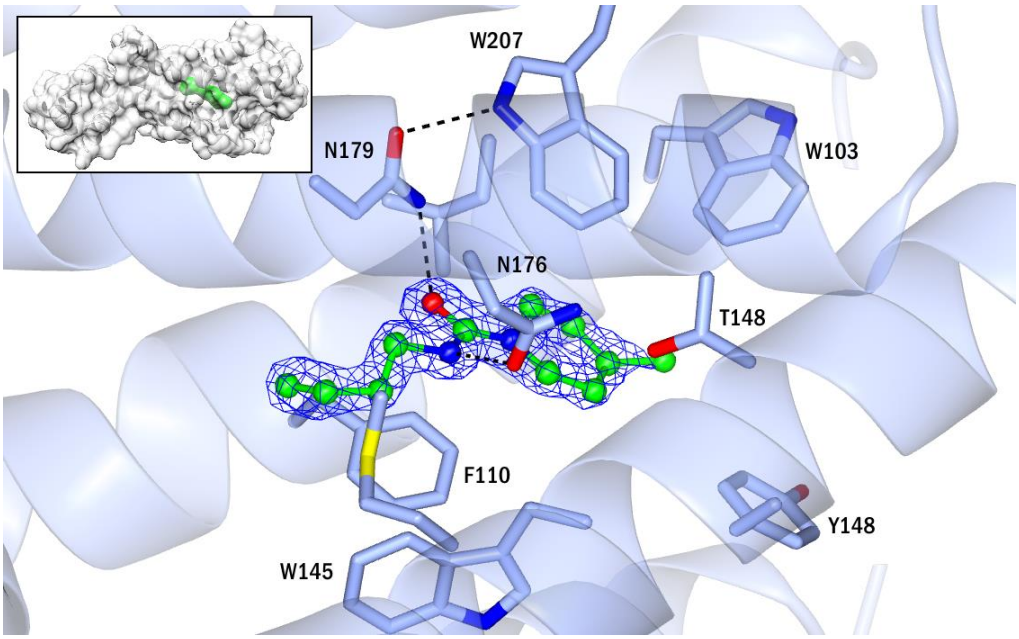


Figure 7.9: Compound 25 crystallised bound to EthR, with ligand density (2Fo-Fc) shown at 2 σ. Inset, surface model of EthR and compound 25 with ligand in green to highlight position.

Despite being the smallest of the small molecules crystallised with EthR, only one molecule binds in the long, EthR binding channel. The ligand itself forms two hydrogen bonds: one to Asn179 at a distance of 2.8 Å via the carbonyl moiety, and a second to Asn176 via the nitrogen of the amide function at a distance of 2.9 Å.

This ligand would make an excellent starting point for a fragment-based approach to develop new scaffolds, as an alternative to the small-molecule methods of discovery described in this thesis. This ligand beautifully exemplifies the common binding mode of many EthR ligands, those previously published and those presented here: binding is anchored by hydrogen bonds to the key asparagine residues in the central region, and supported by non-polar interactions along the rest of the interface.

7.5.5 Compound 42

Diffraction data for compound 42 were obtained to 1.95 Å and the structure was refined with one molecule bound to the protein. It was not immediately clear which ligand orientation best fit the experimental observations. As such both were initially modelled. However, the orientation shown in figure 7.10 was the best interpretation of the observable density.

The structure shows compound 42 binds with the propyl ring toward the deepest part of the pocket and the amide function toward the aperture. This is contrary to what would have been expected, which is that the carbonyl ring would form a hydrogen bond to the key asparagine residue 179; all attempts to model such an interaction resulted in poor density fit and unsatisfactory convergence. This compound makes no observable hydrogen bonds with the protein, which would begin to explain the poor estimated binding affinity observed by SPR (284 μM).

The ligand position results in phenylalanine F184 adopting two conformations, likely to prevent unfavourable clashes in solution. Of the ligands presented here, only compound 85 is capable of occupying the site such that the phenylalanine adopts the open conformation; all other ligands are shorter or adopt lower positions and the phenylalanine remains in its ‘closed’ orientation.

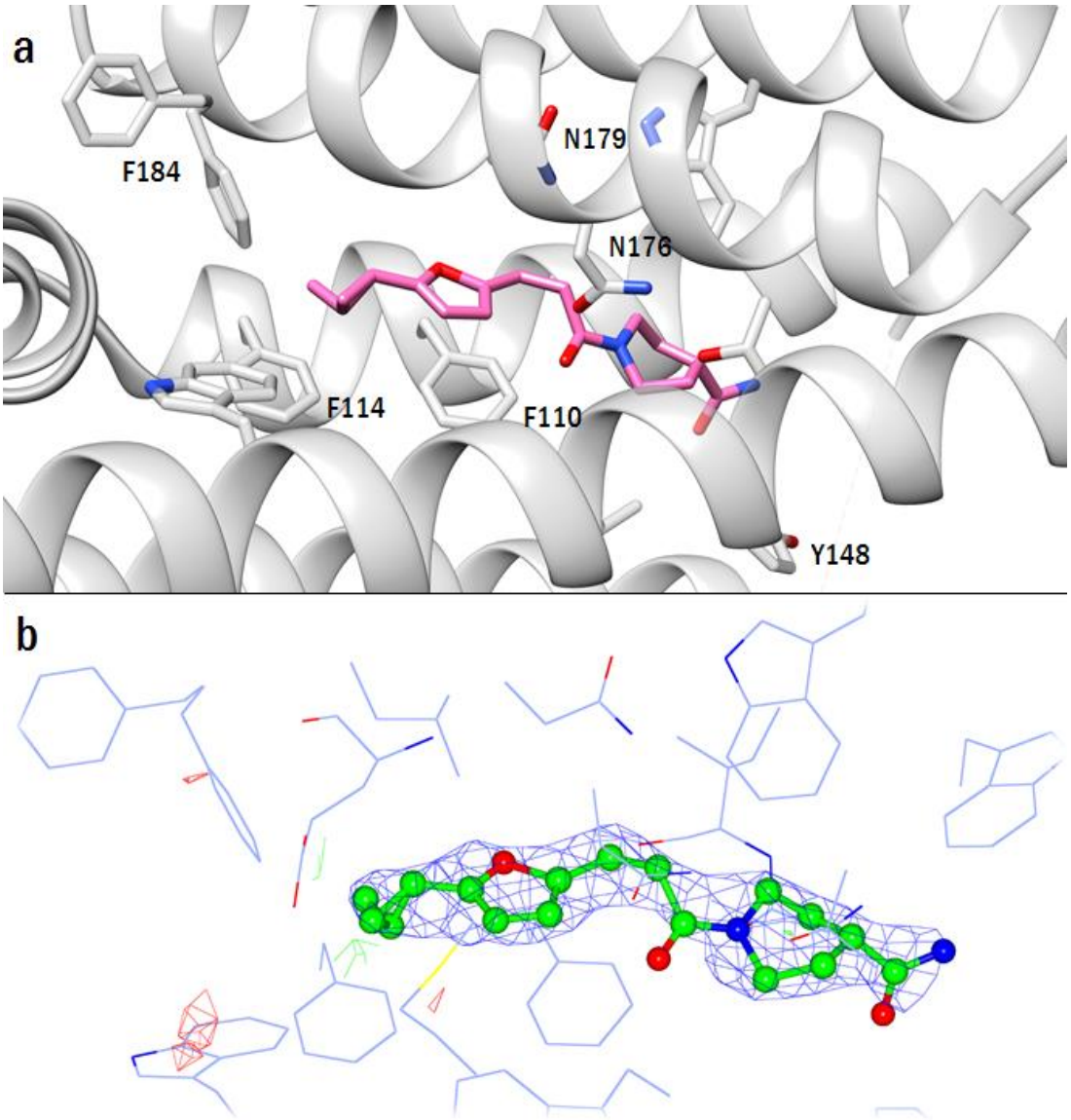
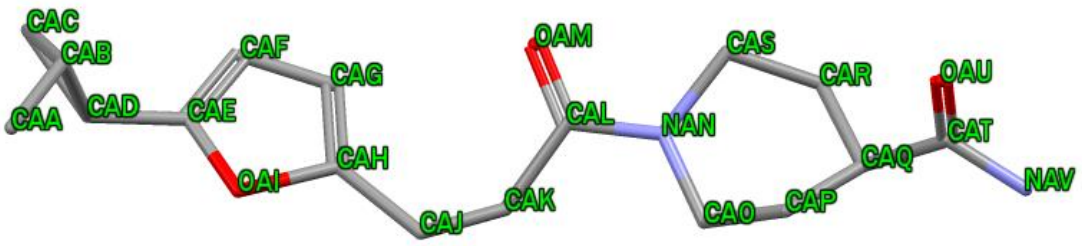


Figure 7.10: Top, compound 42 in the EthR binding site. This ligand makes no observable hydrogen bonds with the protein. Bottom, 2Fo-Fc density contour of compound 42 at 1 σ, with Fo-Fc difference density contoured at 3 σ.

To ensure correct ligand modelling, the refined ligand was assessed by Mogul for torsional strain – particularly around the piperidine which seemed, visually, to adopt an unusual conformation. However, the only torsions which Mogul considered to be unusual were those associated with the propyl ring (table 7.3).

Table 7.3: Mogul torsion check results for compound 42 (labelled by atom designation). Number reflects the number of records found in the database; all other torsions were excluded for poor representation. Minimum and maximum torsion values in the database distribution, the distance to the nearest neighbour (in degrees) and the original query value of the search torsion.



FRAGMENT	Number	Min. / °	Max. / °	| d(min) | / °	Query value / °
CAB CAD CAE CAF	6	6.918	37.046	48.103	-85.149
OAI CAE CAD CAC	5	18.849	132.454	34.141	166.595
OAI CAE CAD CAB	12	9.817	164.882	20.135	94.144
CAC CAD CAE CAF	12	7.071	145.412	1.793	-12.698
OAU CAT CAQ CAP	92	6.098	142.382	0.749	-64.914
CAK CAL NAN CAO	130	0.432	179.913	0.683	-20.1
OAI CAH CAJ CAK	41	21.727	179.74	0.671	164.309
CAK CAL NAN CAS	130	0.432	179.913	0.655	167.107
CAR CAQ CAT NAV	92	40.084	175.405	0.641	-124.873
CAG CAH CAJ CAK	61	0.457	158.203	0.626	-13.383
CAJ CAK CAL NAN	204	52.484	179.845	0.47	128.586
OAM CAL CAK CAJ	251	0.036	127.702	0.197	-70.306
CAH CAJ CAK CAL	53	29.44	179.779	0.109	73.881
OAM CAL NAN CAS	348	0.084	179.698	0.052	6.295
OAU CAT CAQ CAR	92	6.098	142.382	0.024	59.436
CAP CAQ CAT NAV	92	40.084	175.405	0.02	110.777
OAM CAL NAN CAO	348	0.084	179.698	0.009	179.087

Given the CAD-CAE bond is rotatable, it is likely the prolyl ring is fairly flexible in solution, reflected by a diffuse but clear, average density.

Comparison of the experimentally derived ligand conformation was made to the poses derived from virtual screening (figure 7.11). These demonstrate overall very good prediction and agreement. Unusually, no alternate conformation was predicted within the five poses for compound 42. One potential explanation for this is that the very polar carbonyl-amide tail would not be well suited to the more hydrophobic top of the channel, whereas at the aperture position is has more chance of being solvent exposed.

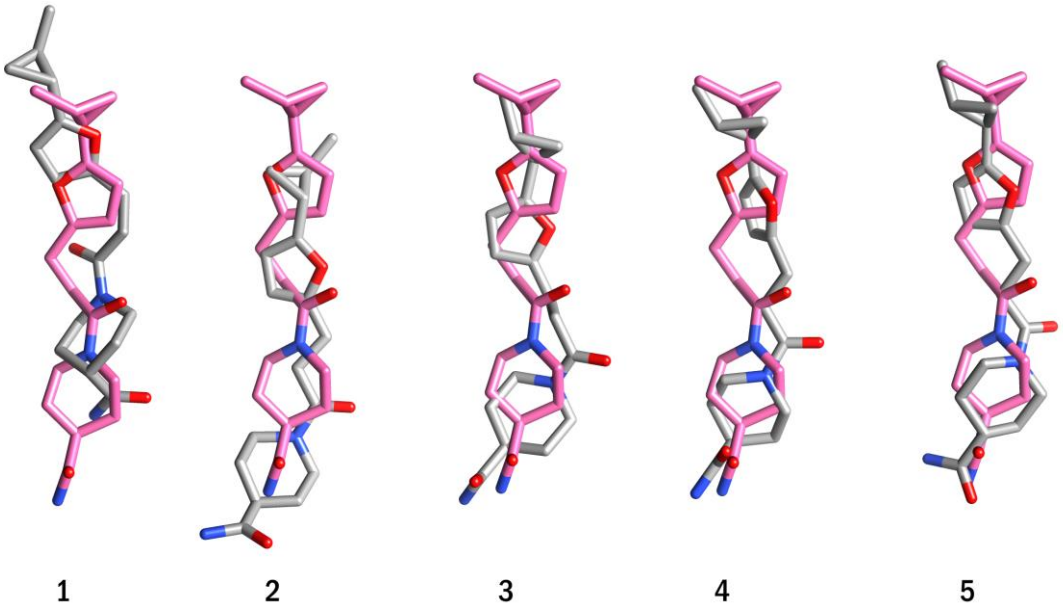


Figure 7.11: Comparison of the ranked ligand binding poses (grey) to the experimentally derived crystallographic ligand position (pink). Overall, this binding position was very well predicted.

Ultimately, the crystal structure of compound 42 shows ligands which do not form hydrogen bonds to EthR are still capable of binding with full occupancy, and that in some cases, EthR clearly supports only one ligand-binding orientation.

7.5.6 Compound 57

Compound 57 was one which showed a strong concentration-dependent shift relationship in the initial thermal shift assays performed with EthR. Diffraction data were refined to a resolution of 2.1 Å by methods previously detailed. The crystal structure of this compound with EthR shows one conformation of the ligand bound, with clearly defined, unambiguous ligand density.

Compound 57 (figure 7.12) forms a hydrogen bond with Thr149 through O22, the alcohol oxygen in the linker region of the ligand, donating a hydrogen to the lone pair of the threonine side chain oxygen at a distance of 2.8 Å. This ligand alcohol oxygen is also able to donate to the carbonyl of the Asn176 side chain at 2.8 Å distance. Finally, as anticipated from ligands bearing an amide carbonyl, the ligand carbonyl acts as a hydrogen bond acceptor to Asn179 at a distance of 2.9 Å.

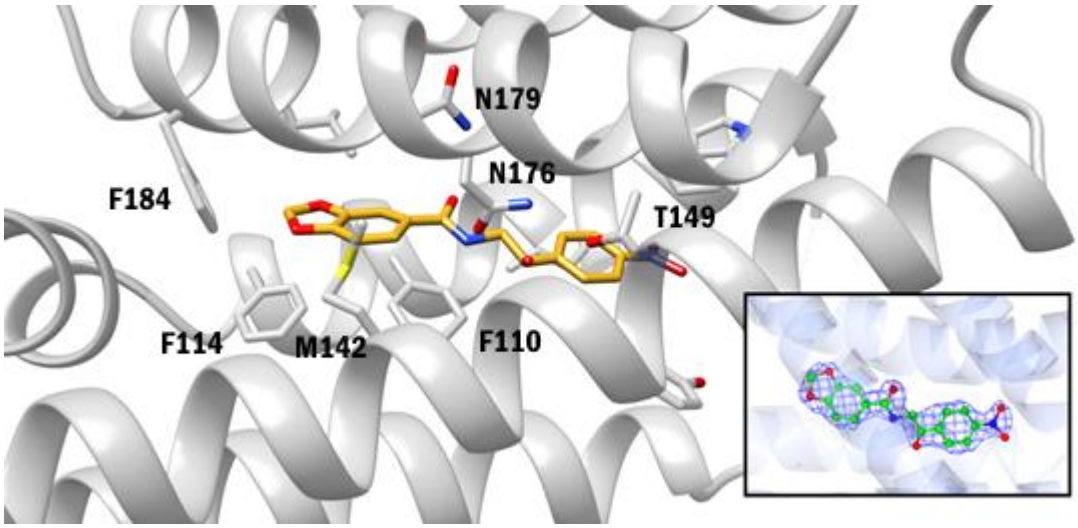


Figure 7.12: Close up image of ligand, compound 57, and the putative weak hydrogen bond formed with methionine via a CH···S interaction.

Additionally, the sulphur atom of Met142 is oriented such that it is in a position to form a C-H···S interaction with the benzyldioxole group at a distance of 3.4 Å (figure 7.13). Although weak interactions, intermolecular C-H···S hydrogen bonds have been observed. They are not well studied due to their infrequency in chemical crystal structures[178] but one notable study utilised experimental and theoretical methods to conclude sulphur is capable of weak interactions with aliphatic groups,[179] and another study utilising *ab initio* and DFT calculations determined such CH···S interactions follow findings on CH···O hydrogen bonds, where interaction strength is greatest for sp3 carbon donors, weaker for sp2 and weakest for sp donors.[180]

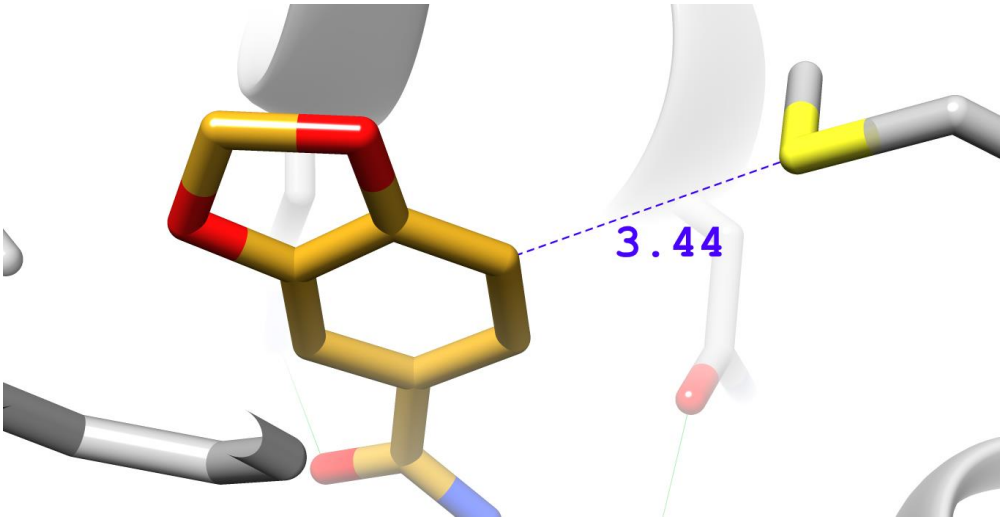


Figure 7.13: Close up image of ligand, compound 57, and the putative weak hydrogen bond formed with methionine residue 142 via a CH···S interaction.

If indeed this ligand is forming a weak hydrogen bond via Met142, it is not likely to greatly affect binding affinity; the energy predicted for CH···S interactions is around 0.25 kcal mol^{-1} (1.04 kJ mol^{-1}) and therefore a mere fraction of the energy gained by a typical NH···O hydrogen bond. However, such interactions are likely to contribute overall to stabilising and supporting the ligand within the binding site, much like hydrophobic/Van der Waals interactions, which are on a similar energetic scale.

Comparisons between the crystallographic ligand and the poses suggested by virtual screening (figure 7.14) show the third-ranked pose is a far better, almost identical, prediction than the top ranked pose, though this too is also very similar. Notably, the second-ranked pose positions the dioxolane oxygens where in the experimental structure these positions are occupied by the nitrile oxygens.

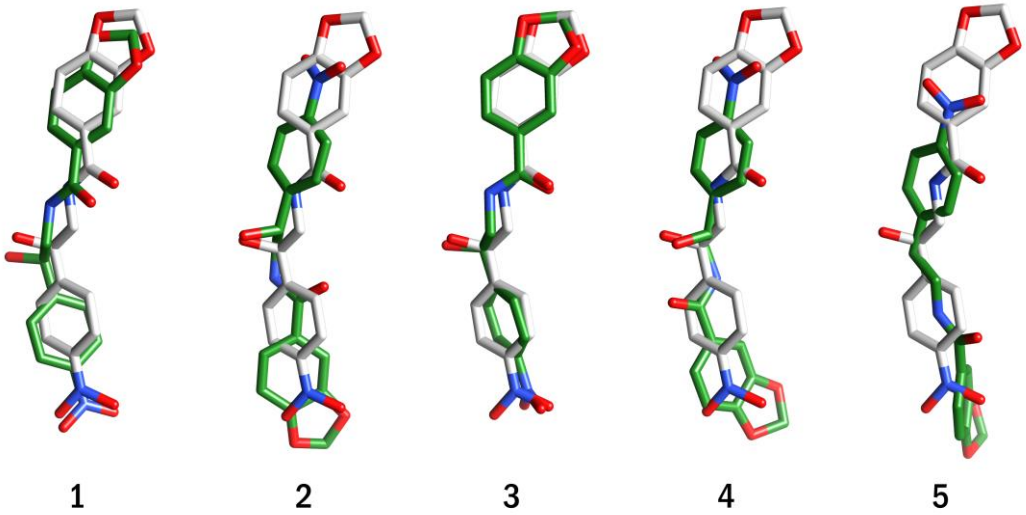


Figure 7.14: Comparison of the experimental ligand position (grey) with the virtual screening poses (green). Poses ranked 1 and 3 are very similar, whereas the lowest ranked poses are vastly different.

Again, two alternate binding modes are suggested, but here only one is confirmed experimentally. As before, this demonstrates excellent binding prediction but also the need for experimental validation.

7.5.7 Compound 60

Diffraction data for compound 60 were obtained to 1.6 Å and processed as previously described. Figure 7.15 shows the unbiased Fo-Fc map after rigid body refinement, water molecule placement and several rounds of restrained refinement with isotropic B-factors.

It was immediately apparent that the compound would not satisfy the density as one singular conformation as the density indicated a molecule longer than compound 60.

(NMR analysis on a sample of compound from stocks confirmed the ligand character, data not shown).

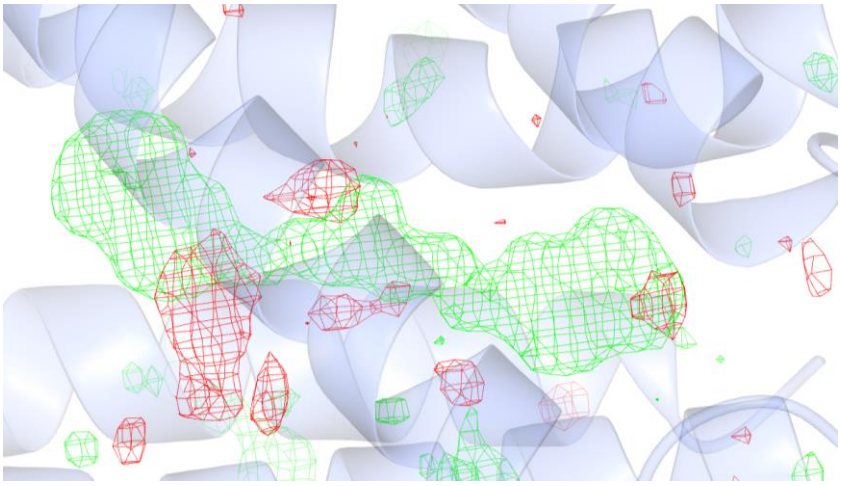


Figure 7.15: Unbiased Fo-Fc map (3σ) of the ligand binding site.

The virtual screening poses, shown in figure 7.16, demonstrate that one pose, with the benzodioxane at the bottom of the channel exposed to solvent is preferred over the alternate conformation.

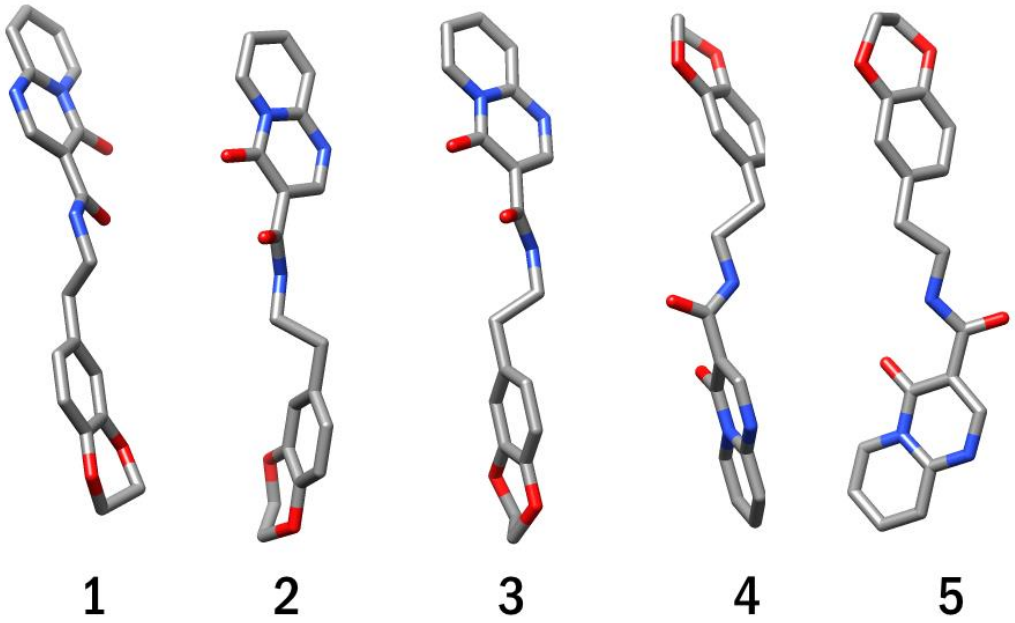


Figure 7.16: Ranked virtual screening poses of compound 60.

Figure 7.17 shows attempts to model both potential orientations of compound 60 in this quite symmetrical ligand density, with some success. However, as with compound 10, positive difference density (unbiased) remained for both models which indicated that

perhaps in this case, as with compound 10, multiple binding modes are present in one structure.

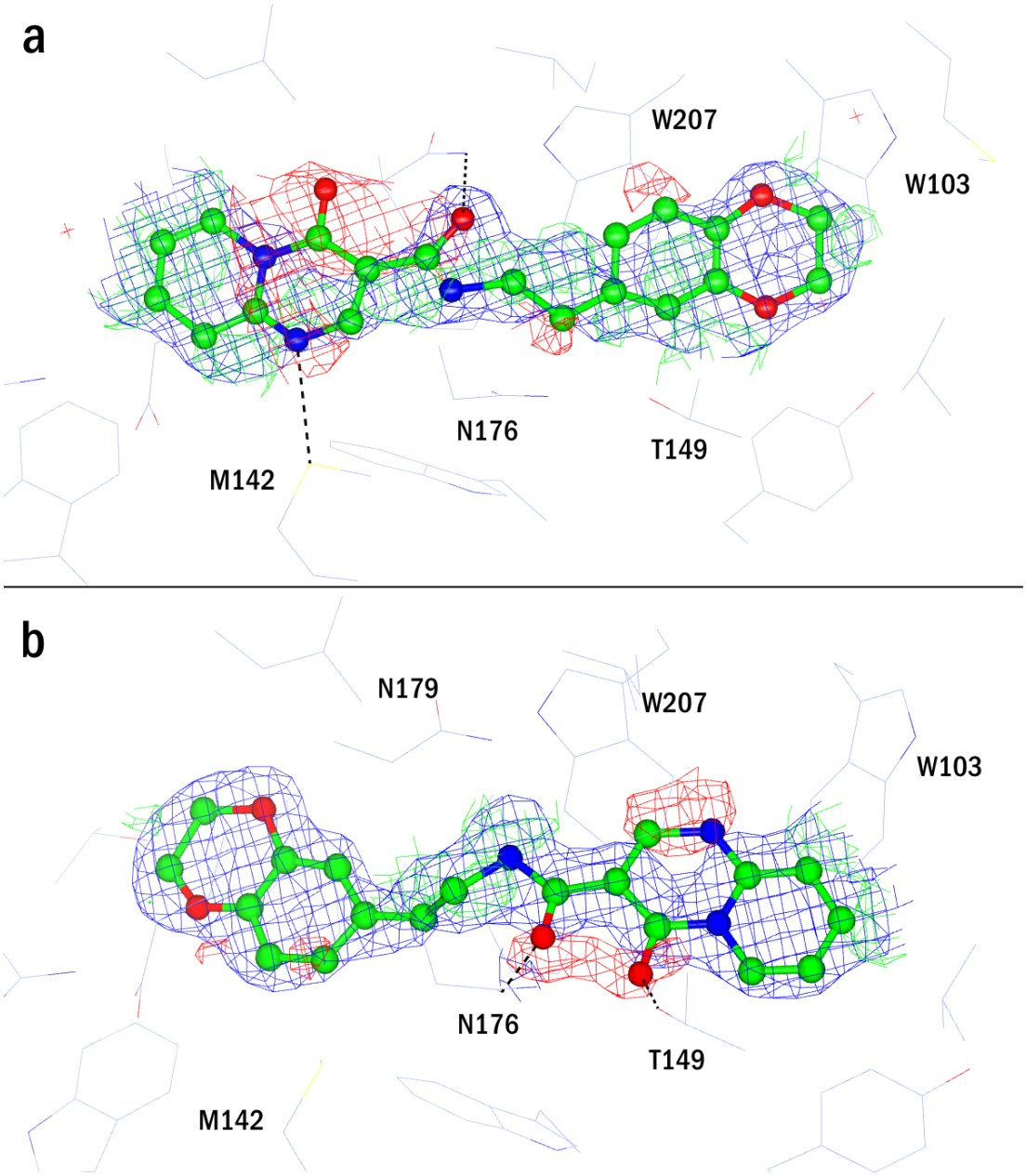


Figure 7.17: Each potential ligand orientation of compound 60 modelled in the density (2Fo-Fc, 1.5σ). The Fo-Fc map is also shown (green/red), contoured at 3 σ. Neither case thoroughly satisfies the density, and unbiased positive density remains.

In one orientation, the nitrogen of the conjugated pyrimidine is within a contact distance of 3.0 Å, and the peptide carbonyl forms the often-observed hydrogen bond to Asn179 at a distance of 2.8 Å. Alternatively, the carbonyls interact with Thr149 (2.7 Å, forming a hydrogen bond) and Asn176 (2.0 Å, a clash with a the carbonyl) in the opposite conformation.

Subsequently, both ligands were modelled simultaneously at 50% occupancy to give a final model which is the best interpretation of the ligand density available, with an Rwork/Rfree of 18.3%/22.1%. This model, however, continues to indicate significant positive difference density which is not covered by the ligands, as well as significant negative density over parts of the modellied ligand, indicating an insufficient fit (figure 7.18).

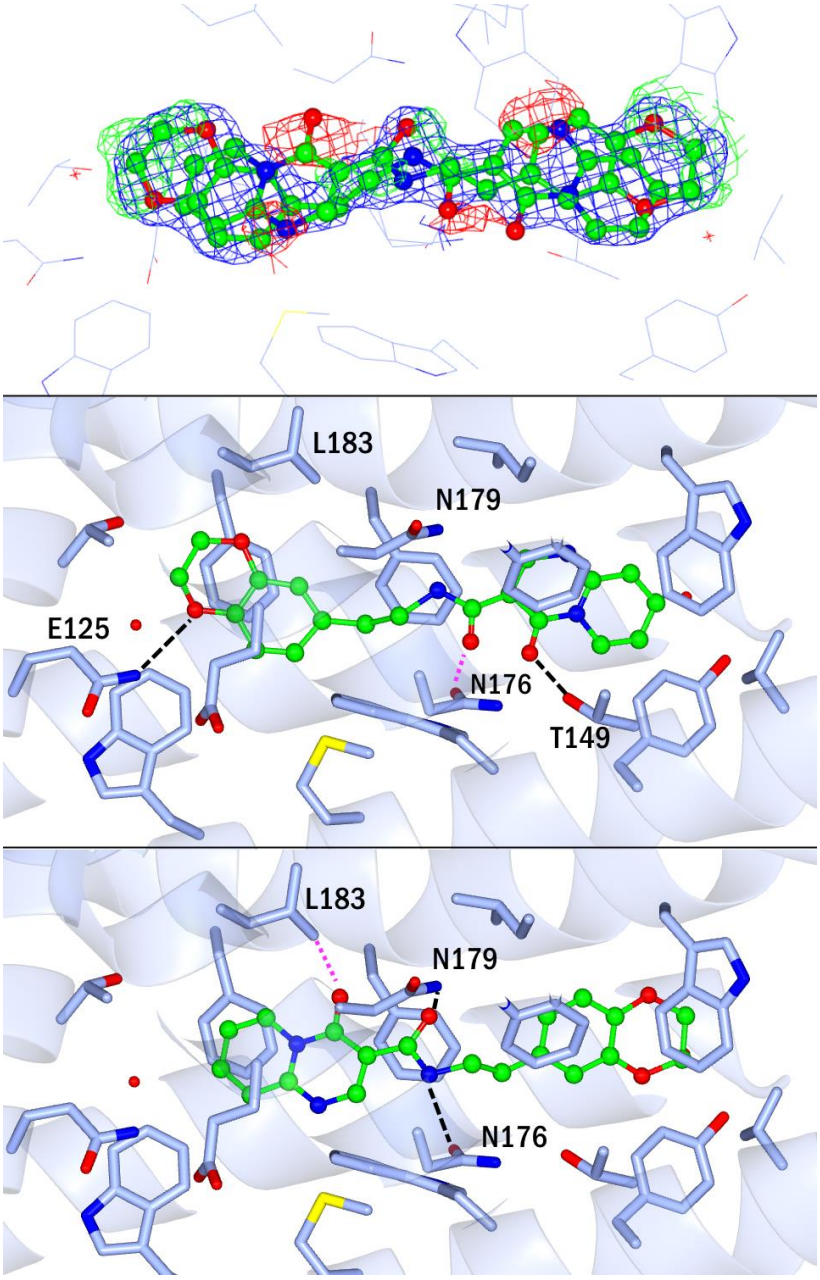


Figure 7.18: Top, ligand 2Fo-Fc map in blue, with unbiased Fo-Fc map in green (positive) and red (negative). The middle and bottom panels show each of the two orientations, with hydrogen bonds in black dashed lines and clashesin magenta dashed lines.

Figure 7.18 also shows this modelling, while also resulting in novel hydrogen bonds, results in close contacts or clashes with the protein. For example, the dioxane oxygens

are capable of hydrogen bonding to Glu125, previously unseen in any crystal structure. The middle panel of figure 7.18 shows this, and a hydrogen bond formed to Thr149. However, this causes a clash with Asn176 by the peptide carbonyl. The bottom panel of figure 7.18 shows the peptide nitrogen in one orientation forms a hydrogen bond to Asn176, and the carbonyl to Asn179. This results in the carbonyl on the conjugated pyrimidine to clash with Leu183.

In conclusion, this model is the best interpretation of the density, however further study (more crystal structures, derived from both co-crystallisation and ligand soaking) would be necessary to give more confidence to the modelling performed here.

7.5.8 Compound 74

For a crystal obtained from co-crystallisation experiments of EthR with compound 74, data were obtained to a resolution of 1.5 Å and refined without ligand as previously described. Difference density was immediately apparent, indicating a bound compound (figure 7.19).

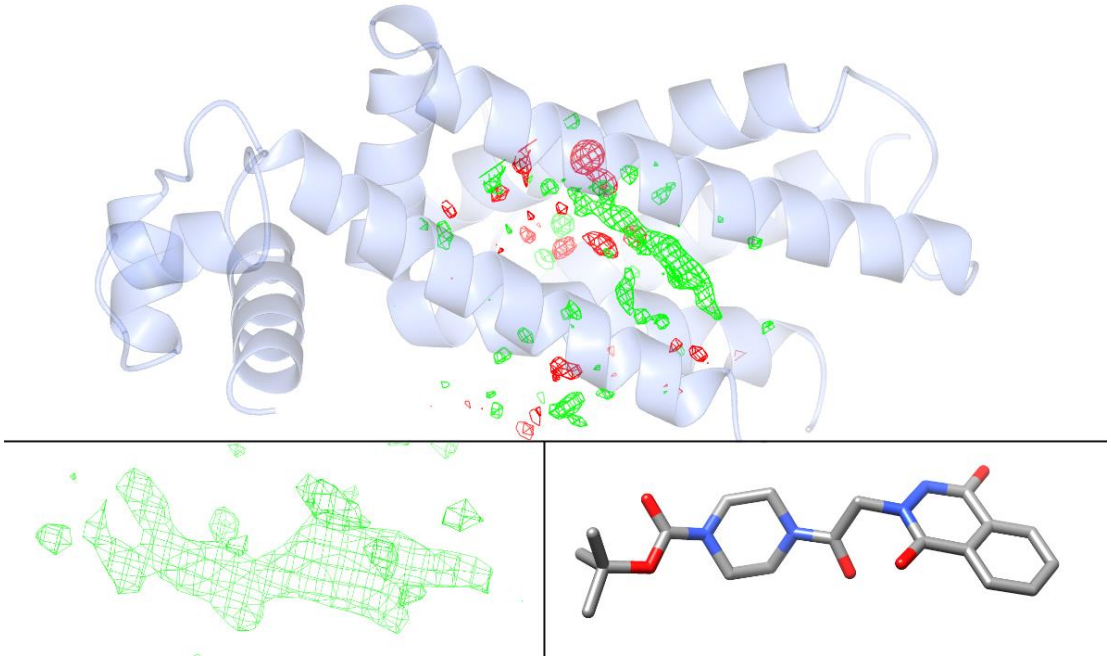


Figure 7.19: Though the difference density (3 σ) appeared to match superficially to compound 74, bottom left, all attempts to model the ligand proved fruitless.

However, it was not possible to fit compound 74 to this density, despite initial optimism. All exploratory attempts to model and refine compound 74 (figure 7.19) did not yield reasonable results in both model fit and chemical sensibility. The two potential

orientations suggested from virtual screening were modelled as proposed orientations (figure 7.20) but were unsuccessful.

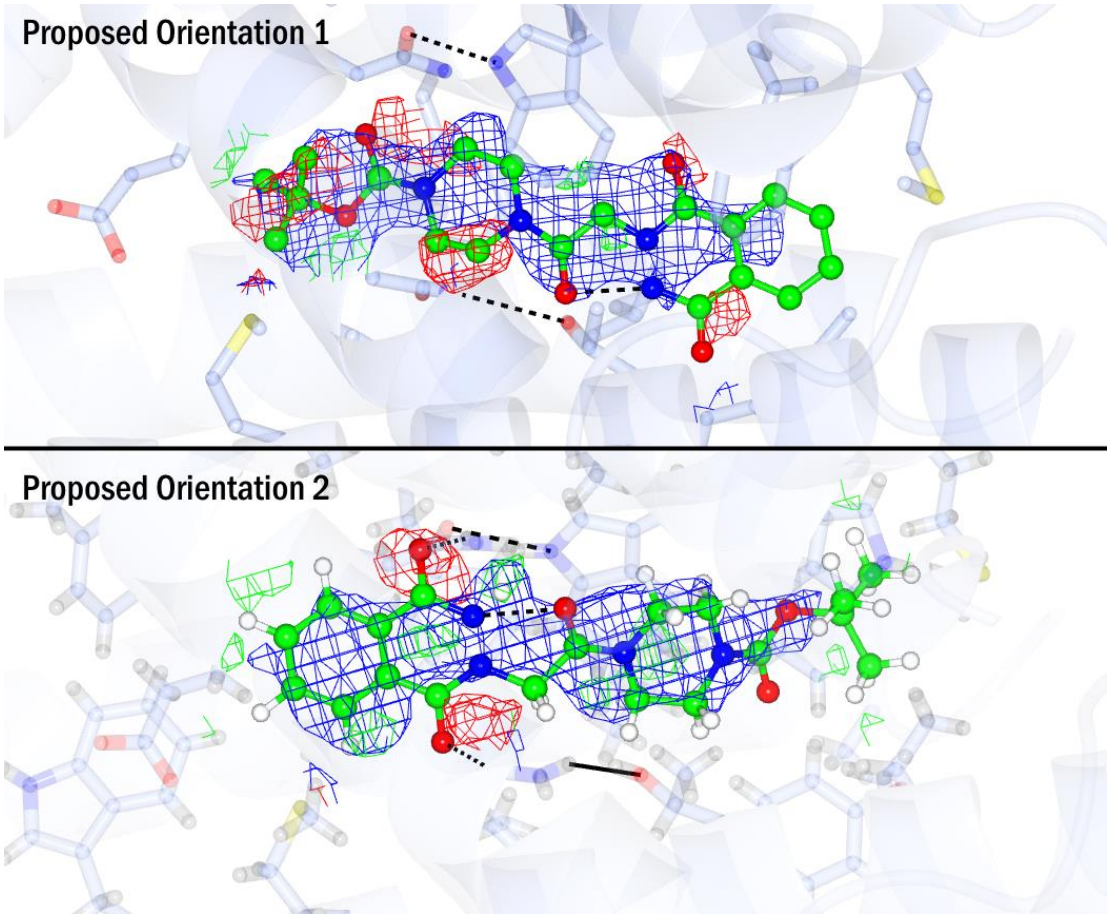


Figure 7.20: Ligand 2Fo-Fc density contoured at 1 σ, with Fo-Fc difference density at 3 σ. Neither ligand orientation (nor attempts to alter position laterally in binding site) were able to satisfy experimental data. Proposed orientation B shows hydrogen atoms to demonstrate ligand chemistry.

The inability to model either orientation satisfactorily suggests there *may* by multiple binding poses, which results in static disorder over the averaging of the unit cell contents. However, this cannot be demonstrated with the data at hand, and as the data were obtained to 1.5 Å it is unlikely higher resolution data will be sufficient to inform this structure further. Alternatively, it could be the result of compound degradation, either in DMSO solution or by radiation damage. Negative difference density can be observed on the sulphur atom of M142, which could be an indicator of the latter.

The results of co-crystallisation experiments with compound 74 are therefore inconclusive.

7.5.9 Compound 80

As with several other ligands, though difference density suggested presence of a ligand, it was not possible to confidently place nor model compound 80 in the structure obtained (resolution 1.7 Å). Figure 7.21 shows the difference density at 3 σ, however with the sigma level reduced as far as 1 σ, it was not possible to identify an orientation or confident position of the ligand for modelling.

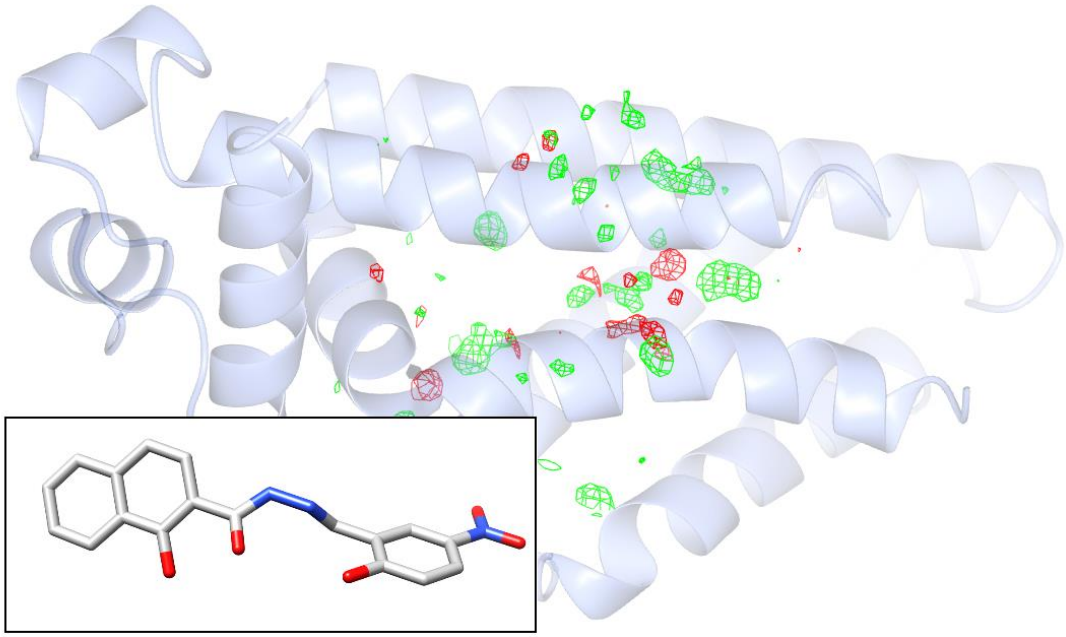


Figure 7.21: Difference density (3 σ) indicates the presence of a ligand in the binding site, but insufficient information can be derived from it to confidently model compound 80 (inset).

7.5.10 Compound 85

Compound 85 was of particular interest throughout the experimental phase of this research due to it being the highest scoring compound of the cohort for testing. This was further exacerbated by compound 85 being the only compound to show a negative, concentration-dependent shift in EthR melting temperature, not just in his cohort but also in screens by our collaborators (private communication). Finally, the SPR pilot experiments showed a positive trend in the dose-response curve, and crystals of EthR in co-crystallisation conditions were smaller and more abundant than observed with other compounds. It was hoped a crystal structure of EthR bound to compound 85 would be achieved which could be compared to the virtual screening results.

The crystal structure of EthR bound to compound 85 was refined to a resolution of 1.8 Å. Initial unbiased difference density was observed for the ligand at 3 σ (figure 7.22, inset) and was sufficient to model the ligand (figure 7.22).

Compound 85 is a long, lipophilic ligand which binds such that the F184 side-chain is forced up into its 'open' conformation to permit the oxane ring. A hydrogen bond is made between the carbonyl of the pyrimidinone to Asn179 at a distance of 2.6 Å. Additionally, a CH···O interaction is formed between the ligand and protein, between CAH and Thr148 (3.2 Å), further implying these weak hydrogen bonds play a role in ligand-binding stabilisation. Finally, the lack of density around the terminal methyl-substituted phenyl ring implies a high degree of rotation which is permissible in this more open region near the channel aperture.

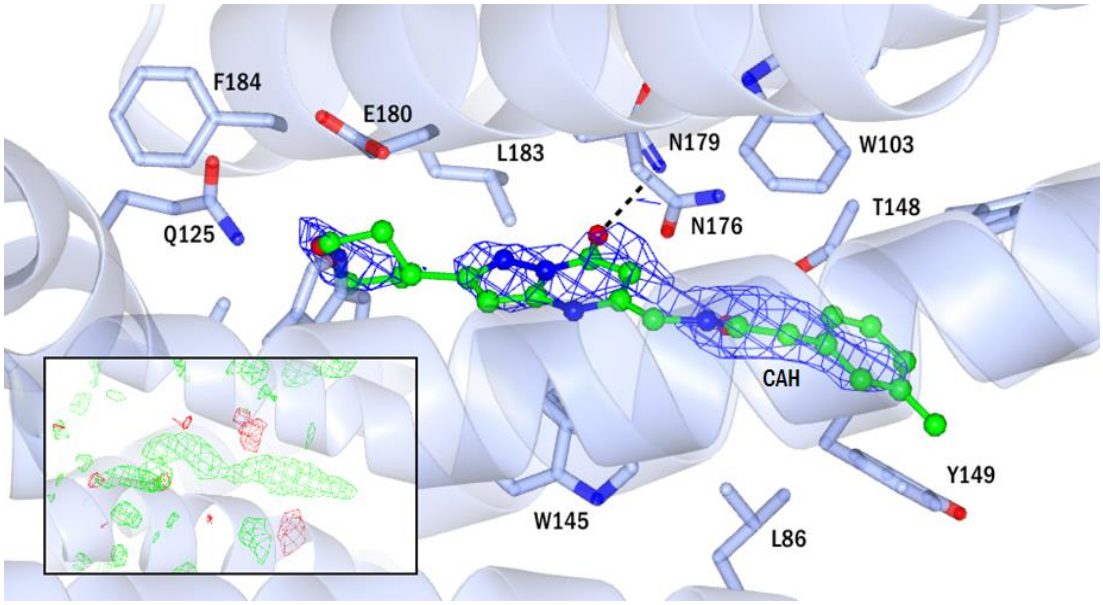


Figure 7.22: Compound 85 modelled with 2Fo-Fc density at 1 σ. Inset, unbiased Fo-Fc map at 3 σ.

Comparison of the crystal structure ligand with the virtual screening poses (figure 7.23) shows poor prediction of the binding mode. Only one ligand orientation is observed in the crystal structure, and only one ligand orientation was predicted in the virtual screening; these are not matched, and as such there is poor agreement between the crystal structure ligand and the predicted poses.

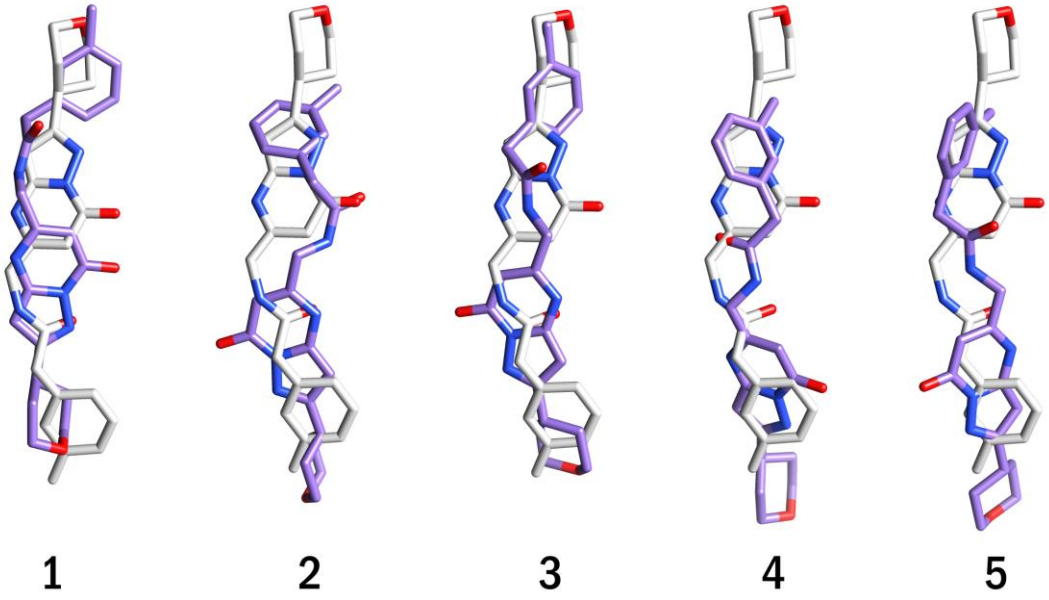


Figure 7.23: Comparison of the experimentally derived position of compound 85 (grey) to the ranked virtual screening poses (purple).

The crystal structure of compound 85 with EthR demonstrates supposedly "destabilising" compounds which cause mildly negative thermal denaturation shifts (chapter five) are no barrier to successful co-crystallisation, despite this seeming somewhat counterintuitive.

Finally, the poor agreement between prediction and experimental observation once again reinforces the need for experimental data over computational modelling when it comes to docking for drug discovery. And yet in this case, *in silico* screening predicted compound 85 to be a binder of EthR; though the screening was incorrect in predicting the mode and pose, the compound indeed binds EthR. This represents a mixed success for the computational modelling, but a great success for co-crystallisation of novel binding compounds with EthR.

7.6 Summary

Compounds with the potential to bind EthR were first identified by virtual screening, and tested by thermal shift assay to determine the likelihood of binding. This yielded 20 potential ligands which were used for co-crystallisation experiments.

Crystallisation of EthR alone was attempted first using known, published conditions. These were very successful and yielded crystals with the correct morphology, the correct unit cells, and which diffracted in-house to 3.5 Å. Co-crystallisation was then attempted with all 20 compounds. Crystals were obtained for many co-crystallisation experiments and were subsequently tested at Diamond Light Source on MX beamlines

I03 and I04-1. Data were obtained in the range of 2.1 – 1.4 Å for the EthR-compound complexes, and of the ten data-sets which initially showed ligand density, all but three were sufficient enough to model with good confidence.

For crystals obtained co-crystallised with compounds 15, 74 and 80, the density is not of sufficient quality to determine the ligand composition, identity, nor orientation with any confidence; as such no ligand can be modelled.

Crystal structures of EthR bound with compounds 3, 10, 25, 42, 57, 60 and 85 have been obtained, and give rise to interesting observations and conclusions. Firstly, compound 3 occupies a very different position to that predicted by virtual screening. This compound sits far lower in the binding channel than all other ligands co-crystallised with EthR, either presented here or published previously.

Most compounds bind EthR in only one orientation, with the absolute pose predicted by virtual screening to some success. Alternatively, compound 10 binds with two observable orientations, which has been suggested in the virtual screening results and docking work with EthR previously.

The crystal structure of compound 25 bound to EthR is the lowest resolution structure of EthR known to date at 1.4 Å. Compound 25 is also the smallest of the ligands presented here and demonstrates the key carbonyl-piperidine motif common to the published series of EthR inhibitors. This compound will therefore be helpful to SAR (structure-activity relationship) studies in altering that key chemistry in the search for a novel, potent series of compounds.

Compound 42 bound to EthR demonstrates hydrogen bonding between the ligand and protein is not necessary in this case to drive binding, due to the highly lipophilic nature of the EthR binding site. However, if this scaffold were to be adapted, it will be necessary to introduce hydrogen bonding between the ligand and protein in order to yield a potent binder.

The crystal structure of compound 57 showed one binding orientation, with hydrogen bonds made to both key asparagine residues. Moreover, there is some evidence for weak C-H···S hydrogen bonds which may contribute to overall binding stability.

Compound 60 was, similar to compound 10, modelled with two ligand conformations, however the hydrogen bonds formed are at the expense of clashes between the ligand and protein. In this case, further studies to repeatedly observe similar binding modes would give more confidence to the modelling performed.

Finally, compound 85, the highest scoring ligand from the virtual screening, was co-crystallised successfully with EthR, despite the thermal shift assay indicating a weak,

destabilising effect on the protein environment. The implied free rotation of the terminal methyl-substituted phenyl shows the area around the channel aperture allows a higher degree of flexibility in some cases, though was successfully occupied by a ligand in the case of compound 3. Compound 85 makes the key hydrogen bond to Asn179, but also is supported by a weak C-H···O interaction, stronger than the C-H···S interaction observed in compound 85 but unlikely to have a strong influence on binding. This further implies a pattern of weak hydrogen bonds between unlikely donors and classical acceptors which have the potential to stabilise ligand binding.

Further work is currently underway with several of these examples to model both ligand orientations by molecular dynamics methods and quantify the binding energy by the linear interaction method. Additionally, the calculation of per-residue interaction energies will allow the pinpointing of those amino acids which contribute to the protein-ligand binding event, and additionally quantify their role. Of particular interest are interactions such as C-H···S and C-H···O weak hydrogen bonds. For EthR, which has few classical hydrogen bonding opportunities in the binding site for targeting in hit-to-lead optimisation, the potential to exploit such interactions, albeit for small benefit, is promising.

In conclusion, the successful co-crystallisation of predicted EthR-binders has not only demonstrated the success of the virtual screening campaign and subsequent hit selection by thermal shift, but has also given rise to new observations (multiple binding modes, alternate binding positions, weak hydrogen bond interactions) which will prove useful and potentially vital to further anti-EthR drug development.

CONCLUSION

The overriding aim of the work presented here was to use computational methods to identify small molecules with unique scaffolds capable of binding EthR, with a view to developing those hits into new inhibitor series distinct from the current clinical lead candidate, BDM41906.[1]

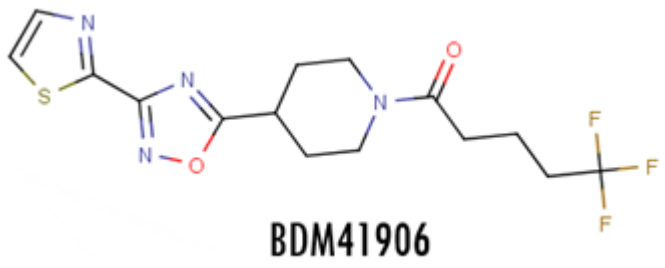


Figure 8.1: Chemical structure of current clinical lead candidate targeting EthR, BDM41906.[1]

To achieve this the ZINC database[2] was utilised as a source of small molecule compounds structures and the docking software GOLD[3–5] was used for virtual screening with those compounds. A protocol was derived through the use of improvised decoy sets (lacking negative data, a similarity set was created to test docking protocols) and knowledge of the protein structure.

Through physical and physiochemical filters implemented in KNIME[6] the initial compound database from ZINC was whittled down to 1.3 million compounds capable of fitting the binding site (ie. a volume < 700 Å^3) and with at least one ring structure, among other attributes. This smaller database of compounds was then clustered by ring chemistry as a method of sampling chemical space within a large set of 'drug like' molecules.[7–9] Roughly equal numbers of compounds were taken from each of the 25 clusters to produce a final screening set of 409,200 compounds.

Virtual screening was implemented in GOLD against EthR, with five poses generated per ligand. The decision was made to implement no additional constraints, for example on hydrogen bond formation, to open up such interactions as filters after docking. Ultimately, a total of over two million poses were generated for 409,195 compounds. The best pose of each ligand was extracted, and the top ten percent of those were imported into GoldMine, which was used in the post-screening phase to filter outliers and unfavourably interacting ligands, pruning the initial set of 40,919 compounds exponentially. A potential biophysical screening set of 284 compounds was

scrutinised manually and with the aid of Mogul to eliminate further those ligands which were, in a variety of ways, unsuitable; this included unusual torsion angles, poorly interacting groups or geometries, and finally duplicate scaffolds. Once those compounds which were no longer commercially available were removed, a cohort of 85 compounds remained for biophysical testing against EthR.

The first method of testing was by thermal shift assay[10] to identify those compounds capable of changing the melting temperature of EthR. This would indicate some interaction with the protein or environment. Of the 85 compounds purchased, only 83 could be tested – one failed quality control at the source and could not be acquired elsewhere (NJT07) and one was totally insoluble in both DMSO and buffer (NJT36). The thermal shift assay utilised four different concentrations of compound representing molar ratios of 2:1, 4:1, 8:1 and 10:1, and compounds which caused concentration-dependent changes in the EthR melting temperature were selected for further analysis. Of the 83 compounds, 20 showed this effect and were taken forward.

Pilot experiments for SPR binding analysis was conducted parallel to co-crystallisation experiments. For the SPR (surface plasmon resonance), previous experiments by a collaborating group[11] were used as a template. Up to six concentrations of each compound were injected over a CM5 surface modified with immobilised EthR, with an unmodified surface used as a control. Several compounds of the cohort of 20 could not be tested by SPR as the last of the compound stock had been used for co-crystallisation experiments. SPR indicated thirteen of the fourteen tested compounds identified by thermal shift assay bind EthR, corresponding to a confirmed hit rate of 16% for the tested biophysical set of 83 compounds.

To co-crystallise EthR with the twenty thermal shift assay hits, 9 mg ml^{-1} EthR was incubated overnight at room temperature with a final concentration of 3.3 mM compound (previously dissolved in 100% DMSO to 33 mM). Crystal structures were obtained for seven complexes with EthR, presented in chapter seven.

Ultimately, the approach utilised in the course of this work has yielded multiple hit compounds which bind EthR. The computational methods provided a cohort rich with true binding compounds, with a confirmed hit rate in the 83 chosen for analysis of 15.6%. This is highly respectable, but may only be the tip of the iceberg, as there are several hits from the thermal shift assay which remain untested, and the preliminary SPR experiments will need to be expanded to truly quantify EthR-ligand binding.

EthR is a target for cocktail therapies against *Mycobacterium tuberculosis* as it is part of the ethionamide bioactivation pathway. For the compounds presented here, two major questions remain: (1) though they bind EthR, do the compounds truly inhibit the

protein; and (2) do they act to reduce the necessary dose of ethionamide required to kill the bacteria.

To address the first question, we can certainly infer this to be the case; the crystal structures obtained so far show EthR bound and adopting the familiar inactive conformation. However to be certain, a further SPR assay with the target sequence of EthR will be necessary. The second question is the true test of the novel scaffolds, and is underway; several compounds from this cohort have been re-purchased and sent to our collaborators at the Institut Pasteur de Lille for testing against *Mycobacterium tuberculosis*. If successful, the compounds will then move forward into structure-activity relationship studies and the resulting optimisation will give rise to new potential clinical candidates for tackling tuberculosis.

REFERENCES

1. Nicolaou, K. C. Advancing the drug discovery and development Process. *Angew. Chem. Int. Ed.* (2014). doi:10.1002/anie.201404761
2. Cole, S. T. Who will develop new antibacterial agents? *Philos. Trans. R. Soc. B Biol. Sci.* **369,** 20130430 (2014).
3. Check Hayden, E. Ebola declared a public-health emergency. *Nature* (2014). doi:10.1038/nature.2014.15689
4. Fischbach, M. A. & Walsh, C. T. Antibiotics for Emerging Pathogens. *Science* **325,** 1089–1093 (2009).
5. Germann, P. G. *et al.* How to create innovation by building the translation bridge from basic research into medicinal drugs: an industrial perspective. *Hum. Genomics* **7,** 5 (2013).
6. Matthew Herper. The truly staggering cost of inventing new drugs. *Forbes* at http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/ Accessed 10th September 2014.
7. Light, D. W. & Warburton, R. Demythologizing the high costs of pharmaceutical research. *BioSocieties* **6,** 34–50 (2011).
8. Mullane, K., Winquist, R. J. & Williams, M. Translational paradigms in pharmacology and drug discovery. *Biochem. Pharmacol.* **87,** 189–210 (2014).
9. Morgan, P. *et al.* Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. *Drug Discov. Today* **17,** 419–424 (2012).
10. Landis, S. C. *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490,** 187–191 (2012).
11. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).
12. Shimura, H., Masuda, S. & Kimura, H. Research and development productivity map: visualization of industry status. *J. Clin. Pharm. Ther.* **39,** 175–180 (2014).
13. Tsukamoto, T. Tough times for medicinal chemists: Are we to blame? *ACS Med. Chem. Lett.* **4,** 369–370 (2013).
14. Six taken ill after drug trials. *BBC* (2006). at http://news.bbc.co.uk/1/hi/england/london/4807042.stm Accessed 22nd August 2014.
15. Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2,** 349–355 (2011).
16. Mannhold, R. & Rekker, R. F. The hydrophobic fragmental constant approach for calculating log P in octanol/water and aliphatic hydrocarbon/water systems. *Perspect. Drug Discov. Des.* **18,** 1–18 (2000).
17. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181,** 662–666 (1958).

18. Muirhead, H. & Perutz, M. F. Structure of haemoglobin: A three-dimensional fourier synthesis of reduced human haemoglobin at 5.5 Å resolution. *Nature* **199,** 633–638 (1963).
19. Levinthal, C., Wodak, S. J., Kahn, P. & Dadivanian, A. K. Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. *Proc. Natl. Acad. Sci.* **72,** 1330–1334 (1975).
20. Beddell, C. r., Goodford, P. J., Norrington, F. E., Wilkinson, S. & Wootton, R. Compounds designed to fit a site of known structure in human haemoglobin. *Br. J. Pharmacol.* **57,** 201–209 (1976).
21. Greer, J. & Bush, B. L. Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl. Acad. Sci.* **75,** 303–307 (1978).
22. Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221,** 709–713 (1983).
23. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T., E. A Geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161,** 269–288 (1982).
24. DesJarlais, R. L. *et al.* Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **31,** 722–729 (1988).
25. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13,** 505–524 (1992).
26. Richards, F. M. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **6,** 151–176 (1977).
27. Wodak, S. J., De Crombrugghe, M. & Janin, J. Computer studies of interactions between macromolecules. *Prog. Biophys. Mol. Biol.* **49,** 29–63 (1987).
28. Langridge, R., Ferrin, T. E., Kuntz, I. D. & Connolly, M. L. Real-time color graphics in studies of molecular interactions. *Science* **211,** 661–666 (1981).
29. Blaney, J. M. *et al.* Computer graphics in drug design: molecular modeling of thyroid hormone-prealbumin interactions. *J. Med. Chem.* **25,** 785–790 (1982).
30. Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **15,** 411–428 (2001).
31. Busetta, B., Tickle, I. J. & Blundell, T. L. DOCKER, an interactive program for simulating protein receptor and substrate interations. *J. Appl. Crystallogr.* **16,** 432–437 (1983).
32. DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D. & Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **29,** 2149–2153 (1986).
33. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).
34. Moore, G. E. Cramming more components onto integrated circuits. *Proc. IEEE* **86,** 82–85 (1998).
35. Leach, A. R. & Kuntz, I. D. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* **13,** 730–748 (1992).
36. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261,** 470–89 (1996).
37. Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **235,** 345–356 (1994).

38. Jones, G., Willett, P., Glen, R. C., Leach, a R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267,** 727–48 (1997).
39. Jones, G., Willett, P. & Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245,** 43–53 (1995).
40. Clark, D. E., Jones, G., Willett, P., Kenny, P. W. & Glen, R. C. Pharmacophoric pattern matching in files of three-dimensional chemical structures: Comparison of conformational-searching algorithms for flexible searching. *J. Chem. Inf. Comput. Sci.* **34,** 197–206 (1994).
41. Kramer, B., Rarey, M. & Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein–ligand docking. *Proteins* **37,** 228–241 (1999).
42. Böhm, H.-J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **6,** 61–78 (1992).
43. Groom, C. R. & Allen, F. H. The Cambridge Structural Database in retrospect and prospect. *Angew. Chem. Int. Ed.* **53,** 662–671 (2014).
44. Knegtel, R., Kuntz, I. D. & Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **266,** 424–440 (1997).
45. Huang, S.-Y. & Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins* **66,** 399–421 (2006).
46. Korb, O. *et al.* Potential and limitations of ensemble docking. *J. Chem. Inf. Model.* **52,** 1262–1274 (2012).
47. Taylor, J. S. & Burnett, R. M. DARWIN: a program for docking flexible molecules. *Proteins* **41,** 173–191 (2000).
48. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614 (2009).
49. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).
50. Rarey, M., Kramer, B. & Lengauer, T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins* **34,** 17–28 (1999).
51. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein–ligand docking using GOLD. *Proteins* **52,** 609–623 (2003).
52. Verdonk, M. L. *et al.* Modeling Water molecules in protein–ligand docking using GOLD. *J. Med. Chem.* **48,** 6504–6515 (2005).
53. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D. & Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **60,** 325–332 (2005).
54. Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P. & DesJarlais, R. L. Docking: successes and challenges. *Curr. Pharm. Des.* **11,** 323–333 (2005).
55. Hartshorn, M. J. *et al.* Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **50,** 726–741 (2007).
56. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49,** 6789–6801 (2006).
57. Durrant, J. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9,** 71 (2011).
58. Cavasotto, C. N. *et al.* Discovery of novel chemotypes to a g-protein-coupled receptor through ligand-steered homology modeling and structure-based virtual screening. *J. Med. Chem.* **51,** 581–588 (2008).

59. Li, Y. *et al.* Discovery of novel checkpoint kinase 1 inhibitors by virtual screening based on multiple crystal structures. *J. Chem. Inf. Model.* **51,** 2904–2914 (2011).
60. Sager, G. *et al.* Novel cGMP efflux inhibitors identified by virtual ligand screening (VLS) and confirmed by experimental studies. *J. Med. Chem.* **55,** 3049–3057 (2012).
61. Izumizono, Y., Arevalo, S., Koseki, Y., Kuroki, M. & Aoki, S. Identification of novel potential antibiotics for tuberculosis by in silico structure-based drug screening. *Eur. J. Med. Chem.* **46,** 1849–1856 (2011).
62. Xu, L. *et al.* Molecular modeling of the 3D structure of 5-HT1AR: Discovery of novel 5-HT1AR agonists via dynamic pharmacophore-based virtual screening. *J. Chem. Inf. Model.* **53,** 3202–3211 (2013).
63. Koch, O. *et al.* Identification of *M. tuberculosis* thioredoxin reductase inhibitors based on high-throughput docking using constraints. *J. Med. Chem.* **56,** 4849–4859 (2013).
64. Sato, H., Shewchuk, L. M. & Tang, J. Prediction of multiple binding modes of the CDK2 inhibitors, anilinopyrazoles, using the automated docking programs GOLD, FlexX, and LigandFit: An evaluation of performance. *J. Chem. Inf. Model.* **46,** 2552–2562 (2006).
65. Huang, S.-Y. & Zou, X. Efficient molecular docking of NMR structures: Application to HIV-1 protease. *Protein Sci.* **16,** 43–51 (2007).
66. Cheng, L. S. *et al.* Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **51,** 3878–3894 (2008).
67. Pauli, I. *et al.* Discovery of new inhibitors of *Mycobacterium tuberculosis* InhA enzyme using virtual screening and a 3D-pharmacophore-based approach. *J. Chem. Inf. Model.* **53,** 2390–2401 (2013).
68. Cambridge Crystallographic Data Centre. GOLD User Guide. at https://www.ccdc.cam.ac.uk/Lists/DocumentationList/gold_52.pdf Accessed 28th August 2014.
69. Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. & Eldridge, M. D. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins.* **33,** 367–382 (1998).
70. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **11,** 425–45 (1997).
71. Korb, O., Stutzle, T. & Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **49,** 84–96 (2009).
72. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multi-threading. *J. Comput. Chem.* **31,** 455–461 (2010).
73. Irwin, J. J. & Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45,** 177–182 (2005).
74. Mazanetz, M. P., Marmon, R. J., Reisser, C. B. T. & Morao, I. Drug discovery applications for KNIME: an open source data mining platform. *Curr. Top. Med. Chem.* **12,** 1965–1979 (2012).
75. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Model.* **43,** 493–500 (2003).

76. Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B* **58,** 380–388 (2002).
77. Liebeschuetz, J., Hennemann, J., Olsson, T. & Groom, C. R. The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *J. Comput. Aided Mol. Des.* **26,** 169–183 (2012).
78. Talele, T., Khedkar, S. & Rigby, A. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* **10,** 127–141 (2010).
79. Scozzafava, A. & Supuran, C. T. Glaucoma and the applications of carbonic anhydrase inhibitors. *Subcell. Biochem.* **75,** 349–359 (2014).
80. Baldwin, J. J. *et al.* Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. *J. Med. Chem.* **32,** 2510–2513 (1989).
81. Loftsson, T., Jansook, P. & Stefánsson, E. Topical drug delivery to the eye: dorzolamide. *Acta Ophthalmol. (Copenh.)* **90,** 603–608 (2012).
82. Pinard, M. A., Boone, C. D., Rife, B. D., Supuran, C. T. & McKenna, R. Structural study of interaction between brinzolamide and dorzolamide inhibition of human carbonic anhydrases. *Bioorg. Med. Chem.* **21,** 7210–7215 (2013).
83. Bissantz, C., Kuhn, B. & Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **53,** 5061–5084 (2010).
84. Varghese, J. N., Laver, W. G. & Colman, P. M. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature* **303,** 35–40 (1983).
85. Woods, J. M. *et al.* 4-Guanidino-2,4-dideoxy-2,3-dehydro-N-acetylneuraminic acid is a highly effective inhibitor both of the sialidase (neuraminidase) and of growth of a wide range of influenza A and B viruses in vitro. *Antimicrob. Agents Chemother.* **37,** 1473–1479 (1993).
86. von Itzstein, M. *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363,** 418–423 (1993).
87. Varghese, J. N., Epa, V. C. & Colman, P. M. Three-dimensional structure of the complex of 4-guanidino-Neu5Ac2en and influenza virus neuraminidase. *Protein Sci* **4,** 1081–1087 (1995).
88. Xu, X., Zhu, X., Dwek, R. A., Stevens, J. & Wilson, I. A. Structural characterization of the 1918 influenza virus H1N1 neuraminidase. *J.Virol.* **82,** 10493–10501 (2007).
89. Vavricka, C. J. *et al.* Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for influenza NA inhibition. *Plos Pathog* **7,** e1002249–e1002249 (2011).
90. Wang, M. Y. *et al.* Influenza a virus n5 neuraminidase has an extended 150-cavity. *J.Virol.* **85,** 8431–8435 (2011).
91. Wu, Y. *et al.* Characterization of two distinct neuraminidases from avian-origin human-infecting H7N9 influenza viruses. *Cell Res* **23,** 1347–1355 (2013).
92. Escuret, V. *et al.* A novel I221L substitution in neuraminidase confers high level resistance to oseltamivir in influenza B viruses. *J.Infect.Dis.* **10**, 1260-1269 (2014).
93. World Health Organization. *Global Tuberculosis Report 2014.* (World Health Organization, 2015).
94. Volmink, J. & Garner, P. Directly observed therapy for treating tuberculosis. *Cochrane Database Syst. Rev.* (2007). doi:10.1002/14651858.CD003343.pub3

95. Zumla, A., Nahid, P. & Cole, S. T. Advances in the development of new tuberculosis drugs and treatment regimens. *Nat. Rev. Drug Discov.* **12,** 388–404 (2013).
96. Villarino, M. E., Geiter, L. J. & Simone, P. M. The multidrug-resistant tuberculosis challenge to public health efforts to control tuberculosis. *Public Health Rep.* **107,** 616–625 (1992).
97. Jain, A. & Mondal, R. Extensively drug-resistant tuberculosis: current challenges and threats. *FEMS Immunol. Med. Microbiol.* **53,** 145–150 (2008).
98. World Health Organization. *Global Tuberculosis Report 2013.* (World Health Organization, 2014).
99. Laxminarayan, R. *et al.* Antibiotic resistance—the need for global solutions. *Lancet Infect. Dis.* **13,** 1057–1098 (2013).
100. Ma, Z., Lienhardt, C., McIlleron, H., Nunn, A. J. & Wang, X. Global tuberculosis drug development pipeline: the need and the reality. *The Lancet* **375,** 2100–2109 (2012).
101. Prasad, R., Verma, S. K., Sahai, S., Kumar, S. & Jain, A. Efficacy and safety of kanamycin, ethionamide, PAS and cycloserine in multidrug-resistant pulmonary tuberculosis patients. *Indian J. Chest Dis. Allied Sci.* **48,** 183–186 (2006).
102. Ormerod, L. P. Multidrug-resistant tuberculosis (MDR-TB): epidemiology, prevention and treatment. *Br. Med. Bull.* **73-74,** 17–24 (2005).
103. Caminero, J. A., Sotgiu, G., Zumla, A. & Migliori, G. B. Best drug treatment for multidrug-resistant and extensively drug-resistant tuberculosis. *Lancet Infect. Dis.* **10,** 621–629 (2010).
104. Steenken, W., Jr & Montalbine, V. The antituberculous activity of thioamide in vitro and in the experimental animal (mouse and guinea pig). *Am. Rev. Respir. Dis.* **81,** 761–763 (1960).
105. Weinstein, H. J., Hallett, W. Y. & Sarauw, A. S. The absorption and toxicity of ethionamide. *Am. Rev. Respir. Dis.* **86,** 576–578 (1962).
106. Wang, F. *et al.* Mechanism of thioamide drug action against tuberculosis and leprosy. *J. Exp. Med.* **204,** 73–78 (2007).
107. Marrakchi, H., Lanéelle, G. & Quémard, A. InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. *Microbiology* **146,** 289–296 (2000).
108. Baulard, A. R. *et al.* Activation of the pro-drug ethionamide is regulated in mycobacteria. *J. Biol. Chem.* **276,** 28326–28331 (2000).
109. Pym, A. S. *et al.* Regulation of catalase–peroxidase (KatG) expression, isoniazid sensitivity and virulence by *furA* of *Mycobacterium tuberculosis*. *Mol. Microbiol.* **40,** 879–889 (2001).
110. Holdiness, M. R. Neurological manifestations and toxicities of the antituberculosis drugs. A review. *Med. Toxicol.* **2,** 33–51 (1987).
111. Frénois, F., Engohang-Ndong, J., Locht, C., Baulard, A. R. & Villeret, V. Structure of EthR in a ligand bound conformation reveals therapeutic perspectives against tuberculosis. *Mol. Cell* **16,** 301–307 (2004).
112. Vilchèze, C. *et al.* Inactivation of the *inhA*-encoded fatty acid synthase II (FASII) enoyl-acyl carrier protein reductase induces accumulation of the FASI end products and cell lysis of *Mycobacterium smegmatis*. *J. Bacteriol.* **182,** 4059–4067 (2000).

113. Dover, L. G. *et al.* Crystal Structure of the TetR/CamR Family Repressor *Mycobacterium tuberculosis* EthR Implicated in Ethionamide Resistance. *J. Mol. Biol.* **340,** 1095–1105 (2004).
114. Brennan, R. G. The winged-helix DNA-binding motif: another helix-turn-helix takeoff. *Cell* **74,** 773–776 (1993).
115. Ramos, J. L. *et al.* The TetR family of transcriptional repressors. *Microbiol. Mol. Biol. Rev.* **69,** 326–356 (2005).
116. Fraaije, M. W., Kamerbeek, N. M., Heidekamp, A. J., Fortin, R. & Janssen, D. B. The prodrug activator EtaA from *Mycobacterium tuberculosis* is a Baeyer-Villiger monooxygenase. *J. Biol. Chem.* **279,** 3354–3360 (2003).
117. Pabo, C. O. & Sauer, R. T. Protein-DNA recognition. *Annu. Rev. Biochem.* **53,** 293–321 (1984).
118. Frénois, F., Baulard, A. R. & Villeret, V. Insights into mechanisms of induction and ligands recognition in the transcriptional repressor EthR from Mycobacterium tuberculosis. *Tuberculosis* **86,** 110–114 (2006).
119. Willand, N. *et al.* Synthetic EthR inhibitors boost antituberculous activity of ethionamide. *Nat. Med.* **15,** 537–544 (2009).
120. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **44,** 235–249 (2000).
121. Engohang-Ndong, J. *et al.* EthR, a repressor of the TetR/CamR family implicated in ethionamide resistance in mycobacteria, octamerizes cooperatively on its operator: Regulation of ethionamide resistance in mycobacteria. *Mol. Microbiol.* **51,** 175–188 (2004).
122. Tatum, N. J. *et al.* Structural and docking studies of potent ethionamide boosters. *Acta Cryst* **C69,** 1243–1250 (2013).
123. Carette, X. *et al.* Structural activation of the transcriptional repressor EthR from *Mycobacterium tuberculosis* by single amino acid change mimicking natural and synthetic ligands. *Nucleic Acids Res.* **40,** 3018–3030 (2011).
124. Willand, N. *et al.* Exploring drug target flexibility using *in situ* click chemistry: application to a mycobacterial transcriptional regulator. *ACS Chem. Biol.* **5,** 1007–1013 (2010).
125. Hu, X. & Manetsch, R. Kinetic target-guided synthesis. *Chem. Soc. Rev.* **39,** 1316–1324 (2010).
126. Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins Struct. Funct. Bioinforma.* **39,** 261–268 (2000).
127. Flipo, M. *et al.* Ethionamide boosters: synthesis, biological activity, and structure–activity relationships of a series of 1,2,4-oxadiazole EthR inhibitors. *J. Med. Chem.* **54,** 2994–3010 (2011).
128. Flipo, M. *et al.* Ethionamid boosters. 2. Combining bioisosteric replacement and structure-based drug design to solve pharmacokinetic issues in a series of potent 1,2,4-oxadiazole EthR inhibitors. *J. Med. Chem.* **55,** 68–83 (2012).
129. Flipo, M. *et al.* Discovery of novel *N*-phenylphenoxyacetamide derivatives as EthR inhibitors and ethionamide boosters by combining high-throughput screening and synthesis. *J. Med. Chem.* **55,** 6391–6402 (2012).

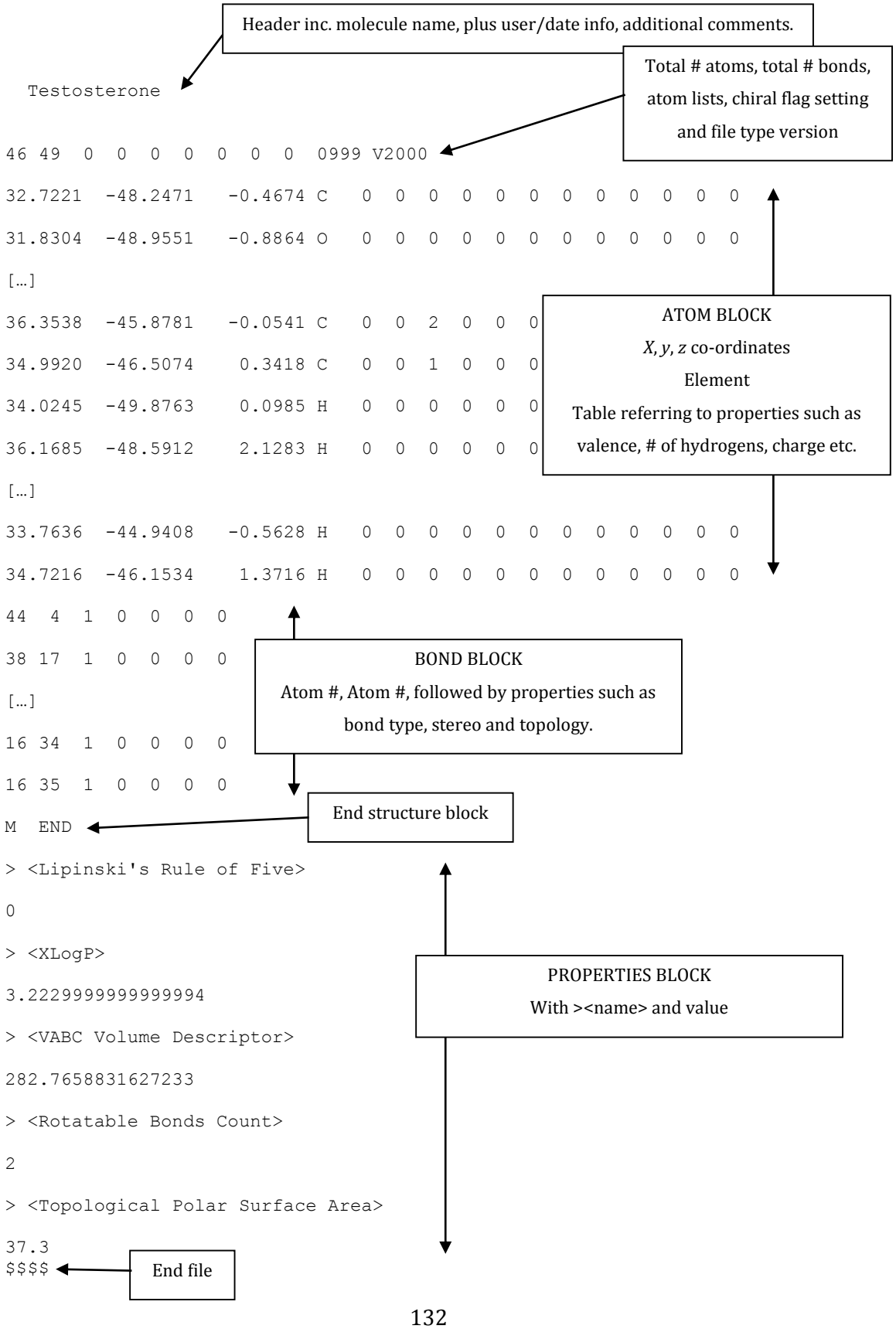
130. Surade, S. *et al.* A structure-guided fragment-based approach for the discovery of allosteric inhibitors targeting the lipophilic binding site of transcription factor EthR. *Biochem. J.* (2013). doi:10.1042/BJ20131127
131. Villemagne, B. *et al.* Ligand efficiency driven design of new inhibitors of *Mycobacterium tuberculosis* transcriptional repressor EthR using fragment growing, merging, and linking approaches. *J. Med. Chem.* **57,** 4876–4888 (2014).
132. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32,** 40–51 (2014).
133. Hendlich, M., Bergner, A., Günther, J. & Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.* **326,** 607–620 (2003).
134. Deng, W., Li, C. & Xie, J. The underling mechanism of bacterial TetR/AcrR family transcriptional repressors. *Cell. Signal.* **25,** 1608–1613 (2013).
135. Flipo, M. *et al.* RCSB Protein Data Bank - RCSB PDB - 3Q0U Structure Summary. (2010). at http://www.rcsb.org/pdb/explore/explore.do?structureId=3q0u Accessed 3rd June 2014
136. Flipo, M. *et al.* RCSB Protein Data Bank - RCSB PDB - 3Q0V Structure Summary. (2010). at http://www.rcsb.org/pdb/explore/explore.do?structureId=3q0v Accessed 3rd June 2014
137. Bergner, A., Günther, J., Hendlich, M., Klebe, G. & Verdonk, M. Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects. *Biopolymers* **61,** 99–110 (2001).
138. Schmitt, S., Hendlich, M. & Klebe, G. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chem. Int. Ed.* **40,** 3141–3144 (2001).
139. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15,** 359–363, 389 (1997).
140. Evers, A. & Klabunde, T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the α1A adrenergic receptor. *J. Med. Chem.* **48,** 1088–1097 (2005).
141. Irwin, J. J. & Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45,** 177–182 (2005).
142. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45,** 2615–2623 (2002).
143. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53,** 2719–2740 (2010).
144. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39,** 2887–2893 (1996).
145. Wallach, I., Jaitly, N., Nguyen, K., Schapira, M. & Lilien, R. Normalizing molecular docking rankings using virtually generated decoys. *J. Chem. Inf. Model.* **51,** 1817–1830 (2011).
146. Cardamone, M. & Puri, N. K. Spectrofluorimetric assessment of the surface hydrophobicity of proteins. *Biochem. J.* **282 (Pt 2),** 589–593 (1992).

147. Takashi, R., Tonomura, Y. & Morales, M. F. 4, 4'-Bis (1-anilinonaphthalene 8-sulfonate)(bis-ANS): a new probe of the active site of myosin. *Proc. Natl. Acad. Sci.* **74,** 2334–2338 (1977).
148. Anraku, M., Yamasaki, K., Maruyama, T., Kragh-Hansen, U. & Otagiri, M. Effect of oxidative stress on the structure and function of human serum albumin. *Pharm. Res.* **18,** 632–639 (2001).
149. Vetri, V. *et al.* Amyloid fibrils formation and amorphous aggregation in concanavalin A. *Biophys. Chem.* **125,** 184–190 (2007).
150. Lindgren, M., Sörgjerd, K. & Hammarström, P. Detection and characterization of aggregates, prefibrillar amyloidogenic oligomers, and protofibrils using fluorescence spectroscopy. *Biophys. J.* **88,** 4200–4212 (2005).
151. Echalier, A., Hole, A. J., Lolli, G., Endicott, J. A. & Noble, M. E. M. An inhibitor's-eye view of the ATP-binding site of CDKs in different regulatory states. *ACS Chem. Biol.* **9,** 1251–1256 (2014).
152. Acharya, P. & Rao, N. M. Stability studies on a lipase from *Bacillus subtilis* in guanidinium chloride. *J. Protein Chem.* **22,** 51–60 (2003).
153. Chang, T.-L. & Cheung, H. C. A model for molecules with twisted intramolecular charge transfer characteristics: solvent polarity effect on the non-radiative rates of dyes in a series of water—ethanol mixed solvents. *Chem. Phys. Lett.* **173,** 343–348 (1990).
154. Das, K., Sarkar, N., Nath, D. & Bhattacharyya, K. Non-radiative pathways of anilino-naphthalene sulphonates: Twisted intramolecular charge transfer versus intersystem crossing. *Spectrochim. Acta Part Mol. Spectrosc.* **48,** 1701–1705 (1992).
155. Grøftehauge, M. K., Hajizadeh, N. R., Swann, M. J. & Pohl, E. Protein–ligand interactions investigated by thermal shift assays (TSA) and dual polarization interferometry (DPI). *Acta Crystallogr. D Biol. Crystallogr.* **71,** 36–44 (2015).
156. Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N. & Nordlund, P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* **357,** 289–298 (2006).
157. Vedadi, M. *et al.* Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 15835–15840 (2006).
158. Genick, C. C. *et al.* Applications of biophysics in high-throughput screening hit validation. *J. Biomol. Screen.* **19,** 707–714 (2014).
159. Crauste, C. *et al.* Unconventional surface plasmon resonance signals reveal quantitative inhibition of transcriptional repressor EthR by synthetic ligands. *Anal. Biochem.* **452,** 54–66 (2014).
160. Rich, R. L. & Myszka, D. G. Higher-throughput, label-free, real-time molecular interaction analysis. *Anal. Biochem.* **361,** 1–6 (2007).
161. Madeira, A. *et al.* Coupling surface plasmon resonance to mass spectrometry to discover novel protein–protein interactions. *Nat. Protoc.* **4,** 1023–1037 (2009).
162. Teng, T. Y. Mounting of crystals for macromolecular crystallography in a free-standing thin film. *J. Appl. Crystallogr.* **23,** 387–391 (1990).
163. Incardona, M.-F. *et al. EDNA* : a framework for plugin-based applications applied to X-ray experiment online data analysis. *J. Synchrotron Radiat.* **16,** 872–879 (2009).
164. Leslie, A. G. W. & Powell, H. R. in *Evolving Methods for Macromolecular Crystallography* (eds. Read, R. J. & Sussman, J. L.) p41–51 (Springer Netherlands,

2007). at http://link.springer.com/chapter/10.1007/978-1-4020-6316-9_4 Accessed 2nd March 2015
165. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 125–132 (2010).
166. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69,** 1204–1214 (2013).
167. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 1131–1137 (2003).
168. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 355–367 (2011).
169. McCoy, A. J. *et al. Phaser* crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
170. Read, R. J. & McCoy, A. J. Using SAD data in *Phaser*. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 338–344 (2011).
171. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501 (2010).
172. Schüttelkopf, A. W. & van Aalten, D. M. F. *PRODRG* : a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 1355–1363 (2004).
173. McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 386–394 (2011).
174. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
175. Aqvist, J. & Marelius, J. The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High Throughput Screen.* **4,** 613–626 (2001).
176. Almlöf, M., Carlsson, J. & Åqvist, J. Improving the accuracy of the linear interaction energy method for solvation free energies. *J. Chem. Theory Comput.* **3,** 2162–2175 (2007).
177. Olson, M. E. *et al.* Oxidative reactivities of 2-furylquinolines: ubiquitous scaffolds in common high-throughput screening libraries. *J. Med. Chem.* **58,** 7419–7430 (2015).
178. Taylor, R. & Kennard, O. Crystallographic evidence for the existence of CH···O, CH···N and CH···Cl hydrogen bonds. *J. Am. Chem. Soc.* **104,** 5063–5070 (1982).
179. Westler, W. M., Lin, I.-J., Perczel, A., Weinhold, F. & Markley, J. L. Hyperfine-shifted 13c resonance assignments in an iron–sulfur protein with quantum chemical verification: aliphatic C–H···S 3-center–4-electron interactions. *J. Am. Chem. Soc.* **133,** 1310–1316 (2011).
180. Domagała, M. & Grabowski, S. J. CH···N and CH···S Hydrogen BondsInfluence of Hybridization on Their Strength. *J. Phys. Chem. A* **109,** 5683–5688 (2005).

APPENDIX A

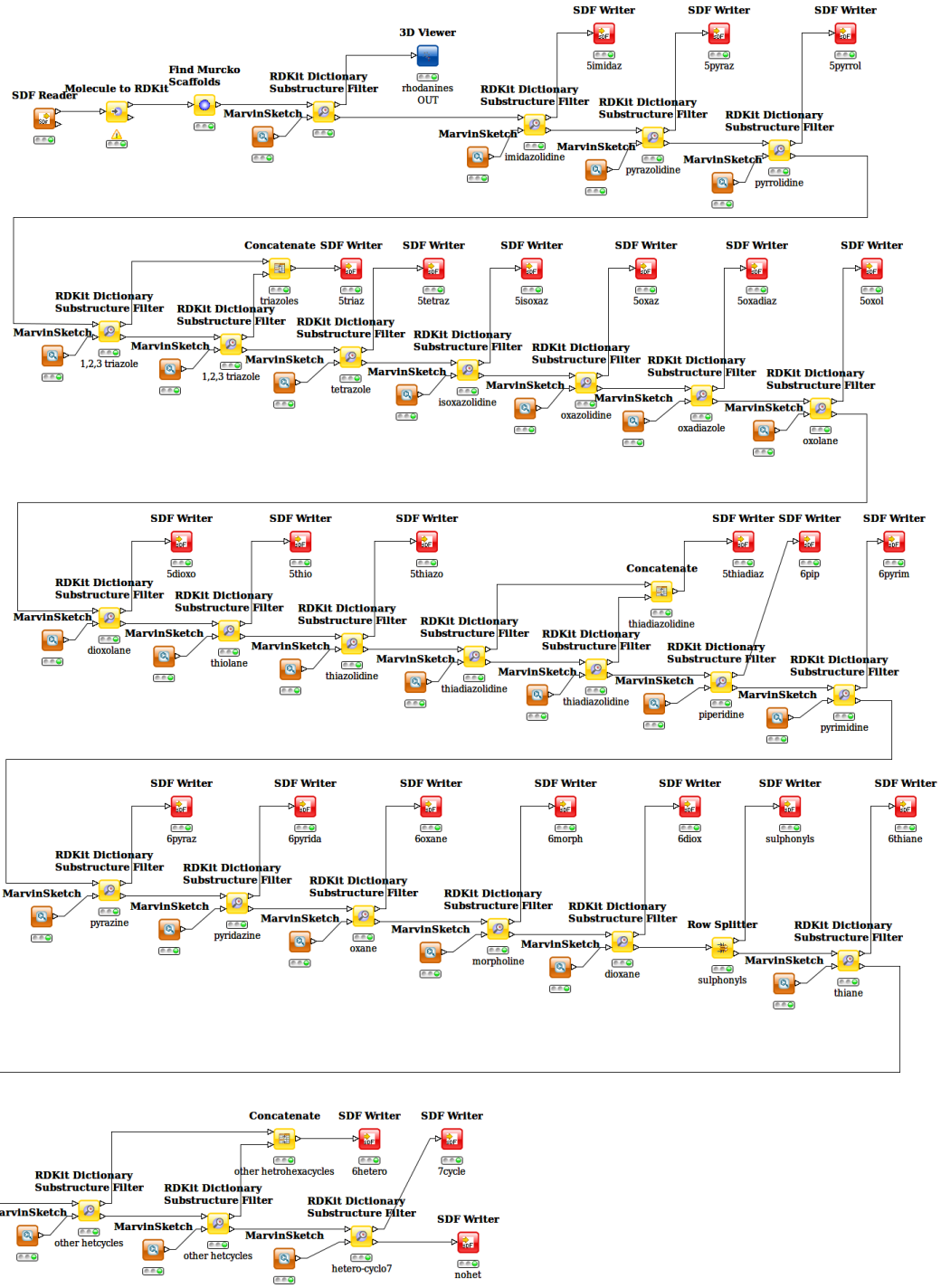
A1 Annotated SDF (Structure Data File), Example: Testosterone




```
  Testosterone

46 49  0  0  0  0  0  0  0  0999 V2000
32.7221  -48.2471   -0.4674 C   0  0  0  0  0  0  0  0  0  0  0  0
31.8304  -48.9551   -0.8864 O   0  0  0  0  0  0  0  0  0  0  0  0
[…]
36.3538  -45.8781   -0.0541 C   0  0  2  0  0  0
34.9920  -46.5074    0.3418 C   0  0  1  0  0  0
34.0245  -49.8763    0.0985 H   0  0  0  0  0  0
36.1685  -48.5912    2.1283 H   0  0  0  0  0  0
[…]
33.7636  -44.9408   -0.5628 H   0  0  0  0  0  0  0  0  0  0  0  0
34.7216  -46.1534    1.3716 H   0  0  0  0  0  0  0  0  0  0  0  0
44  4  1  0  0  0  0
38 17  1  0  0  0  0
[…]
16 34  1  0  0  0  0
16 35  1  0  0  0  0
M  END
> <Lipinski's Rule of Five>
0
> <XLogP>
3.2229999999999994
> <VABC Volume Descriptor>
282.7658831627233
> <Rotatable Bonds Count>
2
> <Topological Polar Surface Area>
37.3
$$$$
```

A2 KNIME Clustering Pipeline

Figure A1: The clustering pipeline created in KNIME for the 1.3 million compounds derived from initial pre-filtering. Orange nodes are SDF Reader input, red nodes are SDF Write output, and yellow nodes are filters or concatenating nodes.



A3 Virtual Screening CONF File

```
  GOLD CONFIGURATION FILE

  AUTOMATIC SETTINGS
autoscale = 0.2

  POPULATION
popsiz = auto
select_pressure = auto
n_islands = auto
maxops = auto
niche_siz = auto

  GENETIC OPERATORS
pt_crosswt = auto
allele_mutatewt = auto
migratewt = auto

  FLOOD FILL
radius = 10
origin = 32.5127 73.6746 11.1142
do_cavity = 1
floodfill_atom_no = 0
cavity_file =
floodfill_center = point

  DATA FILES
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5imidaz.sdf
5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5oxadiaz.sdf
5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5oxaz.sdf 5
ligand_data_file
/home/nataliet/Document                                    niset/5oxodiox.sdf
5
ligand_data_file
/home/nataliet/Document                                    niset/5pyraz.sdf 5
ligand_data_file
/home/nataliet/Document                                    niset/5pyrrol.sdf
5
ligand_data_file
/home/nataliet/Document                                    niset/5tetraz.sdf
5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5thiadiaz.sd
f 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5thiaz.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5thio.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/5triaz.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6diox.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6morph.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6oxane.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6pip.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6pyraz.sdf 5
```

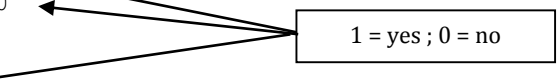
Search efficiency: 20%

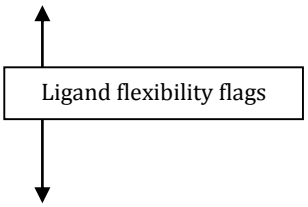
Genetic algorithm settings

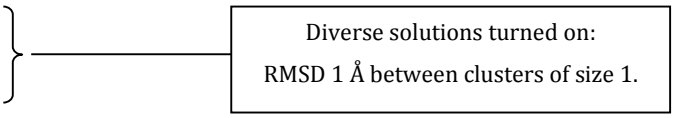
Centre point with *x, y, z*, radius and the option to produce a cavity structure file

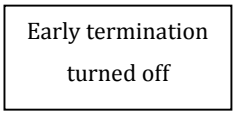
Source of input ligand structures, utilising complete file destination including all directories.
Number denotes how many GA runs to perform, and therefore how many poses to generate per ligand.

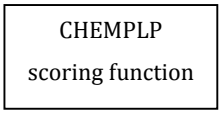
```
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6pyrida.sdf
5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6pyrim.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/6tria.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/7cycle.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/other.sdf 5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/otherhet.sdf
5
ligand_data_file
/home/nataliet/Documents/ZINC_180912_DrugsNow/CLUSTERED/miniset/sulphonyls.s
df 5
param_file = DEFAULT
set_ligand_atom_types = 1
set_protein_atom_types = 0
directory = minidock
tordist_file = DEFAULT
make_subdirs = 1
save_lone_pairs = 1
fit_points_file = fit_pts.mol2
read_fitpts = 0

  FLAGS
internal_ligand_h_bonds = 1
flip_free_corners = 1
match_ring_templates = 1
flip_amide_bonds = 1
flip_planar_n = 1 flip_ring_NRR flip_ring_NHR
flip_pyramidal_n = 1
rotate_carboxylic_oh = flip
use_tordist = 1
postprocess_bonds = 1
rotatable_bond_override_file = DEFAULT
diverse_solutions = 1
divsol_rmsd = 1
divsol_cluster_size = 1
solvate_all = 1

  TERMINATION
early_termination = 0
n_top_solutions = 3
rms_tolerance = 1.5

  CONSTRAINTS
force_constraints = 0

  COVALENT BONDING
covalent = 0

  SAVE OPTIONS
save_score_in_file = 1
save_protein_torsions = 1
output_file_format = MACCS

  FITNESS FUNCTION SETTINGS
initial_virtual_pt_match_max = 3
relative_ligand_energy = 1
gold_fitfunc_path = plp
score_param_file = DEFAULT

  PROTEIN DATA
protein_datafile = NJTflipped_1U9N.mol2
```

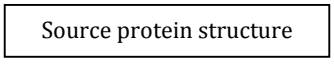












A4 Script for Removal of Problematic Punctuation in SDF Property Headers

In attempting to open a GoldMine of a test set, it was determined with the help of Dr. Richard Sykes (CCDC) that hyphens and parentheses were preventing conversion of SDF headers into the SQLite database fields. Dr. Sykes wrote the following python script to remove the offending characters. It additionally removed any corrupted files.

```
import sys
import os
import string
import shutil

d = 'minidock'
sp = set(string.printable)
for path, dirs, files in os.walk(d):
    for f in files:
        if f.endswith('.sdf'):
            fname = os.path.join(path, f)
            txt = open(fname).read()
            txt = txt.replace('H-bond', 'H_bond')
            txt = txt.replace('(mercaptyl)', 'mercaptyl')
            f = open(fname, 'w')
            f.write(txt)
            f.close()
            sf = set(txt)
            if sf - sp:
                print 'BAD FILE', fname
                shutil.move(fname, fname + '.BAD')
```

A5 Python Script for Filtering top 10% Screening Output by Score

Over 2 million poses had been generated by the virtual screening protocol. With the help of Dr. Tjelvar Olsson, we crafted a script which would access the file which denotes all poses for each ligand, pull out the name of the highest scoring ligand, and if that score placed the ligand in the highest 10% of ligands (ie. NUM_TO_KEEP), it would be written to a file for output.

```
import sys
import os
import os.path

DIR_NAME = "/home/guest2/not-backed-up/screen-1U9N.mol2-1U9N"
NUM_COMPOUNDS = 409195
NUM_TO_KEEP = NUM_COMPOUNDS / 10
best_poses = []

for i in range(NUM_COMPOUNDS):
    db_num = i + 1
    sub_dir = "database_m%i" % db_num
    tot_dir = os.path.join(DIR_NAME, sub_dir)
    try:
        for file_name in os.listdir(tot_dir):
            if file_name.startswith('database'):
               file_path = os.path.join(tot_dir, file_name)
               for i, line in enumerate(open(file_path, 'r')):
                   if i == 4:
                       words = line.split()
                       identifer = int(words[0].strip())
                       score = float(words[1].strip())
                       best_sol_name = 'gold_soln_database_m%i_%i.sdf' %
(db_num, identifer)
                       best_sol_path = os.path.join(tot_dir, best_sol_name)
                       best_poses.append( (score, best_sol_path) )
    except:
        pass

best_poses.sort()
best_poses.reverse()
for i, (score, file_path) in enumerate(best_poses):
    if i < NUM_TO_KEEP:
        file_handle = open(file_path, 'r')
        text = file_handle.read()
        print text,
```

409,195 ligands, each with 5 poses; keep the highest scoring 40,919

This section enters the ranking file to identify the pose number with the highest score

APPENDIX B

B1 Compound Nomenclature and Virtual Screening Score

Table B1: A working code beginning "NJT" was used when handling compounds to avoid long alphanumeric codes which could lead to confusion/mislabelling. This table denotes all compounds used for analysis. Note: NJT07 is not listed as it failed quality control at the supplier and could not be sourced elsewhere. Additionally, NJT25 arose from an ordering error and as such is not listed with other virtual screening results. Finally, NJT37 was completely insoluble, and so no assays were performed with it.

ZINC REFERENCE	NJT REFERENCE	SCORE RANK	COMPANY	COMPANY REFERENCE	RAW SCORE
ZINC12201617	NJT01	22	ChemBridge	86370596	89.1069
ZINC00616109	NJT02	19	ChemBridge	7778963	89.2407
ZINC67692217	NJT03	29	ChemBridge	58238331	87.9500
ZINC65406277	NJT04	33	ChemBridge	50509137	87.2316
ZINC72158865	NJT05	38	ChemBridge	75389916	86.8409
ZINC65507786	NJT06	11	ChemBridge	89710978	91.3825
ZINC67973854	NJT08	18	ChemBridge	94793758	89.3759
ZINC67654656	NJT09	23	ChemBridge	26569393	88.5067
ZINC67974892	NJT10	45	ChemBridge	95652737	85.5077
ZINC71775561	NJT11	46	ChemBridge	96176937	85.4536
ZINC19952823	NJT12	47	ChemBridge	68375090	85.4176
ZINC72172695	NJT13	49	ChemBridge	97057671	85.2338
ZINC65526448	NJT14	55	ChemBridge	92870208	84.5794
ZINC00237604	NJT15	59	ChemBridge	5304371	84.1084
ZINC67955631	NJT16	72	ChemBridge	91757787	83.4029
ZINC32576321	NJT17	73	ChemBridge	64993287	83.3492
ZINC11817892	NJT18	76	ChemBridge	73552030	83.2528
ZINC67673539	NJT19	94	ChemBridge	27987503	81.5626
ZINC20109834	NJT20	95	ChemBridge	69520354	81.5585
ZINC67975203	NJT21	24	ChemBridge	95916741	88.4118
ZINC00132569	NJT22	9	ChemBridge	7618548	92.8909
ZINC67820979	NJT23	12	ChemBridge	72213543	91.2764
ZINC00714237	NJT24	56	ChemBridge	5652915	84.4833
ZINC01216083	NJT26	80	ChemBridge	6088631	82.9549
ZINC02438552	NJT27	6	ChemBridge	9116022	95.2030
ZINC01226689	NJT28	7	ChemBridge	5249858	93.3650
ZINC04828017	NJT29	10	ChemBridge	7962144	92.2616
ZINC20137454	NJT30	43	ChemBridge	9228506	85.8673
ZINC24615603	NJT31	44	ChemBridge	9283033	85.6908
ZINC00425482	NJT32	67	ChemBridge	9331185	83.7565
ZINC09236877	NJT33	91	ChemBridge	6985499	81.7739
ZINC17076365	NJT34	25	Vitas M	STK470583	88.3717
ZINC13758755	NJT35	52	Vitas M	STL050026	84.9144
ZINC71775333	NJT36	58	Vitas M	STL167978	84.1745

ZINC00924619	NJT37	60	Vitas M	STK011909	84.1005
ZINC69462812	NJT38	3	Enamine	Z1025693264	96.1781
ZINC40151811	NJT39	20	Enamine	Z596096640	89.2407
ZINC58327836	NJT40	21	Enamine	Z1101450797	89.2192
ZINC52487551	NJT41	54	Enamine	Z646440516	84.7711
ZINC53275627	NJT42	62	Enamine	Z167023036	83.9472
ZINC46957685	NJT43	51	Enamine	Z422902610	85.0971
ZINC44936750	NJT44	34	Enamine	Z415248090	87.0520
ZINC71839930	NJT45	27	Enamine	Z908777508	88.3083
ZINC69461964	NJT46	42	Enamine	Z1085724378	85.9954
ZINC05187134	NJT47	28	Enamine	Z54071047	88.0150
ZINC69489717	NJT48	85	Enamine	Z1139229987	82.3846
ZINC12876413	NJT49	90	Enamine	Z102922270	81.8401
ZINC12794882	NJT50	71	Enamine	Z119631558	83.5383
ZINC22831673	NJT51	31	Enamine	Z90609697	87.4262
ZINC65575010	NJT52	70	Enamine	Z558478778	83.5945
ZINC12609641	NJT53	57	Enamine	Z279903378	84.4802
ZINC14168858	NJT54	26	Enamine	Z24677903	88.3478
ZINC10777920	NJT55	2	Enamine	Z226510926	97.5458
ZINC08987568	NJT56	68	Enamine	Z224851022	83.6657
ZINC09689958	NJT57	86	Enamine	Z30508852	82.2755
ZINC71795591	NJT58	30	Enamine	Z281523990	87.9420
ZINC71794585	NJT59	16	Enamine	Z257202932	90.7126
ZINC14848503	NJT60	35	Enamine	Z237505970	86.9967
ZINC22281715	NJT61	89	Enamine	Z225731946	81.8996
ZINC45517734	NJT62	36	Enamine	Z351674606	86.9565
ZINC09725714	NJT63	37	Enamine	Z197481400	86.8884
ZINC69416783	NJT64	74	Enamine	Z1001812522	83.3350
ZINC24418394	NJT65	32	Enamine	Z153679482	87.2489
ZINC25187332	NJT66	93	Enamine	Z25094450	81.5864
ZINC23053302	NJT67	81	Enamine	Z281178970	82.8968
ZINC65490004	NJT68	69	Enamine	Z225720568	83.6479
ZINC69669548	NJT69	87	Enamine	Z1139749166	82.0951
ZINC69381663	NJT70	78	Enamine	Z324839670	82.9982
ZINC08706459	NJT71	5	Enamine	Z226452786	95.2065
ZINC03420970	NJT72	8	Enamine	Z24107796	93.0210
ZINC58441335	NJT73	82	Enamine	Z649796534	82.5989
ZINC08980600	NJT74	39	Enamine	Z226124130	86.5464
ZINC58145475	NJT75	63	Enamine	Z846030592	83.9063
ZINC18962750	NJT76	61	Enamine	Z98973900	84.0079
ZINC03905710	NJT77	13	Ambinter	Amb16755507	91.1945
ZINC04370962	NJT78	92	Ambinter	Amb20448045	81.7662
ZINC33004106	NJT79	84	Ambinter	Amb16657739	82.4998
ZINC01502258	NJT80	40	Ambinter	Amb16766591	86.3118
ZINC00499610	NJT81	77	Ambinter	Amb19773311	83.1186
ZINC03903089	NJT82	50	Ambinter	Amb6875571	85.1973

ZINC04870694	NJT83	66	Ambinter	Amb5371151	83.8413
ZINC09087663	NJT84	79	Ambinter	Amb5367264	82.9583
ZINC19234873	NJT85	1	Ambinter	Amb13914491	99.5423

B2 Thermal Shift Raw Data

Table B2: The raw melting temperature data for all tested compounds.

COMPOUND	CONCENTRATION / µM	1	2	3	AVERAGE	SHIFT / °C
NJT01	400	59.4	59.2	59.5	59.37	0.119
NJT01	320	59.2	59.3	59.7	59.40	0.152
NJT01	160	59.5	59.7	59.7	59.63	0.386
NJT01	80	59.6	58.8	59.5	59.30	0.052
NJT02	400	59.4	59.3	59.5	59.40	0.152
NJT02	320	59.3	59.3	59.6	59.40	0.152
NJT02	160	59.5	60.0	59.5	59.67	0.419
NJT02	80	59.6	59.8	59.6	59.67	0.419
NJT03	400	60.5	60.6	60.8	60.63	1.386
NJT03	320	60.6	60.3	60.8	60.57	1.319
NJT03	160	60.1	61.3	60.2	60.53	1.286
NJT03	80	60.2	60.4	59.4	60.00	0.752
NJT04	400	59.8	60.1	60.3	60.07	0.819
NJT04	320	60.0	60.1	60.0	60.03	0.786
NJT04	160	59.9	61.9	60.0	59.95	0.702
NJT04	80	59.6	60.0	59.3	59.63	0.386
NJT05	400	59.2	59.5	59.5	59.40	0.152
NJT05	320	59.4	59.3	59.5	59.40	0.152
NJT05	160	59.4	59.1	59.6	59.37	0.119
NJT05	80	59.5	59.9	58.9	59.43	0.186
NJT06	400	59.0	59.8	59.7	59.50	0.252
NJT06	320	42.1	60.6	59.8	60.20	0.952
NJT06	160	59.6	59.9	59.5	59.67	0.419
NJT06	80	59.5	59.8	59.3	59.53	0.286
NJT07	400	59.3	59.7	59.6	59.53	0.286
NJT07	320	59.2	59.8	59.6	59.53	0.286
NJT07	160	59.3	60.1	59.4	59.60	0.352
NJT07	80	59.5	59.8	59.4	59.57	0.319
NJT08	400	59.3	59.5	59.1	59.30	0.052
NJT08	320	59.3	59.5	59.7	59.50	0.252
NJT08	160	59.3	59.8	59.5	59.53	0.286
NJT08	80	59.6	59.5	59.5	59.53	0.286
NJT09	400	59.2	59.2	59.3	59.23	-0.014

NJT09	320	59.2	59.1	59.3	59.20	-0.048
NJT09	160	31.0	59.7	59.4	59.55	0.302
NJT09	80	59.6	59.7	59.2	59.50	0.252
NJT10	400	61.1	61.1	61.5	61.23	1.986
NJT10	320	60.8	61.0	60.9	60.90	1.652
NJT10	160	60.4	60.3	60.5	60.40	1.152
NJT10	80	60.3	60.1	59.9	60.10	0.852
NJT11	400	59.2	59.2	59.3	59.23	-0.014
NJT11	320	59.2	59.3	59.3	59.27	0.019
NJT11	160	59.5	58.4	59.3	59.07	-0.181
NJT11	80	59.7	59.2	59.2	59.37	0.119
NJT12	400	59.1	59.1	59.2	59.13	-0.114
NJT12	320	59.3	59.3	59.5	59.37	0.119
NJT12	160	59.3	59.4	59.2	59.30	0.052
NJT12	80	59.5	59.4	59.3	59.40	0.152
NJT13	400	82.4	59.3	59.6	59.45	0.202
NJT13	320	59.4	59.3	59.8	59.50	0.252
NJT13	160	59.6	59.4	59.7	59.57	0.319
NJT13	80	58.8	58.4	59.4	58.87	-0.381
NJT14	400	59.6	59.2	59.4	59.40	0.152
NJT14	320	59.3	59.2	59.4	59.30	0.052
NJT14	160	59.0	59.3	59.7	59.33	0.086
NJT14	80	59.5	59.3	60.7	59.83	0.586
NJT15	400	62.9	63.4	61.9	62.73	3.486
NJT15	320	62.3	62.6	62.1	62.33	3.086
NJT15	160	61.3	62.3	61.7	61.77	2.519
NJT15	80	61.6	61.0	61.3	61.30	2.052
NJT16	400	59.3	59.2	59.2	59.23	-0.014
NJT16	320	59.2	59.3	59.7	59.40	0.152
NJT16	160	59.2	59.3	59.1	59.20	-0.048
NJT16	80	58.7	58.8	59.4	58.97	-0.281
NJT17	400	60.0	59.7	59.9	59.87	0.619
NJT17	320	59.5	59.7	59.9	59.70	0.452
NJT17	160	59.2	59.5	59.7	59.47	0.219
NJT17	80	59.6	59.2	59.5	59.43	0.186
NJT18	400	59.5	59.3	59.5	59.43	0.186

NJT18	320	59.2	59.2	59.0	59.13	-0.114
NJT18	160	59.5	59.4	59.4	59.43	0.186
NJT18	80	59.3	59.4	59.5	59.40	0.152
NJT19	400	59.0	59.1	59.0	59.03	-0.214
NJT19	320	59.2	59.2	59.0	59.13	-0.114
NJT19	160	59.3	59.2	59.3	59.27	0.019
NJT19	80	59.5	59.5	59.4	59.47	0.219
NJT20	400	59.8	59.6	59.5	59.63	0.386
NJT20	320	59.7	59.3	59.5	59.50	0.252
NJT20	160	59.2	59.5	59.6	59.43	0.186
NJT20	80	59.6	59.3	59.7	59.53	0.286
NJT21	400	59.5	59.2	59.2	59.30	0.052
NJT21	320	59.5	59.3	59.2	59.33	0.086
NJT21	160	59.2	59.4	59.4	59.33	0.086
NJT21	80	59.4	59.0	59.4	59.27	0.019
NJT22	400	59.6	59.3	59.2	59.37	0.119
NJT22	320	59.2	59.2	59.3	59.23	-0.014
NJT22	160	59.5	59.3	59.4	59.40	0.152
NJT22	80	59.6	59.2	59.5	59.43	0.186
NJT23	400	59.8	59.2	59.1	59.37	0.119
NJT23	320	59.6	59.2	59.2	59.33	0.086
NJT23	160	59.5	59.3	59.3	59.37	0.119
NJT23	80	59.7	59.5	59.4	59.53	0.286
NJT24	400	59.3	58.9	59.4	59.20	-0.048
NJT24	320	59.5	59.5	59.4	59.47	0.219
NJT24	160	59.5	59.2	59.4	59.37	0.119
NJT24	80	59.5	59.2	59.4	59.37	0.119
NJT25	400	68.7	68.0	68.6	68.43	9.186
NJT25	320	68.6	68.3	67.9	68.27	9.019
NJT25	160	67.0	67.4	67.2	67.20	7.952
NJT25	80	66.0	66.3	66.3	66.20	6.952
NJT26	400	59.5	60.0	59.7	59.73	0.486
NJT26	320	59.5	60.0	59.8	59.77	0.519
NJT26	160	60.8	60.2	59.9	60.30	1.052
NJT26	80	59.5	59.3	59.2	59.33	0.086
NJT27	400	89.7	59.3	60.0	59.65	0.402

NJT27	320	59.7	59.5	59.9	59.70	0.452
NJT27	160	60.0	59.7	59.6	59.77	0.519
NJT27	80	59.7	59.5	59.4	59.53	0.286
NJT28	400	59.5	58.8	35.1	59.15	-0.098
NJT28	320	59.3	34.7	27.0	59.30	0.052
NJT28	160	44.6	59.5	60.1	59.80	0.552
NJT28	80	59.7	59.7	59.3	59.57	0.319
NJT29	400	59.6	59.2	59.9	59.57	0.319
NJT29	320	59.6	59.9	59.5	59.67	0.419
NJT29	160	59.7	59.5	59.8	59.67	0.419
NJT29	80	59.8	59.3	59.3	59.47	0.219
NJT30	400	60.0	59.5	60.3	59.93	0.686
NJT30	320	59.9	59.7	60.2	59.93	0.686
NJT30	160	59.3	59.5	59.9	59.57	0.319
NJT30	80	59.9	59.7	59.5	59.70	0.452
NJT31	400	59.9	59.6	60.4	59.97	0.719
NJT31	320	89.1	59.5	60.0	59.75	0.502
NJT31	160	59.7	59.5	59.7	59.63	0.386
NJT31	80	59.8	59.7	59.5	59.67	0.419
NJT32	400	59.3	59.3	59.4	59.33	0.086
NJT32	320	59.4	59.0	59.5	59.30	0.052
NJT32	160	59.5	59.2	59.5	59.40	0.152
NJT32	80	59.8	59.4	59.4	59.53	0.286
NJT33	400	59.1	59.0	59.8	59.30	0.052
NJT33	320	59.4	58.9	59.7	59.33	0.086
NJT33	160	59.4	59.3	59.4	59.37	0.119
NJT33	80	59.6	59.4	59.6	59.53	0.286
NJT34	400	59.4	58.6	59.5	59.17	-0.081
NJT34	320	59.4	57.2	59.5	59.45	0.202
NJT34	160	59.4	59.2	59.5	59.37	0.119
NJT34	80	59.7	59.5	59.2	59.47	0.219
NJT35	400	59.4	59.0	59.5	59.30	0.052
NJT35	320	59.4	59.1	59.7	59.40	0.152
NJT35	160	89.4	59.3	59.5	59.40	0.152
NJT35	80	59.6	59.4	59.1	59.37	0.119
NJT36	400	89.4	59.1	59.5	59.30	0.052

NJT36	320	59.3	59.2	59.5	59.33	0.086
NJT36	160	59.5	59.4	59.7	59.53	0.286
NJT36	80	59.6	59.4	59.4	59.47	0.219
NJT38	400	77.8	59.0	58.8	58.90	-0.348
NJT38	320	60.0	59.2	59.2	59.47	0.219
NJT38	160	60.5	59.3	59.3	59.70	0.452
NJT38	80	60.3	59.5	59.8	59.87	0.619
NJT39	400	59.3	59.0	58.7	59.00	-0.248
NJT39	320	59.7	59.2	59.0	59.30	0.052
NJT39	160	60.0	59.3	59.1	59.47	0.219
NJT39	80	60.1	59.5	59.3	59.63	0.386
NJT40	400	59.9	59.5	58.8	59.40	0.152
NJT40	320	74.2	59.4	59.0	59.20	-0.048
NJT40	160	60.0	59.6	59.3	59.63	0.386
NJT40	80	60.1	59.6	59.3	59.67	0.419
NJT41	400	59.1	59.1	58.7	58.97	-0.281
NJT41	320	59.5	59.1	59.0	59.20	-0.048
NJT41	160	59.8	59.3	59.0	59.37	0.119
NJT41	80	59.8	59.2	59.3	59.43	0.186
NJT42	400	60.7	59.7	59.8	60.07	0.819
NJT42	320	60.6	59.8	59.7	60.03	0.786
NJT42	160	60.2	59.7	59.5	59.80	0.552
NJT42	80	60.1	59.6	59.5	59.73	0.486
NJT43	400	59.5	59.2	58.8	59.17	-0.081
NJT43	320	59.7	59.3	59.0	59.33	0.086
NJT43	160	60.0	59.4	59.4	59.60	0.352
NJT43	80	60.0	59.4	58.5	59.30	0.052
NJT44	400	59.5	59.3	58.9	59.23	-0.014
NJT44	320	59.5	59.4	58.9	59.27	0.019
NJT44	160	59.8	59.5	59.2	59.50	0.252
NJT44	80	59.8	59.6	59.2	59.53	0.286
NJT45	400	74.3	59.3	59.1	59.20	-0.048
NJT45	320	60.1	59.4	59.2	59.57	0.319
NJT45	160	59.7	59.5	59.2	59.47	0.219
NJT45	80	60.0	59.5	59.3	59.60	0.352
NJT46	400	59.6	59.0	58.8	59.13	-0.114

NJT46	320	59.7	59.2	58.9	59.27	0.019
NJT46	160	59.7	59.5	59.2	59.47	0.219
NJT46	80	59.8	59.5	59.2	59.50	0.252
NJT47	400	58.8	59.2	58.7	58.90	-0.348
NJT47	320	59.8	59.4	58.9	59.37	0.119
NJT47	160	60.0	59.4	59.2	59.53	0.286
NJT47	80	60.0	59.6	59.2	59.60	0.352
NJT48	400	60.6	59.4	59.6	59.87	0.619
NJT48	320	60.5	59.5	59.3	59.77	0.519
NJT48	160	60.3	59.5	59.3	59.70	0.452
NJT48	80	60.1	59.4	59.1	59.53	0.286
NJT49	400	73.4	59.6	59.6	59.60	0.352
NJT49	320	60.4	59.4	59.2	59.67	0.419
NJT49	160	60.3	59.3	59.2	59.60	0.352
NJT49	80	59.8	59.3	59.3	59.47	0.219
NJT50	400	59.8	60.1	59.6	59.83	0.586
NJT50	320	60.6	59.9	59.6	60.03	0.786
NJT50	160	60.4	59.9	59.5	59.93	0.686
NJT50	80	60.1	59.6	59.3	59.67	0.419
NJT51	400	59.8	59.7	59.3	59.60	0.352
NJT51	320	59.8	59.9	59.0	59.57	0.319
NJT51	160	59.9	59.8	59.4	59.70	0.452
NJT51	80	59.8	59.6	59.2	59.53	0.286
NJT52	400	59.8	59.0	59.0	59.27	0.019
NJT52	320	60.1	59.2	59.1	59.47	0.219
NJT52	160	59.8	59.3	59.2	59.43	0.186
NJT52	80	59.8	59.3	59.0	59.37	0.119
NJT53	400	59.5	59.2	58.9	59.20	-0.048
NJT53	320	59.8	59.3	58.9	59.33	0.086
NJT53	160	59.7	59.3	59.1	59.37	0.119
NJT53	80	59.8	59.4	59.2	59.47	0.219
NJT54	400	60.0	59.1	59.5	59.53	0.286
NJT54	320	60.3	59.2	59.2	59.57	0.319
NJT54	160	60.1	59.5	59.2	59.60	0.352
NJT54	80	60.3	59.3	59.2	59.60	0.352
NJT55	400	59.9	59.4	59.3	59.53	0.286

NJT55	320	60.2	59.3	59.3	59.60	0.352
NJT55	160	59.9	59.3	59.1	59.43	0.186
NJT55	80	60.2	59.4	59.1	59.57	0.319
NJT56	400	60.3	60.1	59.8	60.07	0.819
NJT56	320	60.2	59.7	59.6	59.83	0.586
NJT56	160	60.3	59.7	59.4	59.80	0.552
NJT56	80	60.0	59.6	59.0	59.53	0.286
NJT57	400	65.8	65.4	65.6	65.60	6.352
NJT57	320	64.4	64.5	65.0	64.63	5.386
NJT57	160	63.7	63.6	63.1	63.47	4.219
NJT57	80	61.5	61.6	61.0	61.37	2.119
NJT58	400	61.4	59.1	59.5	60.00	0.752
NJT58	320	61.0	59.0	59.5	59.83	0.586
NJT58	160	77.8	59.1	59.6	59.35	0.102
NJT58	80	75.4	58.9	58.7	58.80	-0.448
NJT59	400	59.1	58.4	59.0	58.83	-0.414
NJT59	320	59.2	58.4	59.4	59.00	-0.248
NJT59	160	59.0	58.9	59.4	59.10	-0.148
NJT59	80	59.2	60.2	58.9	59.43	0.186
NJT60	400	64.8	66.1	65.6	65.50	6.252
NJT60	320	63.9	65.2	60.4	63.17	3.919
NJT60	160	62.6	63.6	62.3	62.83	3.586
NJT60	80	61.9	61.1	60.9	61.30	2.052
NJT61	400	58.8	58.5	58.9	58.73	-0.514
NJT61	320	58.3	58.7	59.1	58.70	-0.548
NJT61	160	59.0	58.5	58.5	58.67	-0.581
NJT61	80	59.0	58.5	58.6	58.70	-0.548
NJT62	400	58.7	58.3	58.9	58.63	-0.614
NJT62	320	58.7	58.3	59.0	58.67	-0.581
NJT62	160	59.0	58.4	58.5	58.63	-0.614
NJT62	80	59.1	58.5	58.4	58.67	-0.581
NJT63	400	60.0	59.4	59.8	59.73	0.486
NJT63	320	59.9	59.0	59.9	59.60	0.352
NJT63	160	59.4	58.7	58.5	58.87	-0.381
NJT63	80	59.3	58.7	58.2	58.73	-0.514
NJT64	400	61.1	59.9	60.5	60.50	1.252

NJT64	320	61.0	59.2	60.3	60.17	0.919
NJT64	160	60.3	59.4	59.5	59.73	0.486
NJT64	80	59.8	59.6	58.9	59.43	0.186
NJT65	400	58.6	58.0	58.8	58.47	-0.781
NJT65	320	59.3	58.3	59.2	58.93	-0.314
NJT65	160	59.0	58.6	58.2	58.60	-0.648
NJT65	80	59.0	58.7	58.3	58.67	-0.581
NJT66	400	58.7	58.3	59.1	58.70	-0.548
NJT66	320	58.7	58.5	59.2	58.80	-0.448
NJT66	160	59.0	58.5	58.0	58.50	-0.748
NJT66	80	59.1	58.7	58.3	58.70	-0.548
NJT67	400	59.0	58.7	58.7	58.80	-0.448
NJT67	320	58.7	59.0	58.6	58.77	-0.481
NJT67	160	59.1	58.7	58.5	58.77	-0.481
NJT67	80	59.0	58.7	58.6	58.77	-0.481
NJT68	400	59.2	58.7	58.9	58.93	-0.314
NJT68	320	59.1	58.7	58.8	58.87	-0.381
NJT68	160	59.0	58.6	58.7	58.77	-0.481
NJT68	80	59.0	58.7	58.7	58.80	-0.448
NJT69	400	59.2	58.7	58.7	58.87	-0.381
NJT69	320	59.0	58.5	58.4	58.63	-0.614
NJT69	160	59.1	58.7	58.3	58.70	-0.548
NJT69	80	59.1	58.8	58.7	58.87	-0.381
NJT70	400	59.2	58.7	59.0	58.97	-0.281
NJT70	320	59.1	58.5	58.7	58.77	-0.481
NJT70	160	59.0	58.5	58.5	58.67	-0.581
NJT70	80	59.0	59.0	58.7	58.90	-0.348
NJT71	400	59.1	58.5	58.6	58.73	-0.514
NJT71	320	59.0	59.0	58.7	58.90	-0.348
NJT71	160	59.0	58.5	58.6	58.70	-0.548
NJT71	80	59.0	58.5	59.0	58.83	-0.414
NJT72	400	60.0	59.2	60.1	59.77	0.519
NJT72	320	59.8	59.1	59.9	59.60	0.352
NJT72	160	59.6	59.3	59.7	59.53	0.286
NJT72	80	59.3	58.8	59.3	59.13	-0.114
NJT73	400	59.2	59.0	58.7	58.97	-0.281

NJT73	320	59.2	58.8	58.5	58.83	-0.414
NJT73	160	59.3	58.8	58.6	58.90	-0.348
NJT73	80	59.0	58.7	58.6	58.77	-0.481
NJT74	400	60.8	59.5	58.5	59.60	0.352
NJT74	320	60.6	60.1	58.6	59.77	0.519
NJT74	160	60.3	60.3	58.4	59.67	0.419
NJT74	80	60.1	59.3	58.8	59.40	0.152
NJT75	400	59.2	59.3	58.1	58.87	-0.381
NJT75	320	59.2	59.2	58.0	58.80	-0.448
NJT75	160	59.2	60.0	58.3	59.17	-0.081
NJT75	80	59.1	59.1	58.3	58.83	-0.414
NJT76	400	59.2	59.5	58.1	58.93	-0.314
NJT76	320	59.0	58.8	58.1	58.63	-0.614
NJT76	160	59.0	58.6	58.3	58.63	-0.614
NJT76	80	59.0	58.6	58.2	58.60	-0.648
NJT77	400	57.9	57.8	58.7	58.13	-1.114
NJT77	320	57.9	59.2	59.2	58.77	-0.481
NJT77	160	58.3	59.5	58.1	58.63	-0.614
NJT77	80	58.5	59.6	58.6	58.90	-0.348
NJT78	400	57.4	57.8	59.0	58.07	-1.181
NJT78	320	57.9	59.5	59.0	58.80	-0.448
NJT78	160	58.2	59.3	58.1	58.53	-0.714
NJT78	80	58.1	59.4	58.2	58.57	-0.681
NJT79	400	58.4	58.3	58.7	58.47	-0.781
NJT79	320	58.4	60.1	58.7	59.07	-0.181
NJT79	160	58.3	59.4	58.5	58.73	-0.514
NJT79	80	57.9	58.8	58.3	58.33	-0.914
NJT80	400	62.5	62.6	63.0	62.70	3.452
NJT80	320	61.7	64.0	63.8	63.17	3.919
NJT80	160	60.5	63.4	62.0	61.97	2.719
NJT80	80	59.9	62.2	60.1	60.73	1.486
NJT81	400	58.0	58.5	57.7	58.07	-1.181
NJT81	320	57.9	58.3	59.2	58.47	-0.781
NJT81	160	58.2	58.3	59.4	58.63	-0.614
NJT81	80	57.9	58.1	59.6	58.53	-0.714
NJT82	400	57.8	58.2	57.4	57.80	-1.448

NJT82	320	58.0	58.1	59.6	58.57	-0.681
NJT82	160	57.4	58.3	58.6	58.10	-1.148
NJT82	80	58.1	58.5	59.5	58.70	-0.548
NJT83	400	59.9	58.2	59.2	59.10	-0.148
NJT83	320	59.6	61.7	59.3	60.20	0.952
NJT83	160	59.3	60.9	59.3	59.83	0.586
NJT83	80	58.8	60.1	59.5	59.47	0.219
NJT83	400	71.5	71.6	70.8	71.30	12.052
NJT83	320	71.3	72.2	71.2	71.57	12.319
NJT83	160	59.3	72.1	59.3	63.57	4.319
NJT83	80	58.8	60.1	59.5	59.47	0.219
NJT84	400	60.5	60.7	61.0	60.73	1.486
NJT84	320	60.2	60.2	60.4	60.27	1.019
NJT84	160	60.1	62.0	60.9	61.00	1.752
NJT84	80	58.9	61.9	61.0	60.60	1.352
NJT85	400	57.6	57.6	57.9	57.70	-1.548
NJT85	320	57.4	57.9	58.8	58.03	-1.214
NJT85	160	57.5	58.8	59.2	58.50	-0.748
NJT85	80	58.2	59.0	59.2	58.80	-0.448

B3 Tabulated Thermal Shift Data

Table B3: Temperature shift data is derived from the average readings calculated for each compound at each concentration, as displayed in table B2, as well as the average EthR melting temperature of 59.25°C derived from 42 negative control readings.

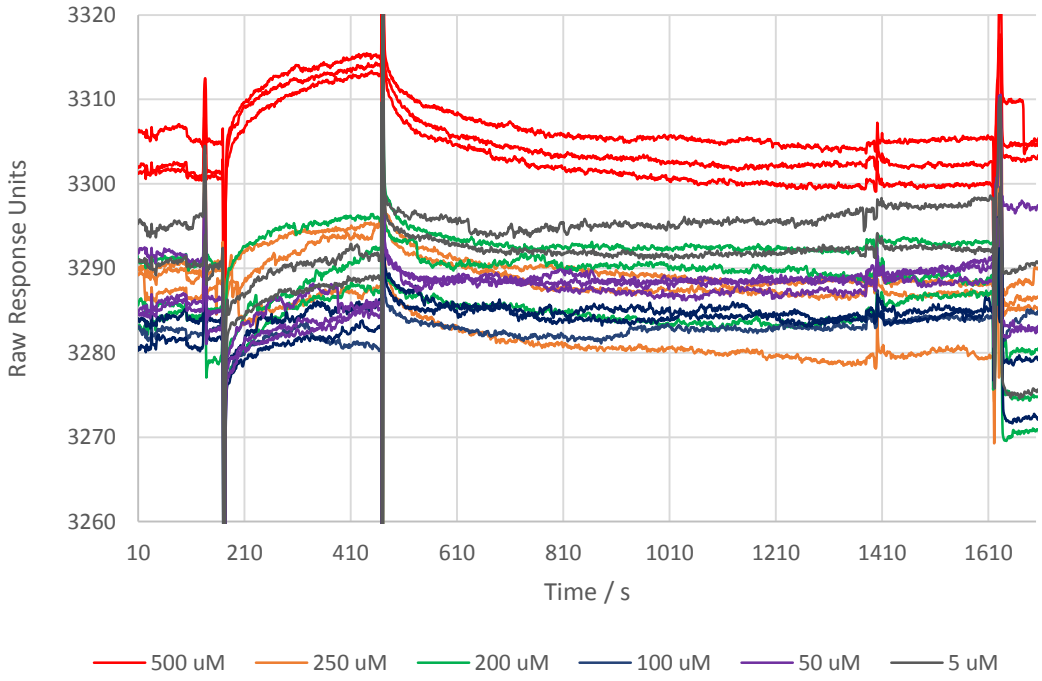
COMPOUND	400 µM	320 µM	160 µM	80 µM
NJT01	0.119	0.152	0.386	0.052
NJT02	0.152	0.152	0.419	0.419
NJT03	1.386	1.319	1.286	0.752
NJT04	0.819	0.786	0.702	0.386
NJT05	0.152	0.152	0.119	0.186
NJT06	0.252	0.952	0.419	0.286
NJT08	0.052	0.252	0.286	0.286
NJT09	-0.014	-0.048	0.302	0.252
NJT10	1.986	1.652	1.152	0.852
NJT11	-0.014	0.019	-0.181	0.119
NJT12	-0.114	0.119	0.052	0.152
NJT13	0.202	0.252	0.319	-0.381
NJT14	0.152	0.052	0.086	0.586
NJT15	3.486	3.086	2.519	2.052
NJT16	-0.014	0.152	-0.048	-0.281
NJT17	0.619	0.452	0.219	0.186
NJT18	0.186	-0.114	0.186	0.152
NJT19	-0.214	-0.114	0.019	0.219
NJT20	0.386	0.252	0.186	0.286
NJT21	0.052	0.086	0.086	0.019
NJT22	0.119	-0.014	0.152	0.186
NJT23	0.119	0.086	0.119	0.286
NJT24	-0.048	0.219	0.119	0.119
NJT25	9.186	9.019	7.952	6.952
NJT26	0.486	0.519	1.052	0.086
NJT27	0.402	0.452	0.519	0.286
NJT28	-0.098	0.052	0.552	0.319
NJT29	0.319	0.419	0.419	0.219
NJT30	0.686	0.686	0.319	0.452
NJT31	0.719	0.502	0.386	0.419
NJT32	0.086	0.052	0.152	0.286
NJT33	0.052	0.086	0.119	0.286
NJT34	-0.081	0.202	0.119	0.219
NJT35	0.052	0.152	0.152	0.119
NJT36	0.052	0.086	0.286	0.219
NJT38	-0.347	0.220	0.453	0.620
NJT39	-0.247	0.053	0.220	0.386
NJT40	0.153	-0.047	0.386	0.420
NJT41	-0.280	-0.047	0.120	0.186
NJT42	0.820	0.786	0.553	0.486

NJT43	-0.080	0.086	0.353	0.053
NJT44	-0.014	0.020	0.253	0.286
NJT45	-0.047	0.320	0.220	0.353
NJT46	-0.114	0.020	0.220	0.253
NJT47	-0.347	0.120	0.286	0.353
NJT48	0.620	0.520	0.453	0.286
NJT49	0.353	0.420	0.353	0.220
NJT50	0.586	0.786	0.686	0.420
NJT51	0.353	0.320	0.453	0.053
NJT52	0.020	0.220	0.186	0.120
NJT53	-0.047	0.086	0.120	0.220
NJT54	0.286	0.320	0.353	0.353
NJT55	0.286	0.353	0.186	0.320
NJT56	0.820	0.586	0.553	0.286
NJT57	6.353	5.386	4.220	2.120
NJT58	0.753	0.586	0.103	-0.447
NJT59	-0.414	-0.247	-0.147	0.186
NJT60	6.253	3.920	3.586	2.053
NJT61	-0.514	-0.547	-0.580	-0.547
NJT62	-0.614	-0.580	-0.614	-0.580
NJT63	0.486	0.353	-0.380	-0.514
NJT64	1.253	0.920	0.486	0.186
NJT65	-0.780	-0.314	-0.647	-0.580
NJT66	-0.547	-0.447	-0.747	-0.547
NJT67	-0.447	-0.480	-0.480	-0.480
NJT68	-0.314	-0.380	-0.480	-0.447
NJT69	-0.380	-0.614	-0.547	-0.380
NJT70	-0.280	-0.480	-0.580	-0.347
NJT71	-0.514	-0.347	-0.547	-0.414
NJT72	0.520	0.353	0.286	-0.114
NJT73	-0.280	-0.414	-0.347	-0.480
NJT74	0.353	0.520	0.420	0.153
NJT75	-0.380	-0.447	-0.080	-0.414
NJT76	-0.314	-0.614	-0.614	-0.647
NJT77	-1.114	-0.481	-0.614	-0.348
NJT78	-1.181	-0.448	-0.714	-0.681
NJT79	-0.781	-0.181	-0.514	-0.914
NJT80	3.452	3.919	2.719	1.486
NJT81	-1.181	-0.781	-0.614	-0.714
NJT82	-1.448	-0.681	-1.148	-0.548
NJT83 (1)	-0.148	0.952	0.586	0.219
NJT83 (2)	12.052	12.319	4.319	0.219
NJT84	1.486	1.019	1.752	1.352
NJT85	-1.548	-1.214	-0.748	-0.448

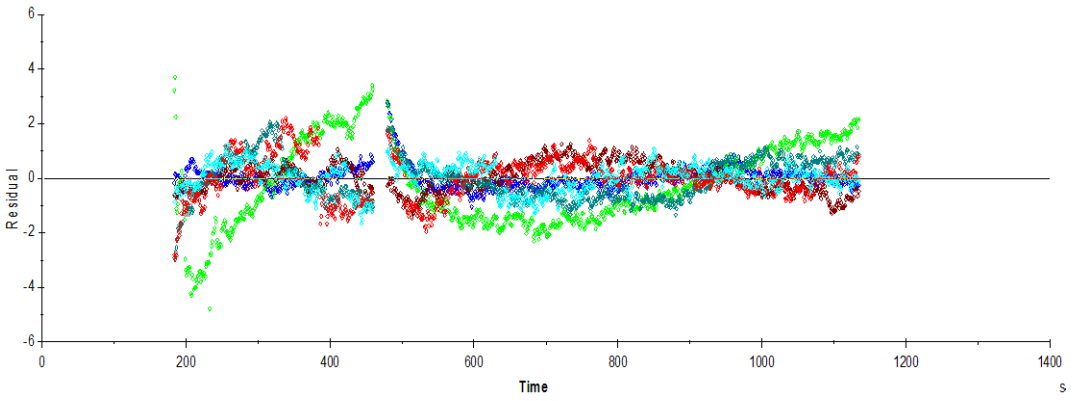
APPENDIX C

This appendix presents the raw SPR data from pilot experiments, with fitting data and residual plots for each. All sensorgrams show injection phases and dissociation phases (in buffer).

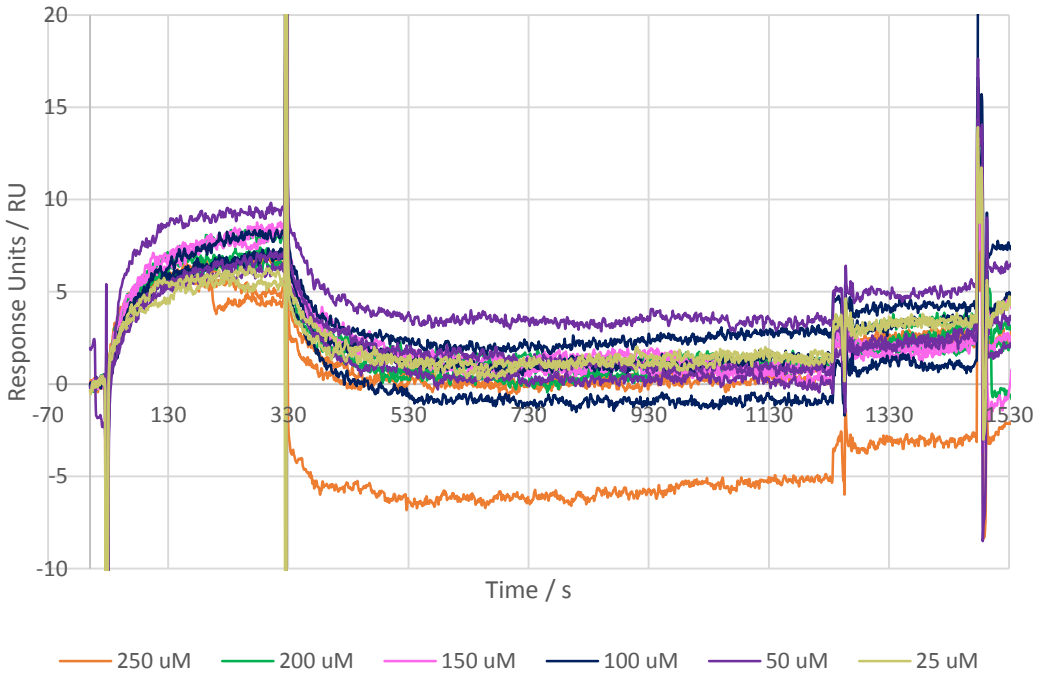
C1 Raw SPR Data: Compound 3



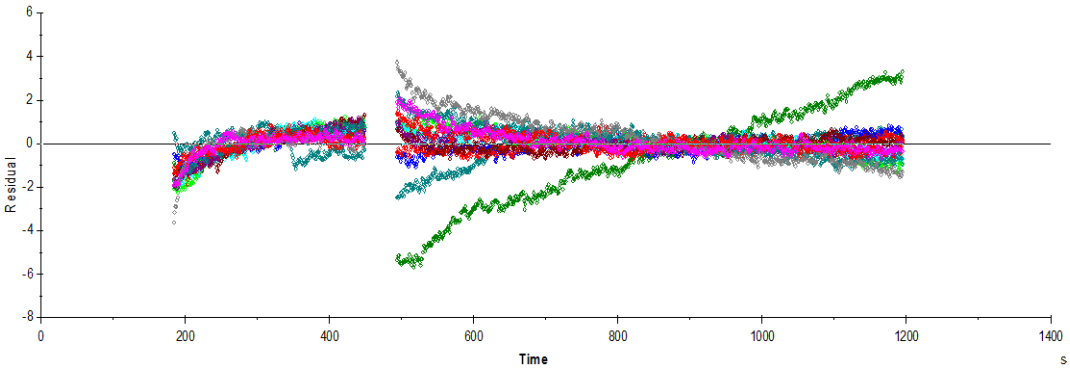
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
40.8	4.31E-03					9.47E+03	1.06E-04			0.725
		12.2	3.59E+00	2.40E-04	2.50E-04			8.59	0.0145	
		0.0498	-1.68	-0.0122	2.00E-04			0.0326	1.25E-02	
		15.5	-1.23	4.75E-03	1.50E-04			9.07	0.0104	
		2.73	-2.14	-8.36E-04	1.00E-04			1.33	8.39E-03	
		12.6	-3.08	4.19E-04	5.00E-05			4.05	6.35E-03	
		158	-3.86	5.81E-03	5.00E-06			7.13	4.51E-03	



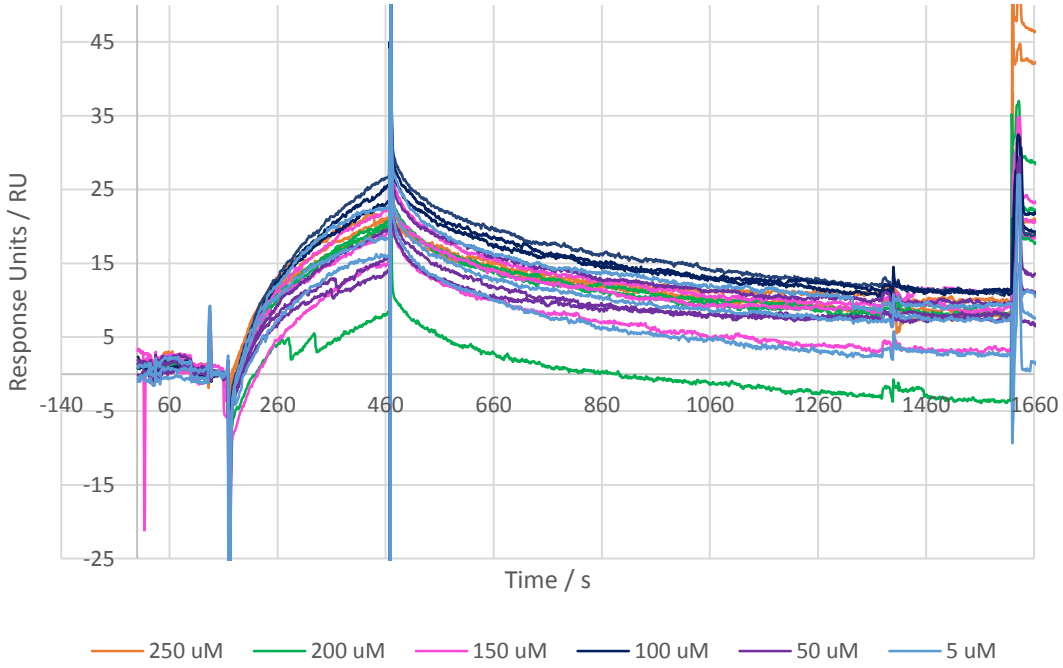
C2 Raw SPR Data: Compound 4



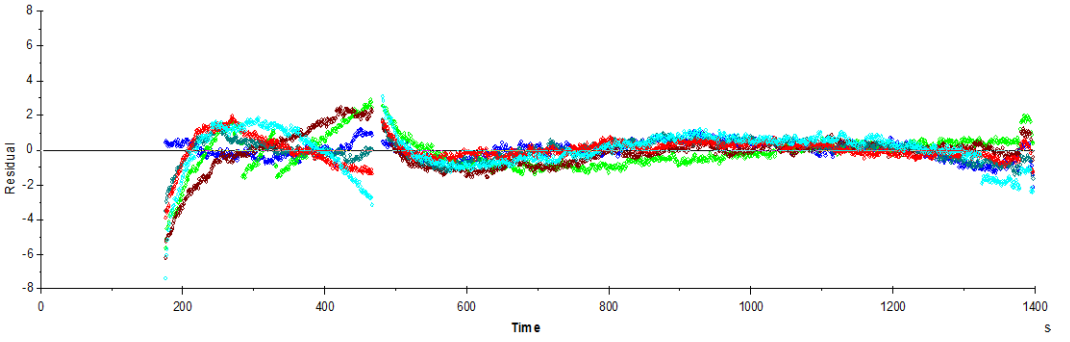
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
187	0.0132	6.91				1.41E+04	7.09E-05			0.69
			1.22	-8.35E-03	2.50E-04			5.38	0.0599	
			-0.444	-3.06E-04	2.50E-04			5.38	0.0599	
			0.768	4.57E-04	2.00E-04			5.1	0.0506	
			1.13	1.28E-03	2.00E-04			5.1	0.0506	
			2.3	1.92E-03	1.50E-04			4.69	0.0412	
			1.99	1.09E-03	1.50E-04			4.69	0.0412	
			2.27	-1.22E-03	1.00E-04			4.04	0.0319	
			1.77	1.40E-03	1.00E-04			4.04	0.0319	
			2.82	1.71E-03	5.00E-05			2.86	0.0226	
			5.18	4.74E-03	5.00E-05			2.86	0.0226	
			3.11	1.42E-03	2.50E-05			1.8	0.0179	
			3.57	1.81E-03	2.50E-05			1.8	0.0179	



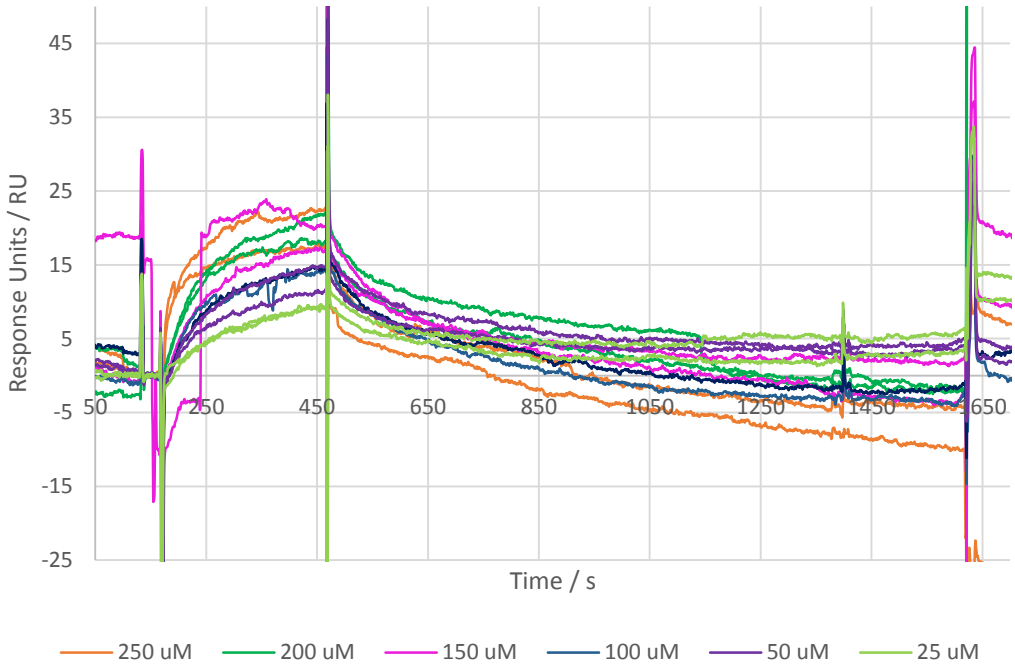
C3 Raw SPR Data: Compound 10



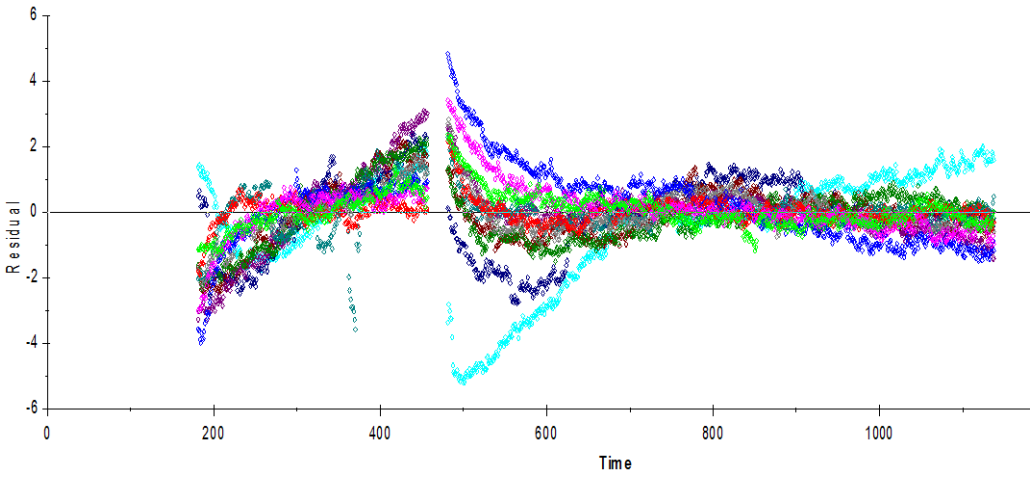
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
33.1	2.37E-03					1.39E+04	7.17E-05			0.708
		23.8	-0.539	6.30E-03	2.50E-04			18.5	0.0106	
		12.5	-1.88	-3.79E-03	2.00E-04			9.19	8.99E-03	
		27.3	-3.58	1.37E-03	1.50E-04			18.4	7.33E-03	
		49.8	-0.145	7.54E-03	1.00E-04			29	5.68E-03	
		70.7	-1.23	6.82E-03	5.00E-05			29	4.03E-03	
		619	2.23	6.79E-03	5.00E-06			40.3	2.54E-03	



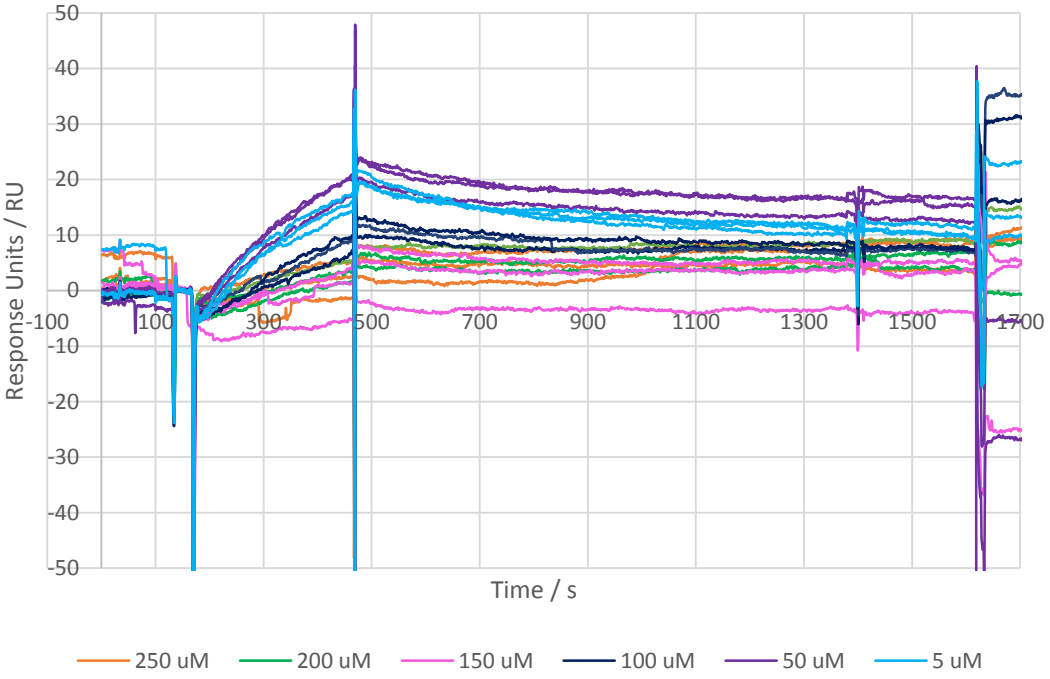
C4 Raw SPR Data: Compound 17



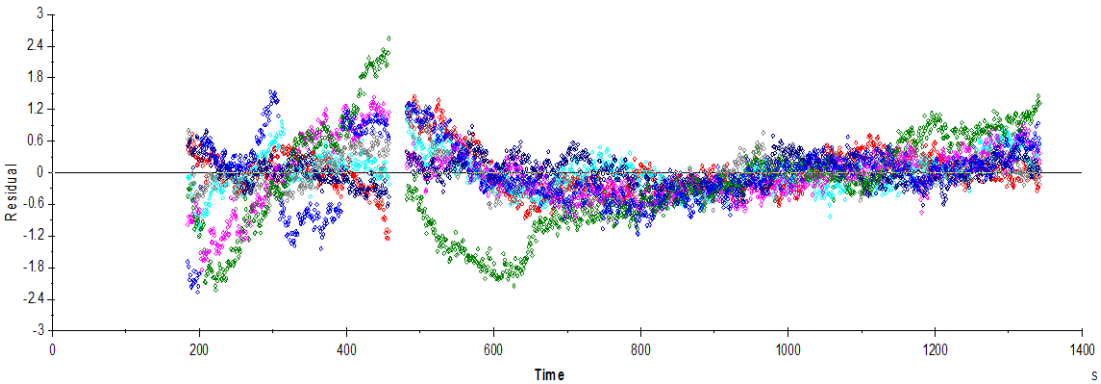
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
187	0.0132	6.91				1.41E+04	7.09E-05			0.69
			1.22	-8.35E-03	2.50E-04			5.38	0.0599	
			-0.444	-3.06E-04	2.50E-04			5.38	0.0599	
			0.768	4.57E-04	2.00E-04			5.1	0.0506	
			1.13	1.28E-03	2.00E-04			5.1	0.0506	
			2.3	1.92E-03	1.50E-04			4.69	0.0412	
			1.99	1.09E-03	1.50E-04			4.69	0.0412	
			2.27	-1.22E-03	1.00E-04			4.04	0.0319	
			1.77	1.40E-03	1.00E-04			4.04	0.0319	
			2.82	1.71E-03	5.00E-05			2.86	0.0226	
			5.18	4.74E-03	5.00E-05			2.86	0.0226	
			3.11	1.42E-03	2.50E-05			1.8	0.0179	
			3.57	1.81E-03	2.50E-05			1.8	0.0179	



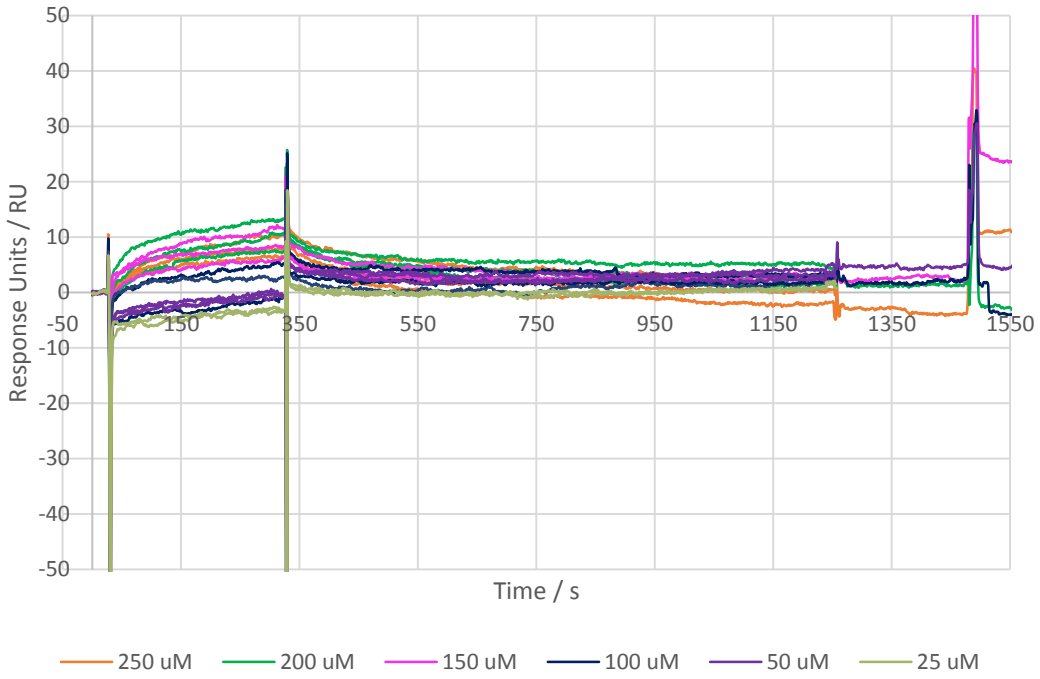
C5 Raw SPR Data: Compound 25



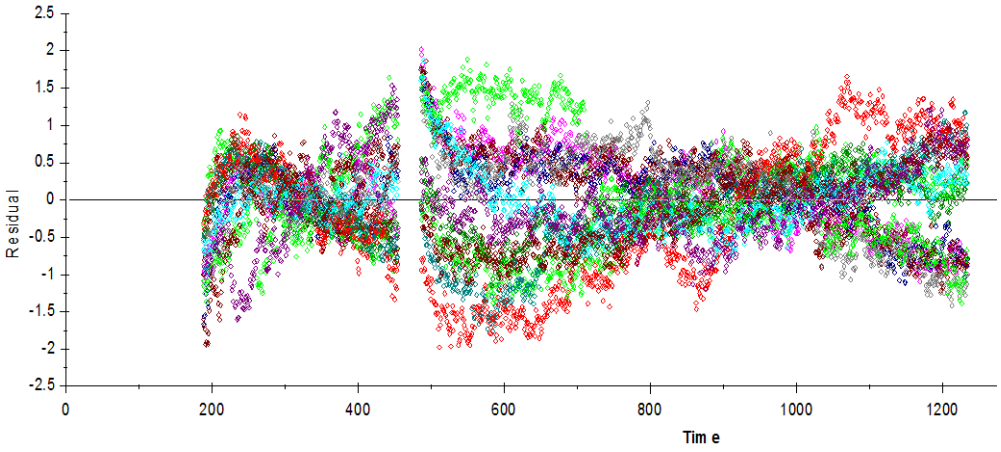
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
0.732	1.10E-05					6.67E+04	1.50E-05			0.303
		118	-0.472	1.80E-03	2.50E-04			112	1.94E-04	
		96.3	-2.24	-3.50E-04	2.50E-04			90.8	1.94E-04	
		129	-2.33	2.23E-04	2.00E-04			120	1.57E-04	
		102	-3.53	-5.10E-04	2.00E-04			94.4	1.57E-04	
		180	-1.75	5.42E-04	2.00E-04			167	1.57E-04	
		2.20E-03	-6.18	-3.52E-03	1.50E-04			2.00E-03	1.21E-04	
		138	-3.36	-9.61E-04	1.50E-04			125	1.21E-04	



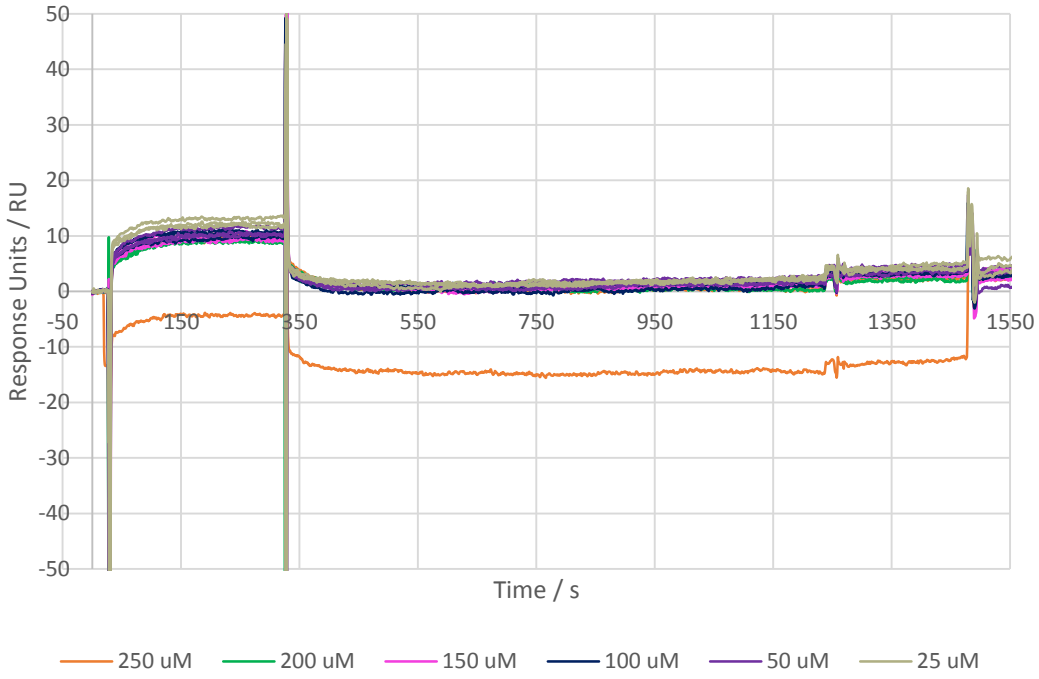
C6 Raw SPR Data: Compound 31



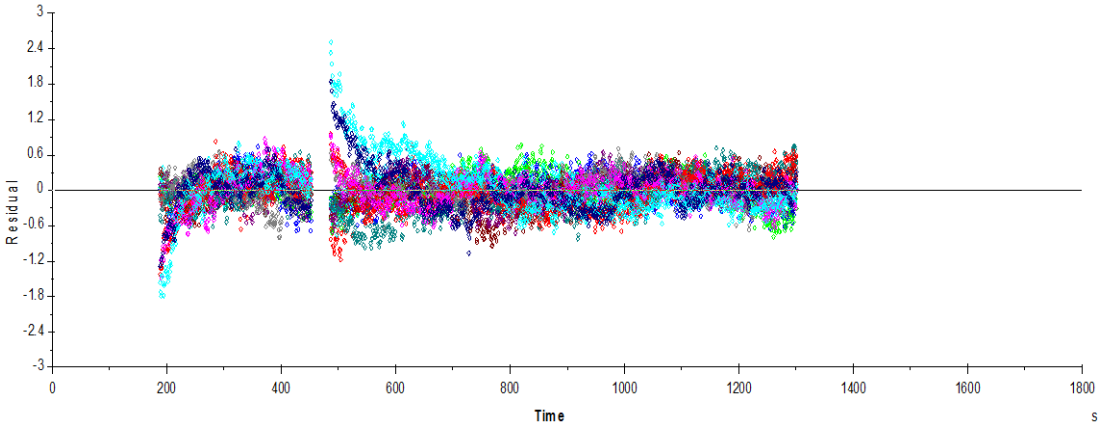
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
9.01	3.88E-03	24.2				2.32E+03	4.30E-04			0.374
			2.1	3.32E-03	2.50E-04			8.88	6.13E-03	
			1.3	-3.33E-04	2.50E-04			8.88	6.13E-03	
			3.13	3.01E-03	2.00E-04			7.67	5.68E-03	
			5.98	5.41E-03	2.00E-04			7.67	5.68E-03	
			3.43	1.93E-03	1.50E-04			6.25	5.23E-03	
			1.03	2.04E-03	1.50E-04			6.25	5.23E-03	
			0.43	3.74E-03	1.00E-04			4.56	4.78E-03	
			-0.0612	6.73E-05	1.00E-04			4.56	4.78E-03	
			-3.45	2.70E-03	5.00E-05			2.52	4.33E-03	
			-3.02	2.69E-03	5.00E-05			2.52	4.33E-03	
			-4.9	-3.02E-04	2.50E-05			1.33	4.10E-03	
			-4.9	-3.78E-04	2.50E-05			1.33	4.10E-03	



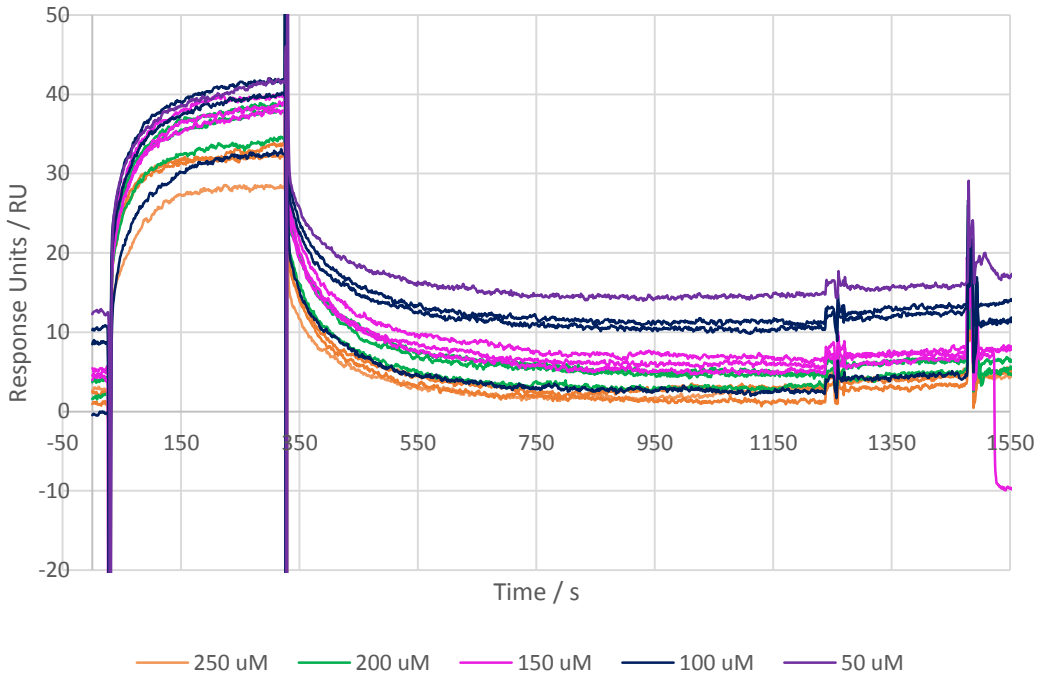
C7 Raw SPR Data: Compound 42



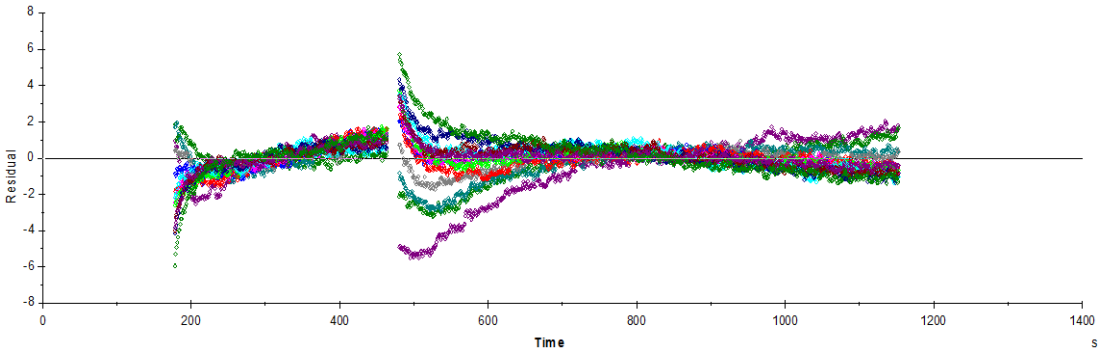
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
71.4	0.0203	14				3.53E+03	2.84E-04			0.0986
			3.12	7.13E-04	2.50E-04			6.55	0.0381	
			2.71	3.40E-04	2.50E-04			6.55	0.0381	
			2.91	1.39E-03	2.00E-04			5.78	0.0345	
			2.98	3.81E-04	2.00E-04			5.78	0.0345	
			4.22	6.83E-04	1.50E-04			4.84	0.031	
			4.23	9.58E-04	1.50E-04			4.84	0.031	
			5.8	1.34E-04	1.00E-04			3.65	0.0274	
			6.53	1.47E-03	1.00E-04			3.65	0.0274	
			7.61	1.04E-03	5.00E-05			2.1	0.0238	
			7.55	1.53E-03	5.00E-05			2.1	0.0238	
			10.3	2.17E-03	2.50E-05			1.13	0.022	
			10.2	1.76E-03	2.50E-05			1.13	0.022	



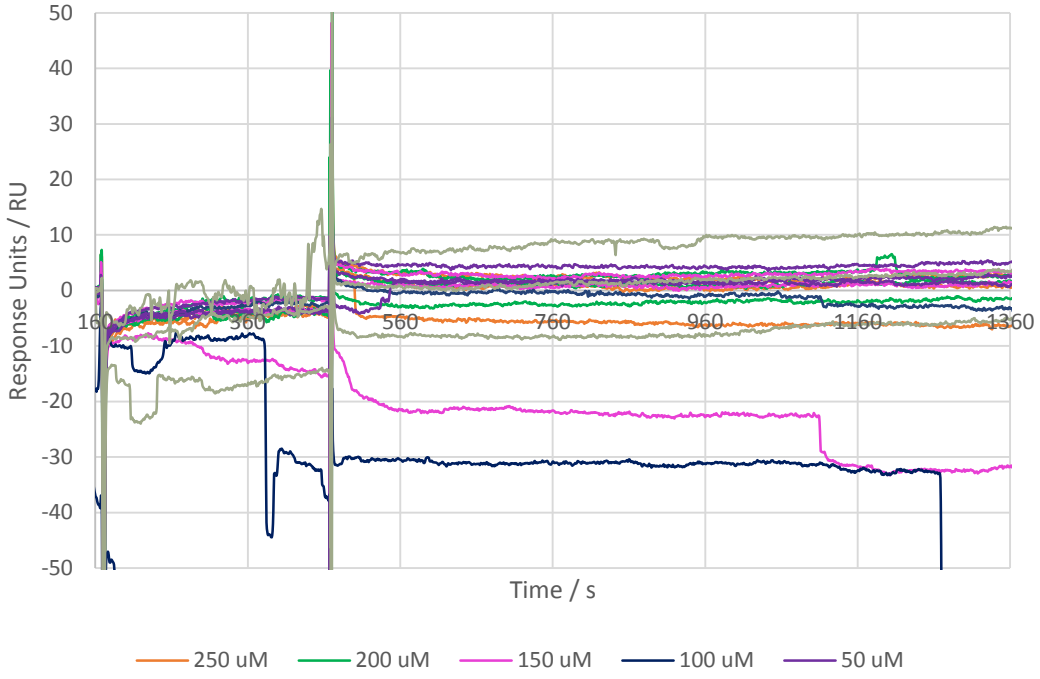
C8 Raw SPR Data: Compound 48



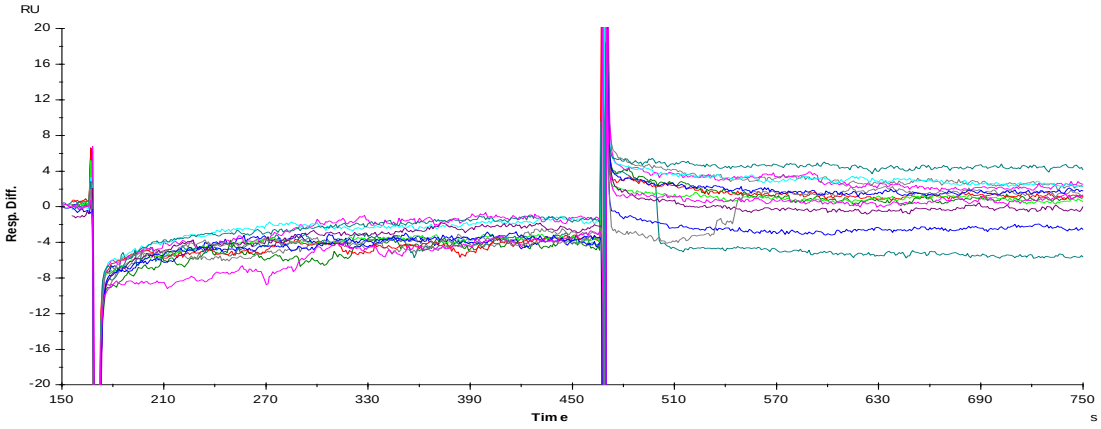
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
116	8.54E-03	23.2				1.35E+04	7.38E-05			0.858
			7.14	-2.75E-03	2.50E-04			17.9	0.0375	
			12	-5.09E-04	2.50E-04			17.9	0.0375	
			13.3	-4.63E-05	2.50E-04			17.9	0.0375	
			14.2	9.16E-04	2.00E-04			16.9	0.0317	
			15.2	1.61E-03	2.00E-04			16.9	0.0317	
			16	7.72E-04	2.00E-04			16.9	0.0317	
			16.3	1.23E-03	1.50E-04			15.5	0.0259	
			16.5	1.67E-03	1.50E-04			15.5	0.0259	
			17.6	2.76E-03	1.50E-04			15.5	0.0259	
			17.6	4.11E-03	1.00E-04			13.3	0.0201	
			16.4	2.63E-03	1.00E-04			13.3	0.0201	
			16.6	1.12E-03	1.00E-04			13.3	0.0201	
			17.8	3.02E-03	5.00E-05			9.35	0.0143	



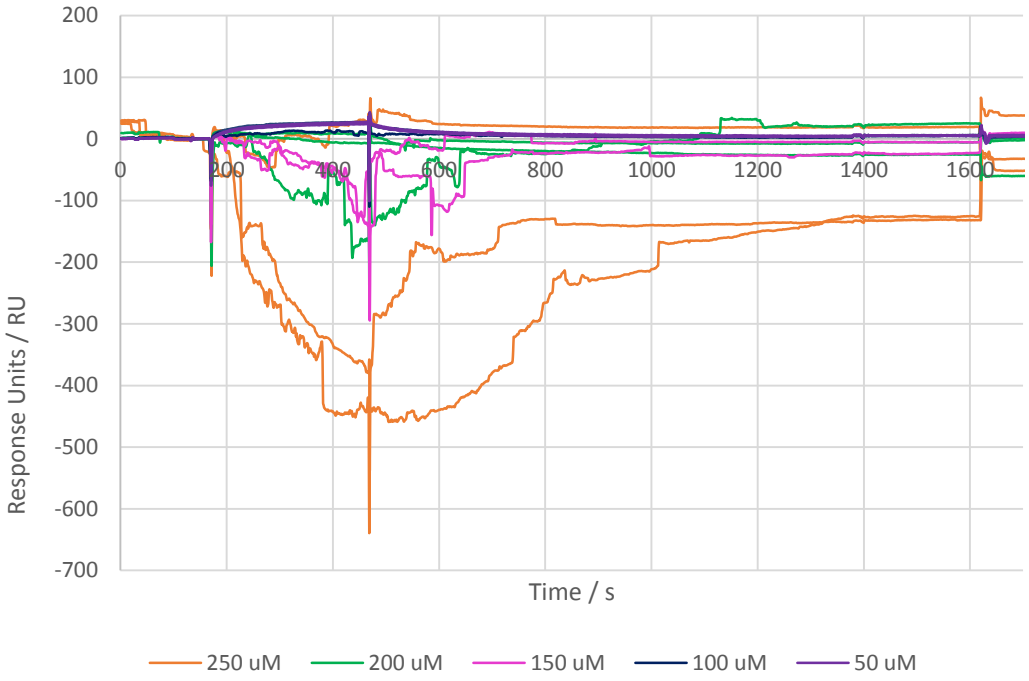
C9 Raw SPR Data: Compound 50



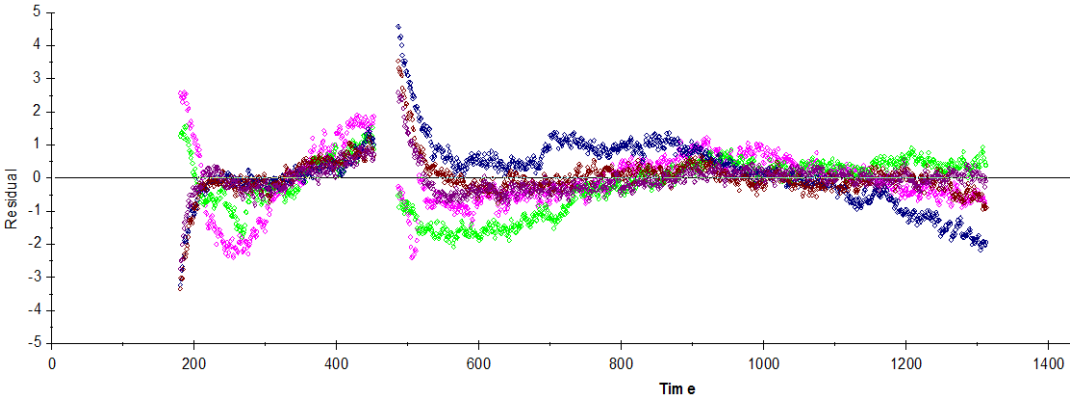
Compound 50 showed no evidence of binding, once uninterpretable sensorgrams were removed (below) therefore no further analysis was carried out.



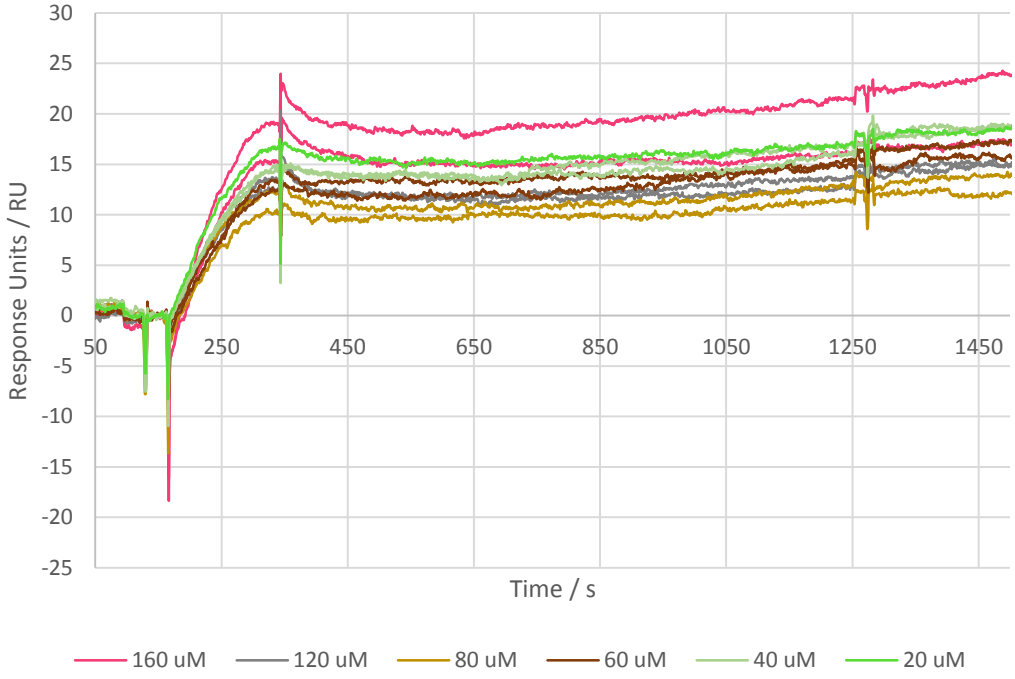
C10 Raw SPR Data: Compound 56



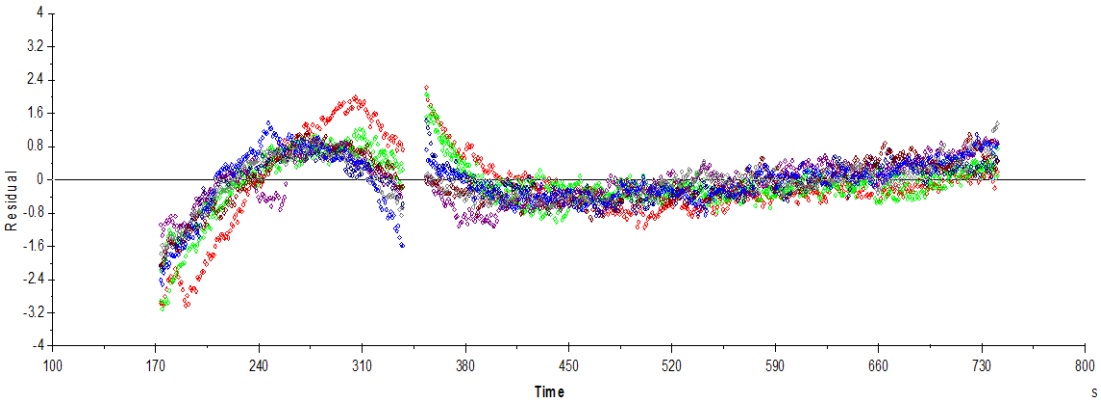
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
399	5.14E-03	25.9				7.77E+04	1.29E-05			0.588
			3.05	3.61E-03	5.00E-05			20.6	0.0251	
			4.86	2.41E-03	5.00E-05			20.6	0.0251	
			7.52	5.85E-03	2.50E-05			17.1	0.0151	
			6.06	3.55E-03	2.50E-05			17.1	0.0151	
			5.19	3.25E-03	2.50E-05			17.1	0.0151	



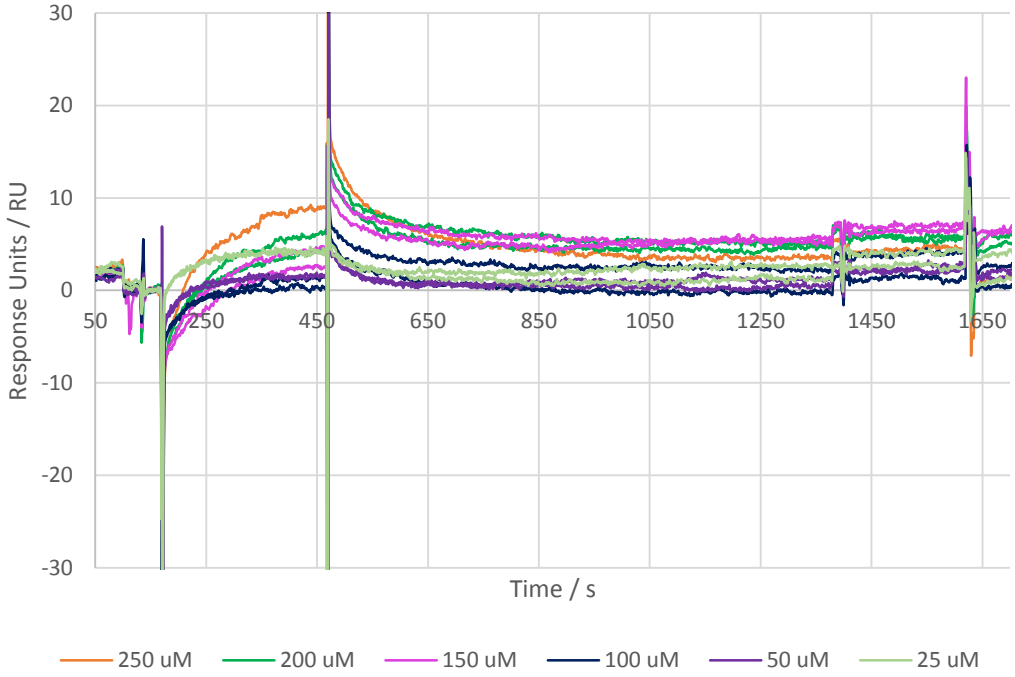
C11 Raw SPR Data: Compound 57



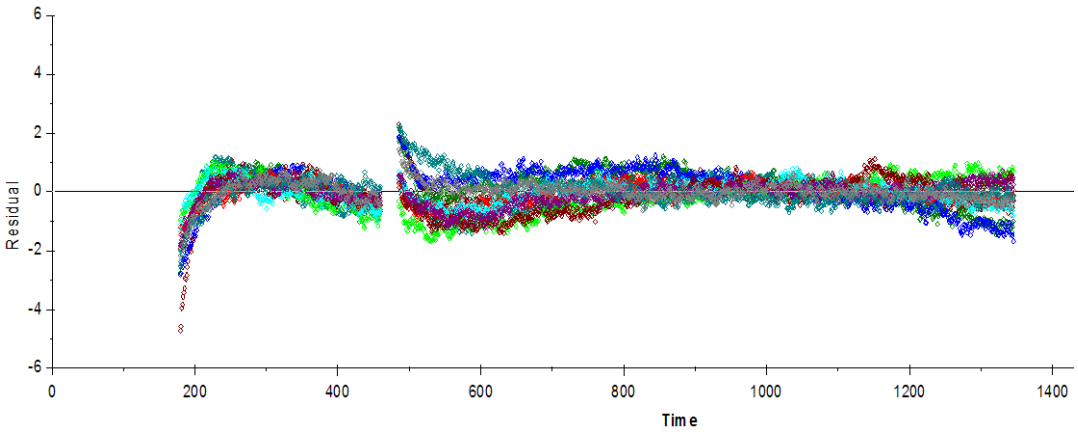
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
58.8	2.94E-04				2.00E+05	4.99E-06			0.412
		20.9	-1.91	1.60E-04			20.2	9.71E-03	
		18.4	0.586	1.20E-04			17.7	7.35E-03	
		18.1	0.0806	1.20E-04			17.4	7.35E-03	
		20.8	1.35	8.00E-05			1.96E+01	5.00E-03	
		31.2	-0.0607	6.00E-05			28.8	3.82E-03	
		28.1	0.0648	6.00E-05			25.9	3.82E-03	
		44.5	0.976	4.00E-05			39.6	2.65E-03	
		89	2.12	2.00E-05			71.2	1.47E-03	



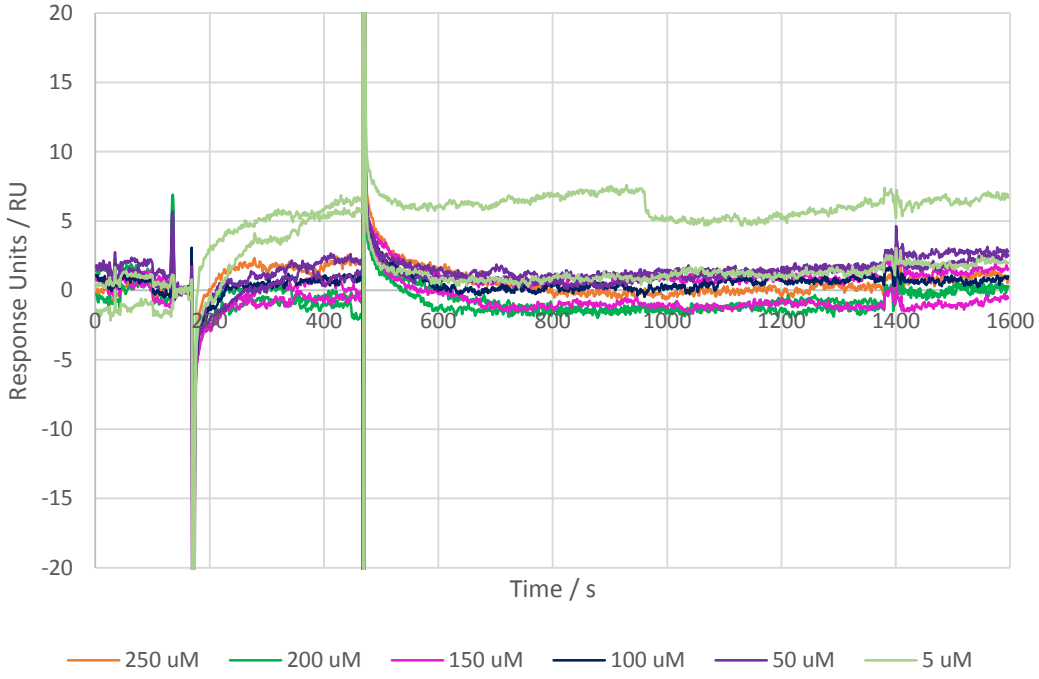
C12 Raw SPR Data: Compound 58



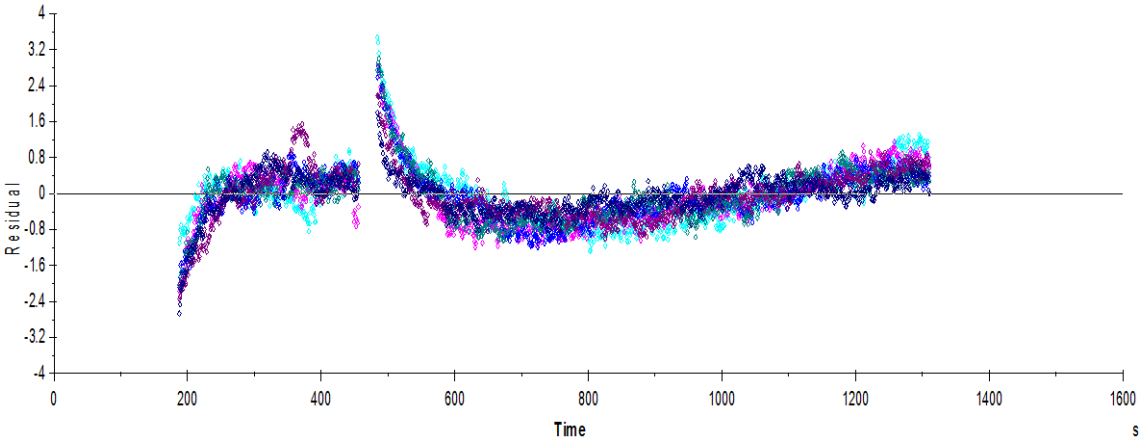
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
10.7	4.98E-03	36.5				2.15E+03	4.65E-04			0.27
			-2.87	3.33E-03	2.50E-04			12.7	7.66E-03	
			-4.37	5.26E-03	2.00E-04			11	7.12E-03	
			-5.65	3.99E-03	2.00E-04			11	7.12E-03	
			-4.38	5.58E-03	1.50E-04			8.89	6.59E-03	
			-5.9	4.99E-03	1.50E-04			8.89	6.59E-03	
			-3.96	-7.95E-04	1.00E-04			6.45	6.05E-03	
			-3.94	2.22E-03	1.00E-04			6.45	6.05E-03	
			-1.18	7.09E-04	5.00E-05			3.54	5.52E-03	
			-0.988	6.58E-05	5.00E-05			3.54	5.52E-03	
			2.34	1.18E-03	2.50E-05			1.86	5.25E-03	
			2.16	2.51E-03	2.50E-05			1.86	5.25E-03	



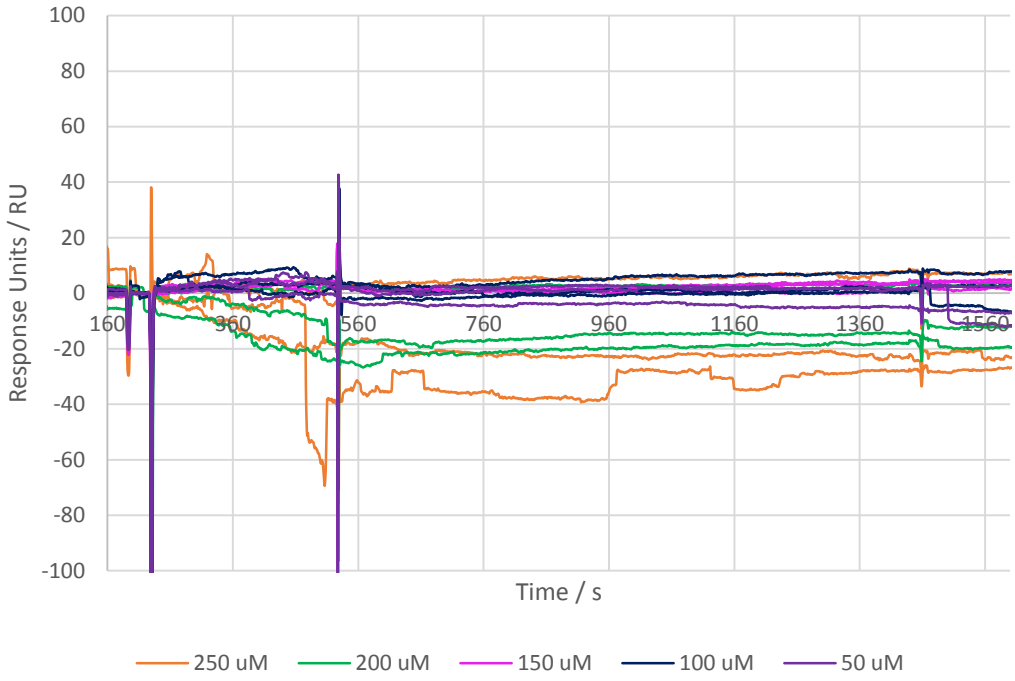
C13 Raw SPR Data: Compound 64



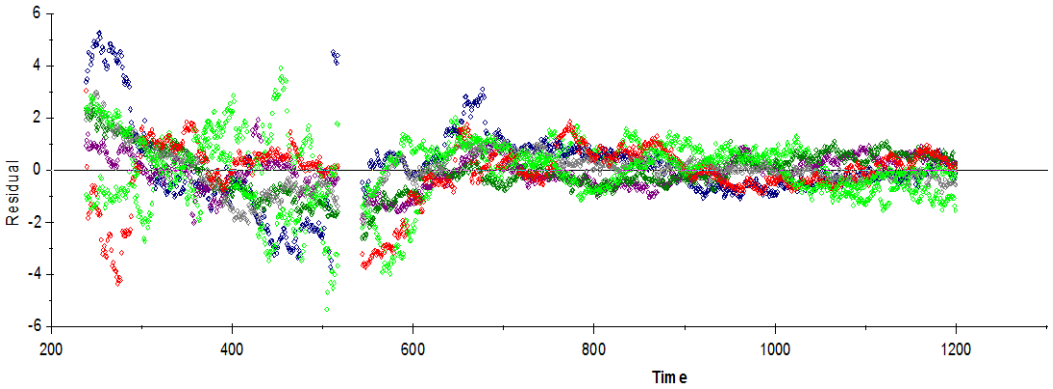
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
69.9	2.64E-04					2.65E+05	3.78E-06			0.356
		2.56	-0.271	-2.43E-03	2.50E-04			2.52	0.0177	
		0.0216	-0.973	-1.39E-03	2.00E-04			0.0212	0.0142	
		0.824	-1.25	-1.95E-03	1.50E-04			0.804	0.0107	
		1.15	-0.283	-4.76E-04	1.00E-04			1.11	7.25E-03	
		1.93	-0.542	-1.86E-05	5.00E-05			1.79	3.76E-03	
		12.3	4.3	5.47E-05	5.00E-06			7.02	6.13E-04	



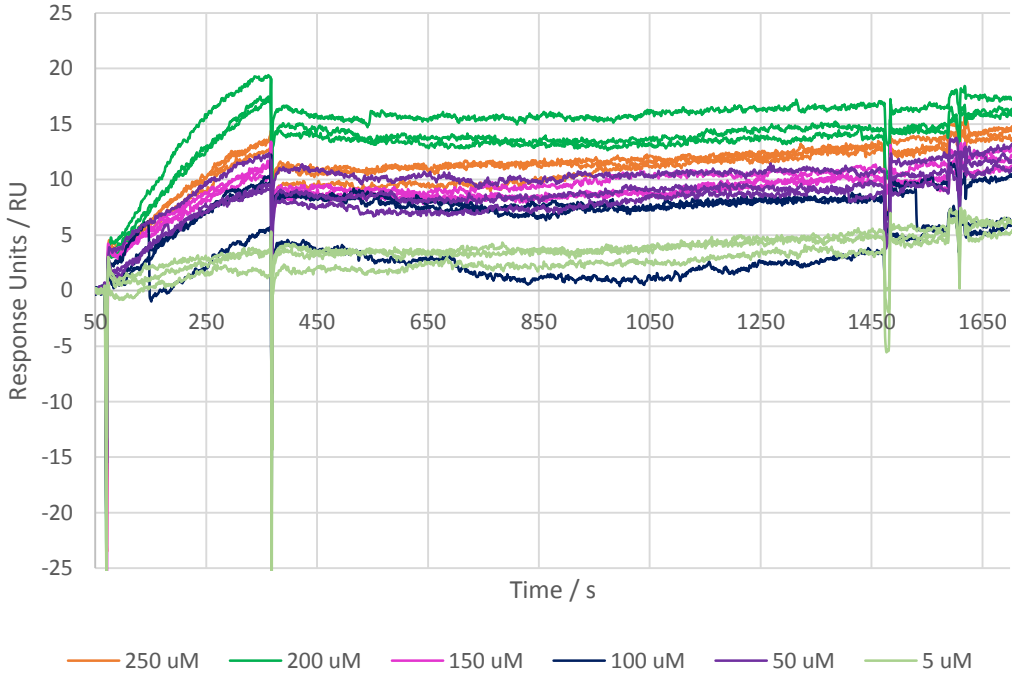
C14 Raw SPR Data: Compound 80



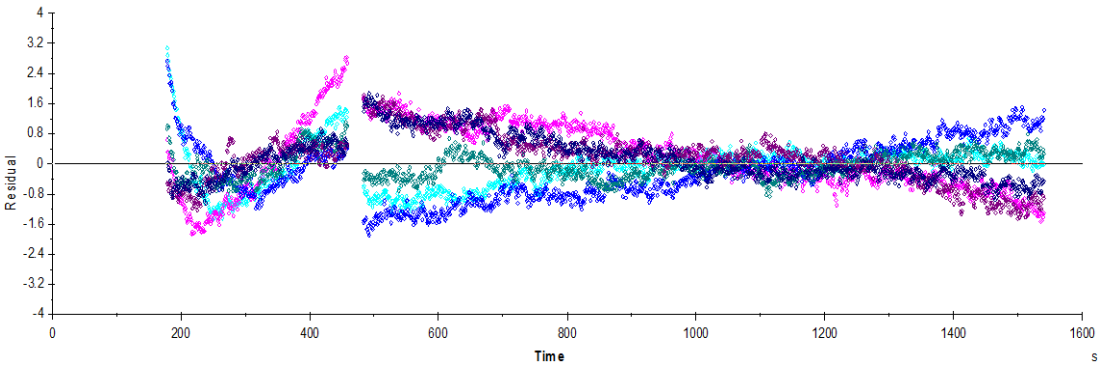
ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
1.05	1.86E-04					5.63E+03	1.78E-04			1.14
		242	2.31	8.39E-03	2.50E-04			141	4.48E-04	
		301	-3.43	-1.14E-03	2.00E-04			160	3.96E-04	
		0.388	-0.816	-3.68E-03	2.00E-04			0.206	3.96E-04	
		0.245	-1.8	-6.46E-04	1.50E-04			0.112	3.44E-04	
		7.88	-2.85	-3.20E-03	1.50E-04			3.61	3.44E-04	
		1.68	-3.4	-6.88E-04	1.00E-04			0.607	2.91E-04	
		44.1	-3.87	-2.94E-03	5.00E-05			9.69	2.39E-04	



C15 Raw SPR Data: Compound 85



ka (1/Ms)	kd (1/s)	Rmax (RU)	RI (RU)	Drift (RU/s)	[Analyte] (M)	KA (1/M)	KD (M)	Req (RU)	kobs (1/s)	Chi2
54.1	1.13E-03	11.2				4.80E+04	2.09E-05			0.475
			-0.167	7.15E-03	2.50E-04			10.3	0.0147	
			1.87	9.57E-03	2.00E-04			10.1	1.19E-02	
			0.185	4.68E-03	1.50E-04			9.81	9.24E-03	
			0.189	4.18E-03	1.00E-04			9.25	6.54E-03	
			1.97	6.54E-03	5.00E-05			7.89	3.83E-03	
			1.1	3.73E-03	5.00E-06			2.16	1.40E-03	



APPENDIX D

D1 Preliminary Data Collection and Refinement Statistics

Table D1: Preliminary data collection and refinement statistics for co-crystal data presented in chapter eight. Note: * number in brackets represents last resolution shell.

Compound	3	10	15 (-ligand)	25	42	57	60	74 (-ligand)	80 (-ligand)	85
Beamline	DLS-I03	DLS-I03	DLS-I03	DLS-I04-1	DLS-I03	DLS-I04-1	DLS-I04-1	DLS-I03	DLS-I04-1	DLS-I04-1
Wavelength [Å]	0.96000	0.96000	0.96000	0.91741	0.96000	0.91741	0.91741	0.96000	0.91741	0.91741
Space Group	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$	$P4_12_12$
a, b[Å]	121.76	122.54	118.76	120.08	120.48	121.41	121.39	121.54	121.54	120.89
c [Å]	33.63	33.82	33.49	33.66	33.65	33.69	33.76	33.78	33.60	33.68
α,β,γ [°]	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Resolution range [Å]	86.10-1.80	86.65-1.50	83.98	84.91-1.40	85.19-1.95	85.85-2.10	85.84-1.60	85.94-1.50	85.94-1.70	85.48-1.80
Wilson B factor [Å]	32.1	24.7	37.8	24.3	34.2	31.7	31.5	26.3	35.3	36.0
I/sigma*	21.3 (4.2)	12.5 (3.2)	21.0 (3.7)	22.6 (3.2)	13.6 (4.0)	14.3 (2.9)	17.5 (3.2)	18.7 (4.6)	20.5 (1.9)	14.9 (3.1)
R_{merge} [%]*	6.0 (52.8)	8.7 (62.2)	5.9 (60.4)	3.8 (45.5)	8.4 (54.6)	11.5 (65.8)	5.3 (52.0)	5.4 (44.8)	4.6 (67.0)	7.2 (61.3)
Completeness [%]*	99.6 (98.1)	99.7 (100)	995. (97.5)	99.2 (96.3)	99.7 (99.5)	99.4 (96.9)	99.9 (100.0)	99.9 (100.0)	98.7 (92.7)	98.4 (95.8)
CC1/2	99.9 (93.5)	99.8 (90.9)	99.9 (92.0)	100 (90.1)	99.7 (88.6)	99.8 (86.7)	99.9 (90.0)	99.9 (95.9)	100.0 (78.1)	99.9 (85.9)
R [%]	18.3	19.0	19.3	19.9	18.5	19.3	18.3	20.1	19.1	21.6
R_{free} [%]	22.6	22.0	21.3	22.1	23.8	23.5	22.1	22.0	22.6	25.9
# Residues	188	188	188	188	188	188	188	188	188	188
# waters	122	93	63	62	86	63	85	93	84	14
RMSD bond length [Å]	0.02	0.02	0.02	0.02	0.02	0.019	0.03	0.03	0.02	0.02
RMSD bond angles [°]	1.81	1.91	1.78	1.84	1.89	1.86	2.54	2.26	1.77	1.89