Fragment Based Ligand Discovery
- Library Design and
Screening by Thermal Shift Analysis -

Michèle Nadine Schulz

PhD
University of York
Chemistry

March 2012

Abstract

The central idea in Fragment Based Ligand Discovery (FBLD) is to identify small, low molecular weight compounds (MW $<$ 250) that bind to a particular protein active site. Hits can be used to efficiently design larger compounds with the desired affinity and selectivity.

Three approaches to FBLD are described in this thesis.

The first topic is the development and assessment of different chemoinformatics procedures to select those fragments that maximally represent the chemical features of a larger compound library. Such a fragment library could be of great value in the so-called “SAR by Catalogue” approach, where the initial stage of fragment growth is by selecting existing compounds that contain sub-structures of the hit fragments. Five schemes implemented in the Pipeline Pilot software are described.

The second project was to develop improved approaches to processing Thermal Shift Analysis (TSA) data. The shift in melting temperature can indicate that a ligand binds and thus stabilises a protein. A program, MTSA, has been written which allows more straightforward processing of the experimental data than existing available software. However, detailed analysis of fragment screening data highlighted difficulties in defining the melting temperature and suggest that TSA is not sufficiently reliable for routine screening use.

Finally, a number of proteins were assessed experimentally for suitability for FBLD: N-myristoyl transferase (NMT), the bacterial homologue of a GlcNAcase enzyme (BtGH84) and the model system hen egg white lysozyme (HEWL). It was not possible to produce suitable NMT material due to the inherent instability of the protein produced in York. The screening results of HEWL with a new Surface Plasmon Resonance (SPR) assay, a cell based activity assay and TSA were inconsistent and difficult to interpret. However, BtGH84 was suitable for screening by both TSA and SPR. The resulting fragment hits are suitable starting points for further evolution.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor Prof. Rod Hubbard for giving me the opportunity and the support to do my PhD in the YSBL.

Many thanks also go to Jens Landström who greatly helped and supported me during his time in our group.

I also would like to thank my group, Kerrin Bright, Marcus Fischer, Abi Bubb, Kamran Haider and Jens Landström for overall discussion, help and support. Eddy Vande Water helped with some library design, Fiona Whelan, Javier Garcia and Glyn Hemsworth assisted with protein production and structure solution. Thanks to Ralph Hyde for cloning, Bailey Massa for her results with BtGH84, Yuan He and Wendy Offen for BtGH84. Shirley Roberts, Sam Hart and Johan Turkenberg helped with crystallisation and data collection, Jim Brannigan with NMT, Seishi Shimizu with thermodynamic discussions and Sally Lewis and Simon Grist for mass spectrometry and general assistance in the lab.

There many, many more people I should mention here who helped and supported me during my project. My thanks go to all members of YSBL, staff from the TF and from Chemistry.

Special thanks also go to Jerome Dabin, Glyn Hemsworth, Fiona Whelan, Cori Dressler, Judith Stepper, Sue Couling and Jane Thomas-Oates for carefully reading this manuscript and giving many useful suggestions. I further want to thank James Tunaley, Dan Peters, Dan Wright, Katie Jameson and Abi Bubb for proofreading.

Last but not least, I thank the BBSRC and Dr. Anthony Wild for financial support.

Author's declaration

I declare that I have written this thesis on my own and clearly indicated all work which was not done by myself.

Several papers were published during this thesis:

SCHULZ, M. N., LANDSTROM, J., BRIGHT, K. & HUBBARD, R. E. 2011. Design of a fragment library that maximally represents available chemical space. J Comput Aided Mol Des, 25, 611-20

SCHULZ, M. N. & HUBBARD, R. E. 2009. Recent progress in fragment-based lead discovery. Curr Opin Pharmacol, 9, 615-21.

SCHULZ, M. N., LANDSTROM, J. & HUBBARD, R. E. 2012. MTSA - A Matlab Program to Fit Thermal Shift Data. Anal Biochem, [Epub ahead of print]

Chapter 1

Introduction

This thesis describes the development and application of methods in Fragment Based Ligand Discovery (FBLD), which has emerged over the past ten years as an innovative approach to the early stages of drug discovery. This first chapter provides a brief historical introduction to drug discovery and describes the origins of, and context for, the fragment based approach. This introduction includes a more comprehensive review of the concepts and approaches to the design of fragment libraries for screening, which is of particular relevance to the work described later in the thesis. The chapter concludes with an overview of the organisation of the thesis and the aims of the research.

1.1 Drug Discovery

1.1.1 A Historical Perspective

A medicine is a drug that is used for the treatment or prevention of a disease or condition. Since ancient times people have used plant extracts for medicinal purposes – discovered accidentally or empirically by testing. Examples of the earliest drugs include salicylic acid (found by chewing willow bark, today used as salicylic acid acetate in Figure 1.1), opium alkaloids, cannabis and ephedrine. All the major civilisations of the ancient world developed the practice of medicine with associated pharmacies which combined extracts from herbs and plants (and occasionally minerals) to produce treatments for various ailments and conditions.

For example around 1300 B.C., the Minoans exported opium to Egypt, where it was used to calm crying children.

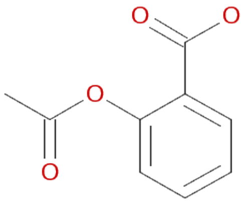
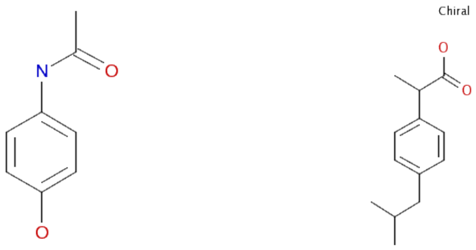


Figure 1.1: **Acetylsalicylic acid.** Marketed under a large number of brand names, famous example: Aspirin (Bayer)

It was not until the 18th century that chemical science developed sufficiently to isolate individual components with scientists such as Scheele and Lavoisier using crystallisation to purify the active ingredients. In addition, metals (such as mercury and antimony) and salts (such as silver nitrate) could now be purified for use as drugs.

However, it was the late 19th century before chemical methods advanced sufficiently to enable the determination of chemical structure and the ability to synthesize individual compounds, giving birth to the modern pharmaceutical industry. Some of the first apothecary companies moving into large-scale drug production were Merck, Schering and Boehringer, followed by chemical companies such as Bayer and Hoechst.

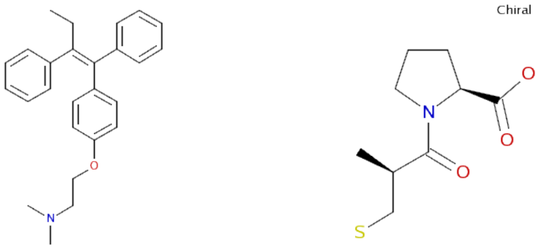


(a) Paracetamol

(b) Ibuprofen

Figure 1.2: **Paracetamol and ibuprofen.** Paracetamol, also acetaminophen: analgesic and antipyretic effects, Ibuprofen: analgesic and antipyretic effects

Until the middle of the 20th century, drugs were discovered by making compounds, and testing for an effect on appropriate animal or cellular models and looking for



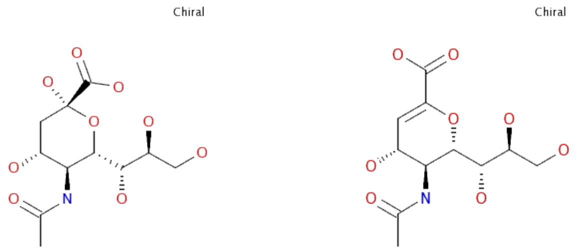
(a) Tamoxifen (b) Captopril

Figure 1.3: **Tamoxifen and Captopril.** Tamoxifen: selective estrogen receptor modulator, Captopril: first orally-active angiotensin-converting enzyme (ACE) inhibitor.

the desired effect or phenotype. The ideas for synthesis came from analysis of known drugs and their metabolites. This is how drugs such as paracetamol and ibuprofen were discovered (Figure 1.2). Alongside such "make and test" strategies, the 1950s to 1970s saw an increased understanding of metabolic pathways, the biochemistry of viruses and bacteria and the identification of some of the key receptors (such as the first steroid receptors). This gradually introduced the idea that modulating the activity of a specific target was an effective strategy for drug discovery for some diseases and conditions. Examples include the development of steroid hormone modulators (e.g. tamoxifen, a pioneering breast cancer treatment, Figure 1.3 a, recently reviewed by Jordan, 2006) and angiotensin-converting enzyme (ACE) inhibitors. Captopril was the first orally-active ACE inhibitor and is used to treat hypertension (Figure 1.3 b, Ondetti et al., 1977; Rubin et al., 1978).

The design of compounds was usually guided by an understanding of the molecular enzymology and making compounds that looked like the substrates or which mimicked the transition state (for enzymes). One such example was an initial attempt at an inhibitor, Neu5Ac2en (Figure 1.4 b), a neuraminidase-transition stage analogue for sialic acid (Figure 1.4 a), for influenza (Meindl and Tuppy, 1969).

However for complex conditions – such as many psychiatric disorders – phenotypic screening continued to be the main way for discovery until very recently. The primary reasons are that therapeutic benefit comes from affecting a whole system of targets and cellular processes. For this reason, although projects are influenced



(a) Sialic Acid (b) Neu5Ac2en

Figure 1.4: **Sialic Acid and Neu5Ac2en.** Sialic acid: also known as N-acetyl neuramic acid (Neu5Ac), Neu5Ac2en: transition state analogue.

by the knowledge of targets, the primary driver for compound optimization is *in vivo* testing.

During the late 1980s, molecular biology revolutionised target discovery and there have been rapid and continuing improvements of our understanding of the molecular basis of diseases. An example is a research programme starting in the 1980s with the intention to finding drugs for the treatment of hypertension and cardiovascular diseases. The scientists at Pfizer tried to find inhibitors of phosphodiesterase (PDE5). The starting point was a weak phosphodiesterase inhibitor found in the literature which was originally an anti-allergy compound. It showed an effect *in vitro* and in an animal model. Computational modelling and rational chemistry increased the affinity and selectivity of the initial compound. Structure-affinity relationships and pharmacokinetics properties were optimised and led to Sildenafil (Figure 1.5). The drug proved not be very effective in the treatment of cardiovascular conditions, but having a positive effect of erectile dysfunction and is now marketed as Viagra (Campbell, 2000). Sildenafil is a selective inhibitor for PDE5. Nevertheless, weak interactions with other isoforms of the protein can lead to side effects. For example transient visual disturbances may have resulted from the interaction with PDE6 (Wallis et al., 1999).

Perhaps the first example of structure based drug design is the discovery of influenza drugs targeting neuraminidase. Analysis of the crystal structure of the enzyme neuraminidase in complex with the transition state mimic Neu5Ac2en (Figure 1.4 b) showed there was an additional pocket at the base of the active site. Computational analysis using some of the earliest fragment ideas (the program GRID; Goodford, 1985; see also Section 1.1.7) identified that a positive amino or

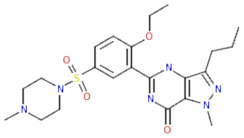
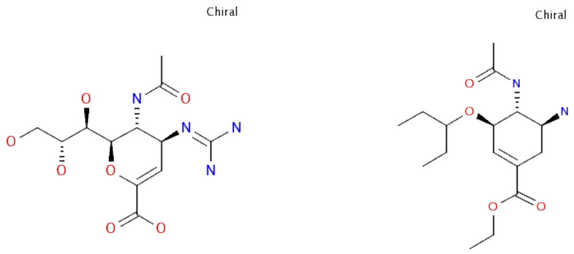


Figure 1.5: **Sildenafil.** PDE5 inhibitor marketed as Viagra.

guanidinium group would have enhanced affinity. Synthesis of such compounds resulted in Relenza (marketed by Glaxo; Figure 1.6 a) with another mimic, Tamiflu (from Gilead Sciences, Figure 1.6 b), closely following (Von Itzstein et al., 1993; Colman, 2006; Lew et al., 2000). The target-oriented approach to drug discovery has dominated for the past 20 years; this has led to most projects following a standardised drug discovery process: the "drug pipeline". Before discussing the pipeline in more detail, it is important first to consider the properties that are required in a drug molecule.



(a) Zanamivir

(b) Oseltamivir

Figure 1.6: **Zanamivir and oseltamivir.** Zanamivir is marketed as Relenza, Oseltamivir as Tamiflu respectively.

1.1.2 Drug-like Molecules

During the 1990s, many drug discovery projects failed in later clinical trials because the drug candidate molecules did not have the right balance of properties for efficacy when given to man. In the late 1990s, Lipinski et al. (2001, first in 1996) analysed the physicochemical properties of known orally available drugs.

The Lipinski *Rule of Five* states that oral absorption and permeation of a compound are better if it has:

- $\leq$ 5 hydrogen bond donors (expressed as the sum of OHs and NHs)
- $\leq$ 10 hydrogen bond acceptors (expressed as the sum of Ns and Os)
- $\leq$ 500 Da molecular weight
- $\leq$ 5 CLogP

(More explanation on molecular descriptors follows in Section 2.4 on page 57.)

The so-called *Rule of Five* was a landmark definition, as although it has occasionally been over-used as a concept (good drugs are coming through which do not obey the rules; some people forget it is just an empirical guide for oral bioavailability), it has increased the attention paid to such properties for drug-like molecules.

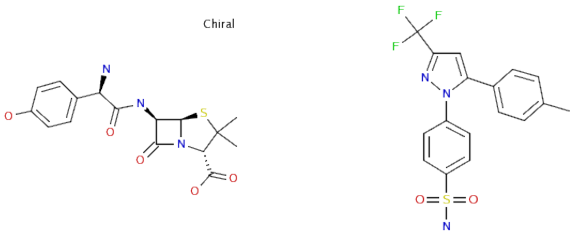
A drug molecule has to achieve a sufficient concentration at the site of action in the body to bind to and modulate the target for a long enough period. A large number of parameters are monitored and optimised during the drug discovery process (see Section 1.1.3 below), the most important of which are:

Affinity

The affinity between a ligand and a protein can be determined with a various number of affinity constants such as K_D, IC_{50} etc. The stronger a drug is binding to a protein (i.e. the smaller the affinity constant), the less amount of drug is needed for the desired effect. This also reduces unwanted side effects. The type of affinity constant obtained depends on the used assay. The dissociation constant K_D is in particular important because it is an universal constant and does not depend on the assay conditions. It corresponds to the concentration of ligand at which the binding site of a protein is half occupied. The IC_{50} is the half maximal inhibitory concentration. It measures the effect of a compound in inhibiting a biological or biochemical function in an assay (for example a dose-response assay), and hence varies between experiments.

Selectivity

A highly selective drug can reduce undesired side effects. However, drugs targeting a certain protein class, e.g. kinases, will all target the same substrate pocket across the species, and selectivity is difficult to achieve. Relenza (zanamivir) selectively inhibits influenza-specific neuraminidase (Figure 1.6 a, Barnett et al., 2000). Sometimes, selectivity is not desired, for example for broad spectrum antibiotics (e.g. amoxicillin in Figure 1.7 a). Binding to different isoforms of a protein can also increase the effect and decrease toxicity, e.g. non-selective COX (cyclooxygenase) inhibitors such as paracetamol (reviewed in Botting, 2006; 1.2 a). In contrast, selective COX-2 inhibitors (e.g. celecoxib) are controversial these days because they might exhibit thrombotic cardiovascular problems (Figure 1.7 b, Mukherjee, 2002).



(a) Amoxicillin (b) Celecoxib

Figure 1.7: **Amoxicillin and celecoxib.**

Efficacy

The measurement of efficacy is therapeutic and often target dependent. During the early stages of discovery, cell assays are used to monitor whether the drug gets into cells. Usually they monitor whether any effect on cell biology can be observed, often seen as an increase or decrease in the presence or modification of some protein in the cell. These so-called pharmacodynamic (PD) markers indicate what effect the compound is having on the body. Often, such markers are also followed during pre-clinical development and clinical trials, to show that the compound is reaching its site of action.

ADME/PK

Pharmacokinetics is the measurement of the effect of the body on the drug. To be effective, drugs need to be absorbed (A), be transported to or diffuse to the site of action (distribution, D), appropriately evade or interact with the body's defence mechanisms for destroying foreign small molecules (metabolism, M) and be excreted (E) at a suitable rate for therapy.

Toxicity

Toxicity is more appropriately known as tolerability. All molecules are toxic at some dose; what is important is that the drug has minimal unwanted side effects at therapeutic doses. This can be characterised *in vitro* by monitoring binding to a selection of receptors, ion channels, transporters and enzymes (the panel provided by Cerep is the most widely used) or by checking mutagenic potential in an Ames test (McCann et al., 1975). The important check is *in vivo*, checking whether the compound can be tolerated by the animal at an escalating dose.

1.1.3 The Drug Pipeline

The drug discovery pipeline can be represented as a linear process, summarised in Figure 1.8. It consists of three major stages: discovery, pre-clinical development and clinical trials. The discovery stage identifies a compound with the appropriate drug-like properties to be taken forward as a clinical candidate. The pre-clinical phase prepares the compound for introduction into man. The clinical trials check first that the compound is safe and then that it is effective and how best to provide it to which patients.

The terminology used to define and the metrics used to delineate the different parts of the discovery stage vary between organisations and also depend on the therapeutic area (for example, many CNS drug discovery pipelines introduce *in vivo* activity requirements quite early in the process; some oncology indications may not show therapeutic effect until in man). However, most have similar characteristics:

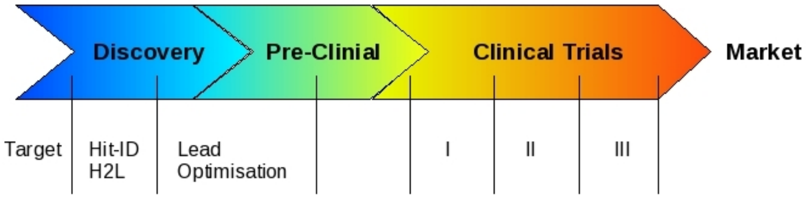


Figure 1.8: **Drug Discovery Pipeline.** The typical drug discovery approach starts with the identification of a target. After hit identification and hit to lead optimisation, the compound enters pre-clinical development and then enters three phases of clinical trials. (Figure adapted from Hubbard, 2006)

The first step is to identify the molecular target (Lindsay, 2003; Egner et al., 2005). This can be a protein, a nucleic acid or a multi-component complex. In many ways, a target is not truly validated until a drug that binds only to that target at therapeutic doses has been successfully used as a medicine for a number of years. For most new targets, the most that can be hoped for is a strong biological rationale, based on an understanding of the disease biology and relevant pathways, usually reinforced by cell or animal experiments where the target is disrupted (siRNA or knock-out), or for some conditions, from identification of particular mutations in patients.

The next step is to find hits that bind to the target and affect its activity. These can be relatively weak (low µM affinity, see Section 1.1.2). A process of exploratory chemistry (known as hits to leads) makes small changes to the hits to establish them as suitable for optimisation. The dominant criteria is establishing some understanding of which parts of the molecule are important for activity and how they can be modified (so-called SAR, structure-activity relationships) and that the compounds are amenable to further synthesis. At this stage, the other "drug-like" properties (see Section 1.1.2) are considered, but mainly to identify what needs to be changed during optimisation.

The most intensive step in discovery is lead optimisation. Here, the structure of the compound is altered to improve its properties, as measured by a range of *in vitro*, cellular and *in vivo* assays. The aim is to generate a clinical candidate with the desired mix of affinity, efficacy, ADME and physicochemical properties. At this point, the compound that could become a drug is fixed.

Pre-clinical development assesses the safety and feasibility of launching clinical trials in man. Synthetic routes are explored for making the large quantities of pure compound needed for clinical trials. The safety of the compound is assessed through an escalating series of dosing of various species as well as checking for activity in a panel of standard safety assays (Ames/Cerep panel etc.). Depending on the therapeutic area, there are continuing *in vivo* trials in animals to identify particular conditions or combination treatments. In addition, there can be considerable work on formulations – that is finding how best to deliver the drug during clinical trials.

Clinical trials are divided into phases I, II and III. In phase I, the safety of the drug is tested on a small number of healthy volunteers. Phase II assesses the actual efficacy of the drug while several hundreds of patients with the relevant disease/condition are given the drug or a placebo. Additional information about safety are also collected. In phase III, the trials are up-scaled to thousands of patients to obtain more information about side effects and the drugs effectiveness. If the drug passes all phases it can finally be launched on the market (Hubbard, 2006). Nevertheless, only one in nine drugs entering clinical development will make it to the market (Paul et al., 2010).

1.1.4 Lead-like Molecules

The late 1990s and 2000s saw an increased focus on compound properties in the drug discovery process. The Lipinski *Rule of Five* informed the criteria for a molecule to be considered drug-like; further analysis attempted to define what properties are desirable in a lead compound to give the maximum chance of retaining drug-like properties in the optimized compound. Such analyses led to the idea of lead-like molecules, with properties as:

- $\approx$ 450 Da MW
- $-3.5 < \text{CLogP} < 4.5$
- $\leq$ 4 rings
- $\leq$ 10 non-terminal single bonds

- < 5 hydrogen bond donors
- < 8 hydrogen bond acceptors

Leads have simpler chemical features to make them suitable for further chemical optimisation, an established structure activity relationship (SAR) series where similar compounds exhibit similar activity and have favourable ADME properties. In addition, patents issues of a scaffold should be avoided (Oprea et al., 2001).

1.1.5 A Matter of Chemical Space

In the 1990s, high throughput screening (HTS) was set in place by most pharmaceutical companies. With HTS, large compound libraries are screened in biochemical assays by robotic systems. The libraries often derive from earlier projects and combinatorial chemistry resulting in most HTS compounds already being drug-like according to Lipinski's *Rule of Five*, which leaves little room for lead optimisation. That also means the HTS libraries cover only a small part of the chemical space. For drug-like compounds in the size of 30 main atoms (C, N, O and S), the chemical space is estimated to comprise 10^{60} molecules (Bohacek, 1996). Even with a library of 1 million compounds, only a very small part of the chemical space can be covered. Another major drawback is the need for a suitable and robust assay. A bigger part of the chemical space can be assessed with virtual screening where compounds are docked into a 3-dimensional model of the protein target. The approach is most successful when the high resolution structure of the protein is known. Nevertheless, restraints are the scoring functions which predict the rank of the most likely binding compound.

Hann et al. (2001) state that the probability if a compound is a hit depends on the complexity of the compounds, rationalising the low success rate of HTS. If it is too small, hits will not be detected in a screen while if they are too complex, the likelihood of finding a hit in a random number of molecules is low. Hann and colleagues suggested screening smaller sized compounds. Today, the suggested smaller compounds are referred to as "fragments". If a fragment is considered to have only 11 main atoms, the fragment space contains 10^7 molecules (Fink, 2005). Comparison of the chemical space of molecules with 11 and 13 main atoms respectively shows that the addition of two heavy atoms increases the chemical universe already from 26.4 million to 977 million compounds, i.e. 37-fold (Reymond et al., 2011). Compared to the 10^{60} molecules in the drug-like space, the number of fragments which have to be tested experimentally is significantly minimized.

Useful starting points for lead identification for most targets can be identified from a relatively small (typically 1000-member) library of low molecular weight compounds in the 120–300 Da range.

1.1.6 Guiding Lead Selection and Optimisation

Since most screening campaigns result in more than one hit, sensitive choices need to be made. A useful metric for lead selection was introduced by Hopkins and colleagues to assess the quality of a binder and guide lead optimisation: The ligand efficiency index (LE) is defined as affinity per size (Hopkins et al., 2004) after an initially suggested concept by Kuntz et al. (1999) (also summarised in Perola, 2010):

$$LE = \frac{\Delta G_{binding}}{N_{Non-hydrogenAtoms}} \tag{1.1}$$

where $N_{Non-hydrogenAtoms}$ is the number of non-hydrogen atoms, and the free binding energy $\Delta G_{binding}$ (Gibbs energy) is defined by:

$$\Delta G_{binding} = -RTlnK_D \tag{1.2}$$

with R the gas constant, T the absolute temperature and K_D the dissociation constant.

A more practical description for LE uses affinity parameters directly measured by most experiments:

$$LE = \frac{pK_i/pK_D/pIC_{50}}{N_{Non-hydrogenAtoms}} \tag{1.3}$$

An alternative description for the LE, but not as frequently used, is referred to as binding efficiency index BEI (Abad-Zapatero and Metz, 2005):

$$BEI = \frac{pK_i/pK_D/pIC_{50}}{MW(kDa)} \tag{1.4}$$

The LE was extended with consideration of the energy of binding per functional group (Hajduk and Sauer, 2008; Verdonk and Rees, 2008) and including a consideration of the lipophilicity. Lipophilic compounds obtain their binding energy by desolvation which is not specific. Thus too lipophilic compounds can be more promiscuous. The lipophilic ligand efficiency LLE was introduced by Leeson and Springthorpe. in 2007:

$$LLE = pK_i/pIC_{50} - ClogP/ClogD \tag{1.5}$$

where $LogP$ is the octanol-water partition coefficient and $LogD$ the octanol-water distribution coefficient. The higher the value, the higher the lipophilicity. Fragments tend to be more polar than other compounds which is another advantage of using fragments (Congreve et al., 2008). The ligand efficiency based on pIC_{50} (Equation 1.3) is typically between 0.3 for initial hits and 1.5 for matured drugs (Siegal et al., 2007). Thus a useful fragment hit should have at least a LE of 0.3 to become a drug obeying the *Rule of Five* (Congreve et al., 2008). A study by Hajduk (2006a) reveals that every added mass unit increases the affinity equally.

1.1.7 Earlier Studies Supporting the Fragment Idea

The first ideas of deriving information for rational drug design from energy information of functional groups binding to a protein target appeared already in the 1980s and early 1990s. In the beginning, the idea of the additivity of intrinsic binding energies was derived (Jencks, 1981; Andrews et al., 1984). The computer program MCSS (Multiple Copy Simultaneous Search; Miranker and Karplus, 1991) places functional groups in the active site of a protein. There are also some modelling approaches of linked-fragments, e.g. LUDI (Böhm, 1992) and CAVEAT (Lauri and Bartlett, 1994). Furthermore, GRID finds binding sites and estimates the binding energy using an interaction grid of various interactions at the surface (Goodford, 1985). SPROUT designs new molecules while adding functionalities to a primary structure which fits to a binding site (Gillet et al., 1993). HOOK builds up on information from MCSS to derive information about functional group sites and links them (Eisen et al., 1994). Complementary to MCSS, the experimental approach MSCS (Multiple-solvent Crystal Structure) was achieved by soaking organic solvents into crystals to map the protein surface (first by Allen et al., 1996; discussed in detail by Mattos and Ringe, 1996; English et al., 1999). These studies reveal that already in the 1980s, the binding of a compound was considered to result from a contribution of all its components.

The first practical approach to screening fragments dates back to the mid 1990s. Abbott Laboratories detected many weak binding compounds for FKBP (immunosuppressant FK506 binding protein) via NMR (Shuker et al., 1996). Two low micromolar fragments binding to adjacent pockets were optimised and linked to a potent inhibitor. This article is nowadays known as the beginning of Fragment Based Ligand Discovery (FBLD) (Figure 1.9).

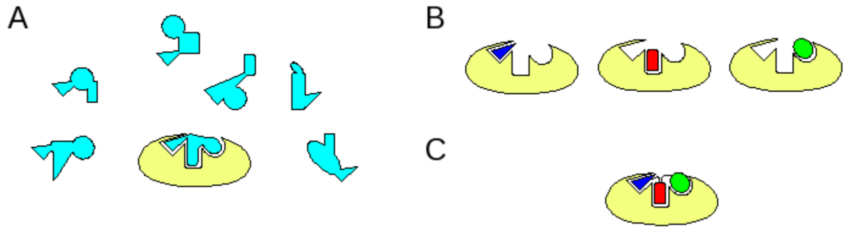


Figure 1.9: **HTS versus FBLD.** (A) A typical HTS library contains millions of complex scaffolds which are screened for fitting in the active site of a target protein. (B) Fragments are less complex and thus more likely to fit into subpockets of the active site. (C) After initial binding of fragments, the hits can be evolved into larger lead compounds

1.2 Fragment Based Ligand Discovery

Over the last fifteen years FBLD has come of age, with a series of compounds now entering the clinic. The main constraints are the need for a method that can reliably detect weak binding and strategies for evolving the fragments into larger lead compounds. Fragments detected binding to the active site can be evolved to larger compounds, either by linking or merging fragments together or by growing the fragments to pick additional interactions (Figure 1.9). In nearly all cases reported to date, this fragments evolution has relied on access to experimentally determined structures of fragments bound to the target either by X-ray crystallography or by high field NMR techniques (Jhoti, 2007a; Pellecchia et al., 2008).

1.2.1 Detecting Fragment Binding

Typically, fragments bind to a target with an affinity (K_D) in the 100 µM to 100 mM range. Detecting such weak binding is a challenge for most binding assays and routine fragment screening has relied on the development of a range of biophysical methods. The first published description of FBLD by the Abbott group was detected by the perturbation of the HSQC (Heteronuclear Single Quantum Correlation) spectrum of isotopically labelled protein (Shuker et al., 1996). Later, commonly used methods to detect binding were also ligand-observed NMR methods such as STD (Saturation Transfer Difference), crystallography, high concentration screening and mass spectrometry (Figure 1.10).

Two new methods have gained importance in recent years. The first is not yet widely published, but is increasingly used. The thermal shift method is based on monitoring the change in the temperature at which a protein unfolds during a binding event, by monitoring the increase in fluorescence from a dye that interacts into the protein as it is heated and unfolds (Kranz and Schalk-Hihi, 2011).

A more widely reported development is the increased use of surface plasmon resonance (SPR) where either the target or a ligand is attached to a surface whose optical properties change with molecular weight (Figure 1.10). Although this technique has been available for some time, the recent increase in use has come not only with improved sensitivity instruments, but also with increased experience of strategies for robust attachment of the target to the surface and a growing experience base for recognising artefacts (developments in Hämäläinen et al., 2008 and in Proll et al., 2009).

In addition, considerations of the kinetics of binding are gaining increased attention in the selection of compounds in drug discovery (Tummino and Copeland, 2008). It remains to be seen whether this kinetic characteristic is retained from a

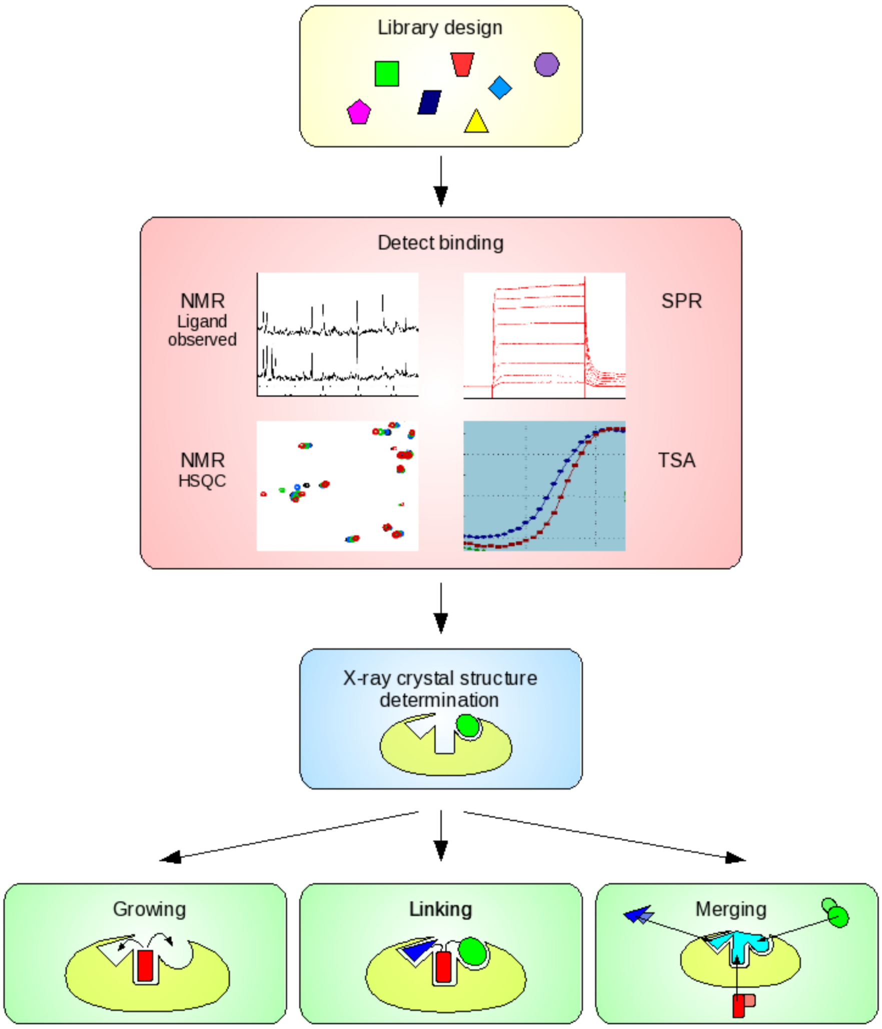


Figure 1.10: **Typical flow from the fragment to the drug.** The three essential elements of a fragment-based discovery platform are a library of suitable fragments, a method for identifying which fragments bind and a strategy for evolving the fragments to larger hits for optimisation to lead compounds.

core scaffold to the final drug candidate – this could then be an additional metric for the selection of fragments to progress.

Most practitioners are now converging on a common approach where a relatively high-throughput technique (ligand-monitoring NMR or SPR) is used to identify fragments that bind. These hits are often (particularly for challenging protein-protein interaction targets) cross-validated by another biophysical technique before being taken forwards for X-ray structure determination. Most are finding a hit rate between 2 and 10%) even with small libraries of only 1,000 or so fragments. This experience contrasts with HTS where 10^6 or more compounds often fail to provide suitable lead compounds.

Despite this success, there are continued efforts to develop improved methods for detecting and characterising fragment binding. A particularly striking method is the TINS approach where the target is immobilised on a resin and ligand-monitoring NMR signals measured (Siegal et al., 2007). The advantage is that low amounts of protein are required and that it may be suitable for membrane proteins. Calorimetric methods (Recht et al., 2008) may eventually gain the sensitivity and throughput for the screening of fragments and consideration of enthalpy/entropy characteristics may become an important mechanism for selecting which fragments to progress (Freire, 2008).

1.2.2 Evolving Fragments

The past few years have seen a rapid increase in the number of papers which describe advanced lead or clinical candidate compounds that have evolved from fragments. All, to date, have relied on structural information to guide optimisation. The three main strategies which can be identified for progressing fragments are linking, merging and growing (Figure 1.10). If more than one fragment is discovered and those bind to different parts of the active site these fragment can either be linked directly or via a suitable linker. The issue is to find the appropriate linking strategy which does not change any distances or angles of the firstly discovered fragments in order to achieve higher potency. This constraint makes that strategy rather challenging. Merging also requires more than one bound fragment. A catalogue of compounds is searched to find compounds which combine functionalities of all compounds in their original position. Growing works by substitution at one or more functional groups of the discovered fragment in order to add groups with additional binding capacity. The directions of substitution are called growth vectors. This method is usually successful and so is the most popular strategy. (Hubbard, 2008; Congreve et al., 2008)

The following paragraphs describe some recent examples.

The original SAR by NMR approach of Abbott (Hajduk, 2006b) relied upon finding suitable chemistry to link two separately identified fragments. There are two recent examples of this approach the Hsp90 programme from Abbott (Huth et al., 2007) and one of the approaches adopted in the discovery of PKB inhibitors at Astex (Saxty et al., 2007). The experience echoes that found by most practitioners, that it is difficult to find effective chemistry to link fragments together that retains the orientation and position of binding of the individual fragments and delivers the expected gain in potency. It remains to be seen whether the further development of computational tools such as CONFIRM (Connecting Fragments Found in Receptor Molecules) will help. A library is searched to find appropriate bridges for fragments binding to a target. The hits are automatically linked and docked to the target protein (Thompson et al., 2008).

A more successful method uses the structure of the fragment in the binding site to guide growth of the fragment. Here, two main ideas are evident. The first is where the structure of the fragment bound to the target is used as a substructure to direct an in silico search of available compounds directories. This method is known as SAR by catalogue (Figure 1.11). The powerful approach is in particular interesting for academic groups who have no access to a team of medicinal chemists but is also realised in Big Pharma where FBLD is integrated in parallel to conventional HTS. In some cases (as in the Hsp90 example form Vernalis; Brough et al., 2008) the compounds identified are further filtered by focussed docking to the target. This idea of using fragments as an initial screen or window into larger collection of compounds is becoming a feature of the continued integration of fragment screening alongside HTS in large pharmaceutical operations. An example is the virtual fragment linking approach at Novartis (Crisman et al., 2008) and there are yet unpublished reports of similar strategies being adopted at other companies.

The second growth method that has been widely reported as successful is where the detail of the structure of the fragment-target complex is used to identify how and where to grow or modify the fragments to increase potency or selectivity. Two very recent examples have been published. In a particularly elegant example, Astex synthesised a small number of compounds to progress rapidly from a weak fragment to a potent inhibitor of Aurora kinase (Howard et al., 2009). A similar approach at SGX, led to the rapid identification of a 78 nM JAK-2 kinase inhibitor

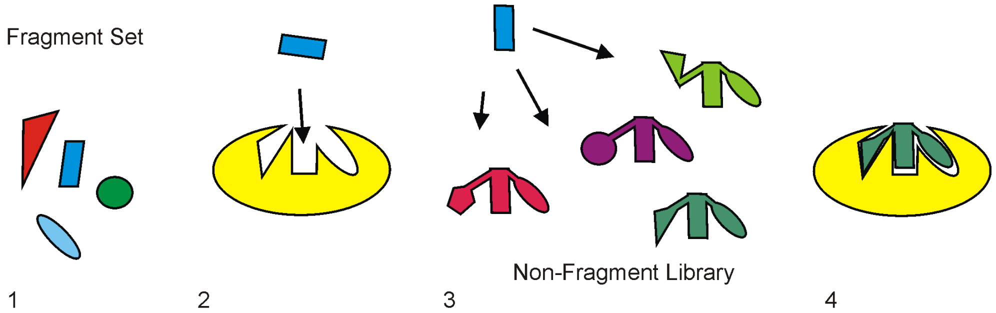


Figure 1.11: **SAR by catalogue.** (1) Screen a Fragment Set (2) Find a hit (3) Find nearest neighbours in a Non-Fragment Library (4) Find a stronger binding hit

(Antonysamy et al., 2009).

The final method of fragment evolution is that of merging structural information about fragments and known ligands. A recent example is the development of PDPK1 (formerly known as PDK1) inhibitors by Vernalis, where the structures of the fragments and initial hits from SAR by catalogue, were combined to generate potent inhibitors which showed activity on cells (Hubbard, 2008).

Many of the examples of FBLD from smaller structure-based companies have targeted kinases. There are historical and commercial reasons for this. The companies had been founded as kinases were identified from cancer genomics and there was seen to be a relatively low barrier to phase I trials in oncology. In addition, this class of proteins has a well-defined and druggable active site and are, in general, structurally tractable. However, there is also an increasing number of reports of fragments being successful on other target classes for application in the range of therapeutic areas. One reason is that fragment-based methods have been used in large organisations as a backup strategy for drug targets that had failed in HTS. The successes (particularly for protein-protein interactions such as the Bcl-2 family; Oltersdorf et al., 2005) have highlighted a major advantage of the fragment approach. Small compounds are more likely to bind into the target binding site than the large, decorated compounds found in most HTS collections. Although there can be challenges in determining crystal structures, many are now finding fragments as an attractive approach for generating initial hits against challenging targets. Recent examples from the literature are for β-secretase (Albert et al., 2007; Congreve et al., 2007), prostaglandin D synthase

(Hohwy et al., 2008) and hepatitis C virus NS5b RNA polymerase (Antonysamy et al., 2008). In addition, the approach can be used on nucleic acid targets, such as tRNALys3 to inhibit the HIV-1 reverse transcriptase initiation complex (Chung et al., 2007).

For well-behaved targets, it is relatively straightforward to identify many dozens of fragments that bind and in appropriate cases, generate crystal structures for many of the fragments bound to the target. One challenge is the selection of fragment(s) to progress. Fragment-based hit identification strategies are now being integrated alongside HTS in many large pharmaceutical organisations.

Nevertheless, the base of every success of fragment hit identification is the selection of an appropriate fragment library. This library must have a reasonable size of compounds – preferably less than 1,000 compounds. At the same time it should represent the chemical space as best as possible and also be as diverse as possible.

1.3 Fragment Library Design

The success of FBLD greatly depends on the fragment library. As in any screening process, the quality of the library dictates the quality of the hits found. Quality in this case means compounds covering a wide range of chemical space with no toxic or reactive groups and having the ability to be developed into drug-like compounds. For fragment screening, there are additional constraints placed by the screen at high concentrations and the need for appropriate synthetic routes to evolve the fragment into lead compounds. The available literature on fragment library design has been sparse during more than one decade of FBLD. However, this is changing with a recent upsurge in publications.

In 2003, Astex introduced the *Rule of Three for Fragments* (Ro3) (Congreve et al., 2003) which is based on Lipinski's *Rule of Five* (Ro5) for drug-like compounds (Lipinski et al., 2001, first in 1996). The Ro3 states that a fragment has:

- $<$ 300 Da MW
- $\leq$ 3 hydrogen bond donors

- ≤ 3 hydrogen bond acceptors
- ≤ 3 ClogP
- ≤ 3 number of rotatable bonds
- ≤ 60 Å^2 polar surface area

Many groups followed these guidelines to design their fragment library. Nevertheless, summaries by Brewer et al. (2008) and Law et al. (2009) reveal that this is mostly considered a rough guide. Often, hydrogen bond acceptors and number of rotatable bonds are higher (up to 8 hydrogen bond acceptors and up to 6 rotatable bonds). However, a fragment should possess a molecular weight of at least 150 Da (Babaoglu and Stoichet, 2006) to minimize reorientation of the bound fragment during its evolution. Very small fragments are also often found to bind in different orientations (Siegal et al., 2007); in addition, below about 120 Da, the fragments bind only at quite high concentrations and at many different sites, as seen in the Multiple Solvent Crystallographic Screen method, pioneered by Mattos and Ringe (1996) and English et al. (1999).

Most libraries are constructed from chemoinformatics pipelines that take available compounds and identify a representative subset that exclude reactive or toxic molecules (Verheij, 2006) and assess solubility and chemical diversity. One of the final steps is the rather subjective selection by medicinal chemists of compounds and identification of a representative subset. Although risk of bias it does ensure eventual hits will be progressed. Finally, rigorous quality control is required for validation and continued curation of the library. The variants to this library generation process include selection of privileged fragments from the analysis of known drug compounds, or generating target-focussed libraries by including a pharmacophore screen.

The following paragraphs give a historical perspective of fragment library design before moving on to give an overall overview of the aim and the chapters in this thesis. The different library design approaches are summarised under the key design strategies:

- Intense filtering, diversity and visual selection criteria

- Selection based on shapes and scaffold
- Libraries for special approaches
- Fragments based on known drugs
- Focussed fragment sets

1.3.1 Selection Based on Intense Filtering, Diversity and Visual Inspection

Most published procedures of fragment library generation are based on selection on physiochemical properties, the Ro3, intensive filtering of an input set to reduce size, diversity criteria and visual inspection. A similar approach seems to be applied by most Big Pharma. Several examples are listed below.

However, it is difficult to characterise and compare the quality of the different fragment libraries from different organisations as comparative screening data is not available. The only metric available is the reported hit rate; however that depends on the sensitivity and cut-off of the assay used. However, in general, it can be seen that the hit rate decreases as the average molecular weight of the libraries increases beyond 200 Da, which is as expected from the arguments of complexity of Hann et al. (2001).

Vernalis Library

A detailed description of fragment library generation is published by Vernalis (Baurin et al., 2004). For Vernalis, the requirements of a library are a 2 mM aqueous solubility and 200 mM solubility in DMSO for stocks. The compounds must be stable in stocks and experiments. Furthermore, the fragment library has to balance costs and practical issues with chemical diversity. The overall library comprises 1,300 compounds. The first steps were achieved by automated filtering. Unwanted groups such as anhydrides, aziridines or epoxides were removed. Then, the compounds were filtered by wanted functionalities, e.g. at least one ring of five or more members or one of specified functionalities such as a carboxylic group. Compounds were selected to be diverse and to contain appropriate physicochemical properties for the screening method, ligand-observed

NMR. Compounds should also be easily tractable for chemical evolution. However, the final compounds were selected by medicinal chemists based on visual inspection which introduced a subjective choice.

Astex Library

The screening process from Astex called Pyramid primarily uses X-ray crystallography (Hartshorn et al., 2005) to screen two complementary sets of fragments. Both sets contain relatively simple molecules of mostly between 100 and 250 Da. The first of the two sets is a drug fragment set. From known drugs, low molecular weight ring systems and simple carboxylic and heterocyclic ring systems were selected. The latter two were used for combination with side chains frequently occurring in drugs, lipophilic chains and a set of N-substituents to generate a virtual library. The library with 4,513 compounds was translated into SMILES (introduced in Section 2.2 on page 52) and searched for commercially available compounds. After manual inspection in order to remove toxic groups, a final set of 327 compounds was purchased. The second set comprised of fragments targeted against particular proteins and protein classes. The sets were constructed by a virtual screen. A database with 3.6 million compounds of chemical suppliers was filtered for desired physicochemical properties such as the Ro3. The obtained compounds were docked into several protein conformations to obtain a protein targeted set. The protein targeted sets were constructed using information of literature and patents. After further evaluation such as enumeration with known drugs, available compounds were purchased.

SGX Pharmaceuticals Library

SGX Pharmaceuticals (Blaney et al., 2006) selected their library based on the minimisation of molecular weight, CLogP (Chapter 2) and complexity. Furthermore, the fragments must be accessible by rapid synthetic optimisation. Therefore the compounds contain two to three chemical handles. A high proportion of their compounds contain bromide whose anomalous dispersion signal helps for structure validation when screening with X-ray crystallography. Compounds containing non drug-like properties (Hann, 1999) were excluded if there were no specific chemical handles. To maximise the diversity of their library the compounds were selected based on diverse ring system and 4-point pharmacophore analysis (Mason, 1999). They also included ring systems of known drugs which are contained in the MDDR (Elsevier MDL, 2004). Additionally, the compounds must be soluble in high concentrations to facilitate screening by X-ray crystallography. Finally, about 1,000 compounds were selected for the screening library.

Roche and GSK Libraries

Roche and GlaxoSmithKline are also using fragment screening in their lead discovery projects. However, not much is known about the design of their libraries. Roche selected their 2,000 strong fragment library based on the Ro3 and is screening with SPR (Hubbard et al., 2007). GSK has a set of compounds of reduced complexity of in-house and purchased compounds. The compounds are slightly bigger with a heavy atom count up to 21 atoms. The compounds were filtered for non lead-like properties and their diversity was determined based on 3D pharmacophore keys (Leach, 2006).

Evotec Library

Evotec has two separate fragment libraries which were constructed on different principles (Brewer et al., 2008). The first was built for high concentration screening (HCS) with biochemical assays, the second for NMR screening. Both libraries comprise about 20,000 compounds. The build-up of the HCS set started from a database of in-house and commercially available compounds. Toxic and reactive compounds were removed. Compounds were further filtered for a predicted aqueous solubility higher than 1 mM, followed by a filter for physicochemical properties such as number of hydrogen acceptors and molecular weight. The fragments were selected based on UNITY fingerprints (introduced in Section 2.3.1 on page 54) to obtain high diversity. All compounds were finally visually inspected by medicinal chemists. The selection for the NMR set originated from a collection of commercially available drug- and lead-like compounds. The collection was filtered based on equal representation of substructures. Toxic and reactive compounds were removed. The compounds were further filtered for molecular weight, CLogP and solubility and for physicochemical properties such as number of rotatable bonds (more in Section 2.4 on page 57). The fragments were then validated for bioactivity and the availability of analogue compounds and synthetic tractability was guaranteed. No unnecessary chemical restraints were introduced such as bromine containing compounds for X-ray screening. The size and diversity of the screened subset depends on the costs of the assayed protein. For costly proteins, a diverse subset of 5,000 compounds is selected using a chemical dissimilarity algorithm.

AstraZeneca Library

One year later, AstraZeneca published their approach (Blomberg et al., 2009) to design a 20 k generic library and a 1.2 k generic NMR screening library. Both libraries were assembled from the AZ in-house library and from vendor catalogues. The selection of compounds from these libraries followed the Core and Layer method (CaL). That procedure selects from pools of structures and uses decreasingly strict selection criteria. The first selection of compounds is based on substructural requirements as the number of heavy atoms. The next step applies practical properties of the sample (availability of solid and availability of analogues). If a structure has a high number of nearest neighbours it is likely to represent a prototype and to be synthetically easily accessible.

Pfizer Library

Recently, Pfizer described their approach to fragment library design (Lau et al., 2011) as three stages consisting of filtering corporate and commercial libraries for physicochemical properties and unwanted groups, applying diversity criteria and finally a visual inspection of the compounds.

BioFocus Library

The BioFocus fragment library was selected based on the Ro3 (Pollack et al., 2011) whereof compounds with unwanted functionalities were removed. The library was assessed for coverage of the chemical space and for diversity. 190,000 compounds from a the ChEMBL database containing bioactive drug-like small molecules were extracted based on the binding constant to their target. Compounds were condensed into 1,500 clusters and the coverage with the fragment library was assessed based on FCFP_6 fingerprints. 52% of the BioFocus library were represented as substructures in the ChEMBL set and in total 1,308 clusters were covered.

1.3.2 Selection Based on Shapes and Scaffolds

Another big group of common library design strategies is based on a shape or scaffold selection.

Vertex Library

One of the first descriptions of how to design such a library was Vertex SHAPES method (Fejzo et al., 1999). Their collection was derived from shapes which are most commonly present in known drugs. The aim of their library for NMR screening was the assembling of small molecules that optimised many factors at the same time such as solubility, cost, synthetic tractability diversity and separation of NMR peaks. Known therapeutics were decomposed into substructures consisting of side chains, linkers and rings. Linkers and rings together were handled as frameworks. An analysis of the CMC database (Comprehensive Medicinal Chemistry, MDL Informatics Systems) showed that only 41 of these frameworks including atom type and bond order described about 24% of all known drugs. They combined those 41 frameworks with the 30 most common drug side chains. Those compounds were used as templates for a substructure search of the ACD database (Available Chemicals Directory, MDL Informatics Systems). The results were further filtered for beforehand specified side chains, solubility and inherent synthetic complexity or for representation of frequently occurring classes in the CMC but not being within the 41 frameworks. The resulting compounds were commercially available, soluble, pure and not reactive.

Plexxikon Library

The scaffold-library by Plexxikon started with the selection of compounds in the range of 120–350 Da from 17 different suppliers (Card et al., 2005). Compounds with reactive groups were removed. The remaining compounds were fragmented into smaller substructures through cutting of rotatable bonds. These substructures were then classified by chemical similarity. All compounds with > 0.85 Tanimoto similarity (more about similarity coefficients in Section 2.5 on page 58) to another compound were removed. The process resulted in a fragment library of 20,360 compounds which represent about 80% of the scaffold space.

ZoBio and Pyxis Discovery Libraries

The compounds of the fragment library of ZoBio and Pyxis Discovery (Siegal et al., 2007) obey the Ro3 and four other themes: diversity guaranteed by the scaffold-based classification approach (SCA) and by shape, amino acid derivates and scaffolds present in natural products. Out of a pool of 70,000 fragments, the first 500 fragments were selected for maximal diversity with the SCA approach. For this purpose, molecules were fragmented in their ring systems and side chains. With four descriptors (maximum number of smallest set of smallest rings, number of heavy atoms, sum of heavy atomic numbers, number of bonds) the complexity of the scaffolds was represented. The ratio of ring atoms to side chains gave a cyclicity score. For the amino acid theme, the motifs of the 20 amino acids connected to small rings were used. Scaffolds present in orally available drugs and commercially available natural products were used as templates to search representative molecules in the 70,000 fragment library to represent the natural products theme. The selected fragments were compared with the whole pool and 500 fragments were selected to optimise the overall shape of the library. The resulting final fragment library contained about 2000 compounds.

Broad Institute Library

Another article in 2011 describes the application of diversity-oriented synthesis to fragment library design in order to create more three-dimensional fragments (Hung et al., 2011). A fragment set with more sp^3-rich compounds covers a bigger part of the chemical space. In addition, more three dimensional compounds have more suitable growth vectors to grow into sites otherwise possibly inaccessible. The synthesised compounds further possess chemical handles. They do not only facilitate chemical optimisation, but also improve solubility and binding potency due to more hydrogen bond donors and acceptors.

1.3.3 Specifically Purposed Fragment Libraries

Some companies have developed distinctive strategies for fragment screening and evolution that require specialised libraries.

Sunesis Pharmaceuticals Library

In 2000, Sunesis Pharmaceuticals Inc. published their "tethering" approach of FBLD (Erlanson et al., 2000). Binders are discovered by the formation of a disulphide bond between the ligand and a cysteine residue of the target. Thus, the Sunesis' fragment library comprises disulphide containing compounds and a library of about 1,200 members was synthesised. The library are screened in pools by mass spectrometry.

Graffinity Pharmaceuticals Library

Graffinity Pharmaceuticals published their approach by SPR imaging (Neumann et al., 2007). The possibilities with SPR allowed the company to design a diverse library of 20,000 fragments. The Graffinity approach is based on coupling fragments onto gold chips. Therefore compounds were synthesised on long linkers with a thiol group at the terminus which is for the covalent coupling.

Novartis Library

Novartis published the design of a fluorinated fragment library which they are using for their NMR approach (Vulpetti et al., 2009).

1.3.4 Fragments Based on Known Drugs

A newer academic approach is based on known drugs. Gianti et al. used Pipeline Pilot to implement protocols that identify substructures of known drugs and drug-like virtual screening sets. The resulting virtual library was used as template to generate a privileged fragment library which was further filtered for desired physicochemical properties and unwanted functional groups. The procedure led to a final collection of 29,500 compounds (Gianti and Sartori, 2008).

1.3.5 Focussed Fragment Libraries

More recently, smaller fragment libraries for individual targets or classes or targets appeared by academic groups. A retrospective analysis of kinase inhibitors helped to design new kinase focussed libraries (Akritopulou-Zanze and Hajduk, 2009). Analyses of this type will be used by others to design novel fragments to increase intellectual property freedom or increase coverage of particular chemotypes. One example is the synthesis of a focussed fragment library of nine compounds used to discover an inhibitor of Mcl-1 and Bcl-XL (Prakesch et al., 2008). Agrawal et al. (2010) designed a library of small molecule chelators for the screening of metalloproteins with promising success. Chelators demonstrate suitable binding affinities and good platforms for optimisation.

1.3.6 Relevance of the Rule of Three

The *Rule of Three* is widely quoted as the criteria for designing a fragment library. The phrase attracted much attention, as it echoes the *Rule of Five* from Lipinski et al. that had a real impact on medicinal chemistry practices in the early 2000s. However, the *Rule of Three* is not useful and has attracted some criticism. For example Köster et al. (2011) from the Klebe group assembled a fragment library that did not conform to the Ro3 and screened it against endothiapepsin. Only four of their eleven hits obeyed the *Rule of Three*.

In general, the trend of recent years (following the chemical space analysis of Reymond et al., 2011) is for the average molecular weight of libraries to fall - many have an average molecular weight of around 200 Da. It is the physical properties of the compounds and their potential for chemical evolution that is more important.

1.3.7 Conclusion and Thesis Outlook

Naturally there are many more fragment libraries in use but not published. The section of the published fragment libraries summarised above reveals that although characteristics are known and discussed, a simple and straightforward guide to library design is often missing. There is significant bias in the selection of compounds with the constraints of the assay conditions as a major consideration. To date, there has been no description of a library which has been designed to represent the available chemical space. If the members of a fragment library are representative substructures of the chemical space (or respectively all available compounds), a superstructure search can be a quick way of fragment evolution. The first part of this thesis targets this problem and introduces a number of protocols for fragment library generation (Chapter 3 on page 62).

1.4 Aims in The Thesis

The work presented in this thesis is in three parts.

The first project was to implement an automated procedure to design and to validate a fragment library that maximally represents superstructures and pharmacophores. The aim is to have a fragment library from which hit fragments can be readily evolved by purchase from available chemical databases. Such methods will be of use to academic groups embarking on fragment based approaches to identify tools for chemical biology. In addition, such methods may be of value to institutions or companies which maintain large physical libraries of compounds, where the fragment library can be designed to maximally represent the compounds available.

The second part of the work was to test the fragment library on a number of protein targets. The main focus of screening methods was on Thermal Shift Analysis which led to the implementation of a new analysis program MTSA, developed to improve the quality and ease of analysis of results generated by that technique. The program MTSA and further analysis of the thermal shift technique represent the third theme of this thesis.

The protein targets were N-myristoyl transferase (NMT), Hen egg white lysozyme

(HEWL) and the GlcNACase enzyme, BtGH84. NMT is an enzyme established by Prof. D. Smith (Biology) as a potential drug target for a number of tropical diseases (in collaboration with Prof. Tony Wilkinson and Dr. Jim Brannigan at the University of York, various scientists at Imperial College, Dundee University and the National Institute for Medical Research NIMR and with some input from the pharmaceutical company, Pfizer). However, it proved difficult to generate stable NMT for fragment screening. In an attempt to work with a more tractable target (the protein can be purchased and it crystallises readily), some work was performed with HEWL, for which a lot of literature is available. However, screening experiments with the enzyme were not reproducible. Another carbohydrate binding target, BtGH84, was under study by a visiting post-doctoral fellow (Dr. Jens Landström) and following his initial work, further screening and analysis was performed. Many compounds were screened, and inhibitors, inhibitor enhancers and what appear to be activity enhancers were obtained.

The organisation of the thesis is as follows:

Chapter 2 introduces relevant chemoinformatics methods.

Chapter 3 describes the developed fragment library protocols and their evaluation.

Chapter 4 introduces the thermal shift analysis and explanation of the implemented program MTSA.

Chapter 5 gives an overview about the protein target NMT which was subcloned to attach a double His$_6$ tag and which was further tested for fragment screening suitability.

Chapter 6 deals with HEWL for which several screening techniques were set up, but delivered inconsistent results.

Chapter 7 summarises the results of screening 500 compounds against BtGH84 with thermal shift analysis with cross validation by surface plasmon resonance and initial crystallisation trials.

Chapter 8 discusses the melting point definition used for thermal shift analysis and questions the suitability of the method to screen low affinity binders with that method.

Chapter 9 summarises and discusses the previous chapters and gives and outlook for future work.

Chapter 10 lists materials and methods deployed in this work.

Chapter 2

Introduction to Chemoinformatics

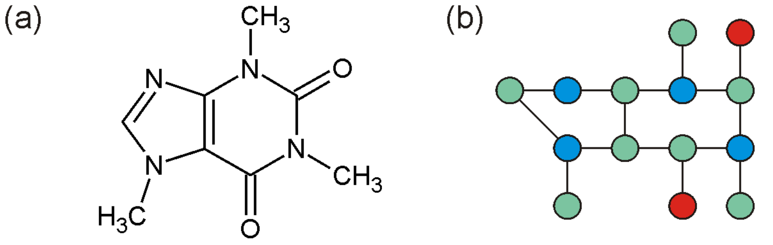
The fragment library design part of this project (Chapter 3, page 62) relies on a range of chemoinformatics methods. This chapter introduces computational handling of chemical structures, description of physiochemical properties and similarity methods are introduced.

2.1 Chemoinformatics

Although many of the methods and ideas had been in use for decades, the term *Chemoinformatics* was first introduced by Dr. Frank Brown in 1998. He described it as "the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization." (Brown, 1998) The term is also known as *cheminformatics*, *chemical informatics* or *chemiinformatics*. The phrase is used widely and the discipline needs to be distinguished from computational chemistry. The main focus in computational chemistry is constructing a model of a molecular system (at the orbital, atomic or molecular level) from which properties of the system can be calculated. In contrast, chemoinformatics is the handling of all sorts of information about molecules. This may be derived from computation but also covers everything from experimental data to where the sample is stored. There are three main aspects to chemoinformatics: the different representations of chemical structure and handling, the calculation and representation of molecular properties and the informatics methods that are used to store, search and analyse the structures and properties (Engel 2006, Chen 2006, Leach and Gillet, 2007). The following sections provide an introduction and summary of these different aspects.

2.2 Representation of 2D Structures

There are many possible ways to computationally represent chemical structures. One can imagine an image file (Figure 2.1 a) or the IUPAC name (International Union of Pure and Applied Chemistry) (Figure 2.1 c). However, to search substructures and to perform calculations, these notations are not very helpful. The most sensible way to approach that problem is by graph theory: A graph contains nodes (corresponding to the atoms) connected by edges (bonds). The nodes and edges can have additional properties. For a chemical structure that could be a certain atom type or the type of bond (Figure 2.1 b). The problem of whether two graphs are the same is known as “graph isomorphism”. The information of a molecular graph needs to be communicated to and from the computer. One way is via so called connection tables. These tables may include atoms, bonds, coordinates and more. Connection tables can be included in SD files (structure data files) which were developed by MDL Information Systems (Dalby et al., 1992) and contain information about atoms, bonds, connectivity and coordinates (Figure 2.2). Alternatively, linear notations can be used. One of the earliest linear representations was the Wiswesser Line Notation (WLN) (Wiswesser, 1954). However WLN was surpassed by the more recent SMILES notation (Simplified Molecular Input Line Entry Specification) (Weininger, 1988). In SMILES, the atoms of a molecule are represented as letters of their atomic symbol. Upper and lower case letters express aliphatic and aromatics atoms respectively. Several other symbols describe bonds and their character, branches are indicated with brackets (Figure 2.1d). For example C stands for methane and $CC(=O)O$ stands for acetic acid. In most cases, many different ways of representation exist for one and the same molecule, however for computational handling a canonical (uniform) representation is necessary. In contrast to SMILES, InChI (IUPAC International Chemical Identifier) produces a unique string for every structure that is based on an extremely formalized version of IUPAC names (www.iupac.org/projects/2000/2000-025-1-800.html).



(c) 1,3,7-trimethyl-3,7-dihydro-1*H*-purine-2,6-dione

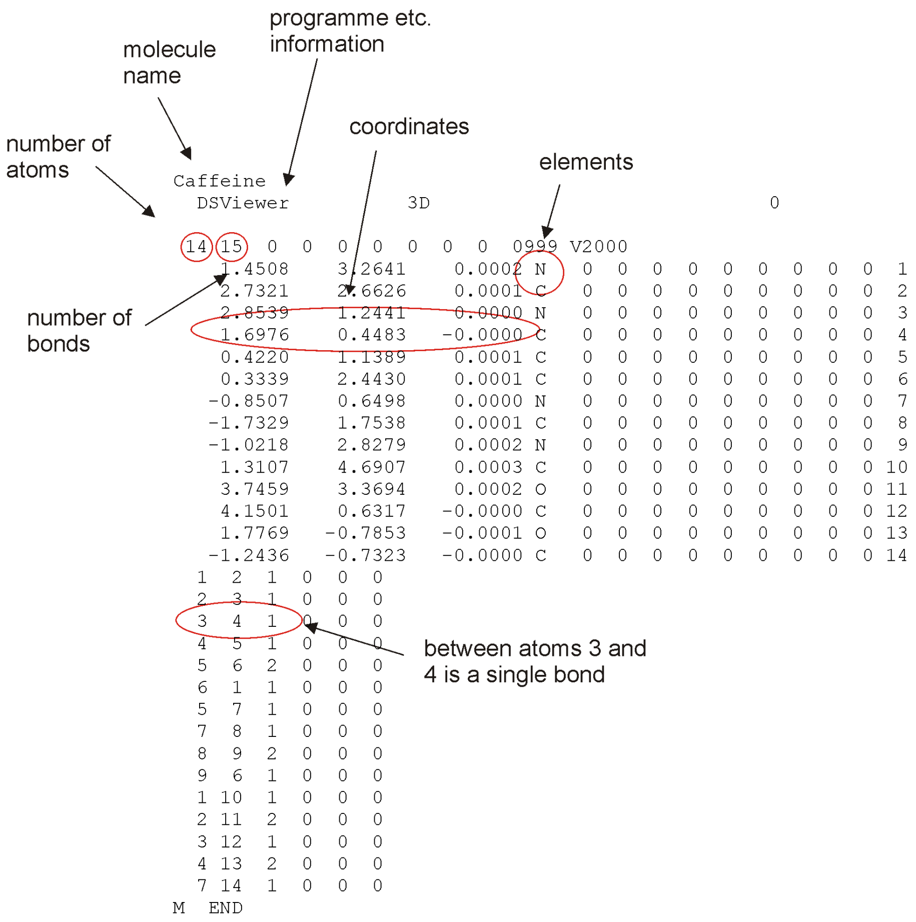
(d) O=C2c1n(C)cnc1N(C)C(=O)N2C

(e) 1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Figure 2.1: **Caffeine in different representations.** The molecule caffeine is represented as (a) an image file, (b) as a graph, (c) with its IUPAC name, (d) in SMILES and (e) in InChI.

2.3 Searching Structures

Databases are used to store, manipulate and search molecules and additional information such as name, physicochemical properties, assay data etc. The simplest task is to search a complete molecule in the database. The wanted molecule is translated into a canonical representation and the database is either simply searched from the beginning or with the help of a hash key which directs straight to the physical location on the hard drive. A hash key is a short integer which is generated of a longer string by a hash function. With hashing, data can be indexed and retrieved faster because the hash key is shorter than the original value. More difficult is the search for substructures. A complex query can be transmitted in SMARTS (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html) which is the query extension of SMILES. Substructure searching can be performed with graph theoretical methods. For large databases, a two-step mechanism is mostly applied to make searches fast and efficient. First, the structures which cannot match the query are eliminated (ideally at least 99% of all



```
Caffeine
  DSViewer          3D                              0

 14 15  0  0  0  0  0  0  0  0999 V2000
    1.4508    3.2641    0.0002 N   0  0  0  0  0  0  0  0  0  1
    2.7321    2.6626    0.0001 C   0  0  0  0  0  0  0  0  0  2
    2.8539    1.2441    0.0000 N   0  0  0  0  0  0  0  0  0  3
    1.6976    0.4483   -0.0000 C   0  0  0  0  0  0  0  0  0  4
    0.4220    1.1389    0.0001 C   0  0  0  0  0  0  0  0  0  5
    0.3339    2.4430    0.0001 C   0  0  0  0  0  0  0  0  0  6
   -0.8507    0.6498    0.0000 N   0  0  0  0  0  0  0  0  0  7
   -1.7329    1.7538    0.0001 C   0  0  0  0  0  0  0  0  0  8
   -1.0218    2.8279    0.0002 N   0  0  0  0  0  0  0  0  0  9
    1.3107    4.6907    0.0003 C   0  0  0  0  0  0  0  0  0 10
    3.7459    3.3694    0.0002 O   0  0  0  0  0  0  0  0  0 11
    4.1501    0.6317   -0.0000 C   0  0  0  0  0  0  0  0  0 12
    1.7769   -0.7853   -0.0001 O   0  0  0  0  0  0  0  0  0 13
   -1.2436   -0.7323   -0.0000 C   0  0  0  0  0  0  0  0  0 14
  1  2  1  0  0  0
  2  3  1  0  0  0
  3  4  1  0  0  0
  4  5  1  0  0  0
  5  6  2  0  0  0
  6  1  1  0  0  0
  5  7  1  0  0  0
  7  8  1  0  0  0
  8  9  2  0  0  0
  9  6  1  0  0  0
  1 10  1  0  0  0
  2 11  2  0  0  0
  3 12  1  0  0  0
  4 13  2  0  0  0
  7 14  1  0  0  0
M  END
```

programme etc. information
molecule name
number of atoms
coordinates
elements
number of bonds
between atoms 3 and 4 is a single bond

Figure 2.2: **Caffeine as an sd file.** The structure data file of caffeine contains a connection table. The individual features are highlighted and annotated.

molecules). Secondly and more computationally intense, a subgraph isomorphism algorithm (the most known being the adjacency matrix based algorithm by Ullmann, 1976) is performed to identify if the substructures are the same.

2.3.1 Fingerprints

The first step of database screening is usually performed using a bit string representation of the molecules where a "1" stands for the presence of a certain feature, e.g. a structural motif or a certain element, and a "0" for its absence (Figure 2.3). These bit strings are called fingerprints. These fingerprints are also used for similarity and clustering methods. There are two most commonly used types of bit strings: structural key fingerprints and hashed fingerprints.

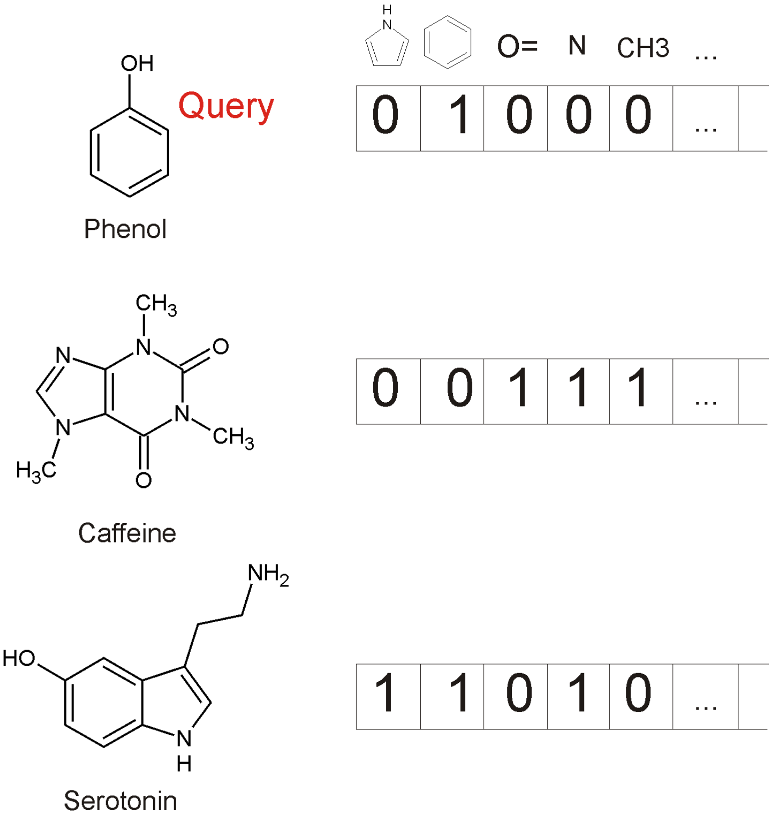


Figure 2.3: **Example of a bit string screen.** In this example phenol is the query. A possible bit string representation is shown on the right. If caffeine and serotonin are entries in the database, it is impossible for caffeine to match because the bit set in the query bit string is not set in the caffeine bit string. However it is set in serotonin which would thus pass the initial screening filter.

Structural Key Fingerprints

The position of the key in a structural key fingerprint corresponds to the presence or the absence of a certain structural motif such as “oxygen” or “at least one ring of size 6”. The quality of that method depends on the accuracy of the structural motif library. Some structural key fingerprints use a predefined library of structural motifs like for example the MDL (Molecular Design Limited) keys. MDL offers a limited amount of public keys which are released to the public and a bigger number of private keys which are proprietary. The 166 public keys are referred to as MDL Public Keys. MDL structural keys are developed for rapid substructure searching in smaller to medium-sized databases.

Hashed Fingerprints

The second method is the use of so called hashed fingerprints which do not need a predefined library. Thus they are universally applicable. Path based hashed fingerprints generate all possible paths of connected atoms in the molecule up to a defined length. For example the molecule OC=CN has the paths C, O and N of length zero, OC, C=C and CN of length one, OC=C, C=CN of length two and OC=CN of length three. All these paths are used to set bits in a bit string. They are hashed which also means that every bit in the final fingerprint can set by more than one feature. This collision can result in a higher number of false positives when screening. These paths are hashed to set the final bit string. Hashed fingerprints can result in a higher number of molecules to be searched with the slow graph isomorphism algorithm since the hashing method allows the same bit to be set by different features resulting in a collision. The major representative hashed fingerprints are those from Daylight Chemical Information Systems Inc. (Daylight).

Other common classes of hashed fingerprints are the Extended Connectivity Fingerprints (ECFP) and its variant the Functional Class Fingerprints (FCFP) (SciTegic) which use initial atom identifiers for atoms and their environment (Rogers and Hahn, 2010). Both are very effective in similarity searches. ECFP is based on atom types and FCFP on generalised types (where for example all halogens appear as equivalent). ECFPs and FCFPS represent a much larger set of features than substructure based MDL fingerprints. They also can be processed quicker because they are fast to calculate. ECPFs and FCFPs are generated in a similar way. The only difference lays in the initial assignment step for each atom of the molecule. The initial assignment of an ECFP gives each atom of the molecule an identifier which is based on information of the atom type, atomic mass, functional group etc. For the FCFPs the functional class of the atom serves for the initial identifier. The assignment generates an initial code which is hashed into a single 32-bit integer value which corresponds to the initial atom identifier. From there the extended connectivity code calculates the next generations and removes duplicates. The first iteration considers the next neighbours for each atom and updates the code. For the second iteration the next sphere of neighbours is considered and so on up to the maximum diameter.

Sometimes a combination of structural keys and hashed fingerprints is employed, for example in the UNITY system (Unity).

The following chapter uses clustering and similarity searches with the program Pipeline Pilot (Accelrys), where the default fingerprint FCFP_4 (Functional class extended-connectivity fingerprint of maximum diameter 4) was used.

2.4 Molecular Descriptors

To manipulate and analyse chemical structures, a diverse range of molecular descriptors is used. The next two sections describe the most important descriptors which were used for the subsequent fragment library generation in Chapter 3.

2.4.1 Property Count

The simple property count counts, for example, the occurrence of a certain element or hydrogen bond acceptors and donors. In the following, hydrogen bond acceptors are defined as oxygen, nitrogen, sulphur or phosphorus with one or more ion pairs. However, some groups are excluded such as: positively charged atoms, amides and pyrrole-type nitrogens as well as aromatic oxygen and sulphur in heterocyclic rings. The hydrogen donors in this thesis are defined as oxygen, nitrogen, sulphur or phosphorus when they have at least one hydrogen atom attached.

2.4.2 Physicochemical Properties

One of the most important physicochemical properties is the partition coefficient between n-octanol and water: P. Usually, P will be expressed in its logarithmic form $LogP$. The reason why $LogP$ is so important when working with drugs is that drugs need to be water soluble enough to be transported to the cells, but also hydrophobic enough to pass through the cell membrane. Commonly used algorithms to determine $LogP$ are $CLogP$ (a program developed by Leo and Hansch, 1993) and $ALogP$ by Ghose and Crippen (Ghose and Crippen, 1986; Ghose and Crippen, 1987; Ghose et al., 1988; Viswanadhan et al., 1989). $CLogP$

works on fragmentation of the molecule into parts of which the solubility is known whereas *ALogP* is atom based.

The next important physicochemical parameter is the aqueous solubility S which also influences the bioavailability of drugs in the body. In this thesis, it is calculated with the help of a multiple linear regression model based on electrotopological state (E-state) indices (Hall et al., 1991; Huuskonen, 1999; Tetko et al., 2001). The index includes information about the electronic state of the atoms influenced by the other atoms of the molecule in the context of the molecular skeleton (Hall and Kier, 1995). The molecular solubility is expressed as *LogS* with S being in the units mol/l.

Also the surface area and volume of a molecule are used as descriptors for fragments and non-fragments. Here, the total surface is calculated using a 2D approximation.

2.5 Similarity Coefficients

A search of a database against a defined query or similar compounds requires a method for calculating similarity (SIM). The searches are usually based on 2D fingerprints. Many different similarity and distance measures for binary fingerprints are available. If a are the bits set to 1 in molecule A, b the bits set to 1 in molecule B and c the bits set common in A and B and n is the total number of properties, then the formula for the binary variables of the commonly used Tanimoto coefficient (Jaccard, 1901; Tanimoto, 1957) is:

Tanimoto (Jaccard) coefficient $SIM_{AB} = \frac{c}{a+b-c}$

The Tanimoto coefficient compares all features (bits) present in two molecules. Because smaller compounds contain less features (and thus have less bits set to 1), the Tanimoto coefficient automatically punishes smaller compounds when compared to larger ones. A fragment which is fully included in another compound (i.e. being the exact substructure of another molecule) would not appear as 100% similar. This is known as an asymmetric problem. The solution which is applied in this thesis is the use of an asymmetric similarity coefficient, the Tversky index (Tversky, 1977; Bradshaw, 1997).

2.5.1 Tversky Coefficient

The general definition of the Tversky coefficient is given in Equation 2.1.

$$SIM_{Tversky} = \frac{c}{c + \beta * (a - c) + \gamma * (b - c)} \tag{2.1}$$

where β and γ are additional weighing factors. If γ is set to 0 and β is to 1, features only present in the superstructure compound are not taken into account and leads to the equation:

$$SIM_{Tversky} = \frac{c}{a} \tag{2.2}$$

The Equation 2.2 now represents the formula of a superstructure search. This means a fragment is 100% similar to another compound if the latter fully includes the fragment how graphically represented in Figure 2.4.

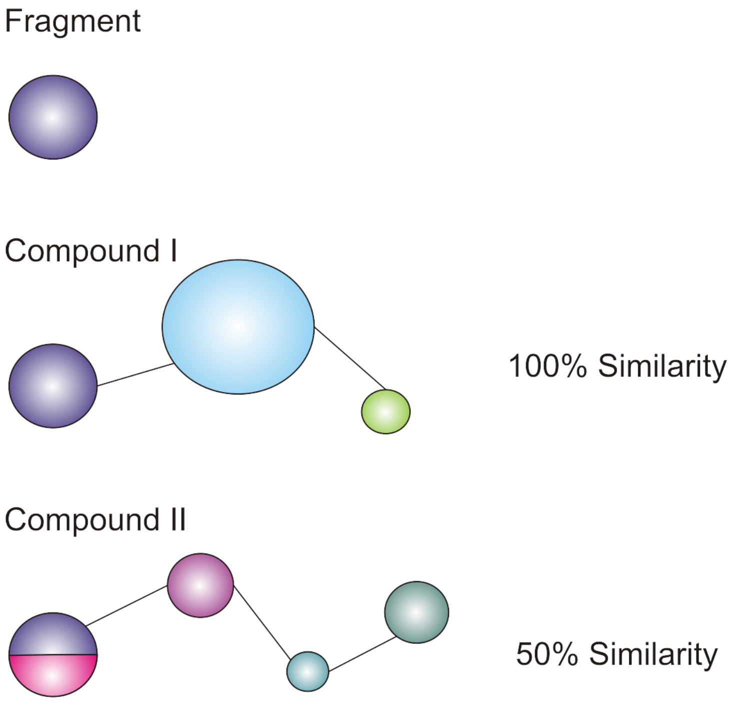


Figure 2.4: **The Tversky coefficient.** The Tversky coefficient produces a 100% similarity if the fragment is fully included in the bigger reference compound. It results in a 50% similarity if only half of the fragment is included. Thus, it corresponds to a superstructure search.

However, setting both weighing factors to 1 ($\beta = 1$ and $\gamma = 1$) results in the Tanimoto coefficient (Jaccard, 1901; Tanimoto, 1957) which is used to compare molecules of similar size. Equation 2.3 represents a similarity between two structures.

$$SIM_{Tanimoto} = \frac{c}{a + b - c} \tag{2.3}$$

2.6 Clustering

Clustering itself is widely used in many disciplines. There are many different approaches to clustering which can be grouped into hierarchical and non-hierarchical clustering. Hierarchical methods divide a set into ever-smaller regions or merged sets as they move up the hierarchy. To rapidly cluster large data sets, the program used in the following chapter (Pipeline Pilot) uses a partitioning method which is non-hierarchical. A set of compounds is divided into ever smaller subsets based on a maximum dissimilarity method: A random molecule is chosen and named the cluster centre. The most distant molecule to this one is the next cluster centre. The next cluster centre is formed by the molecule the most dissimilar to both chosen cluster centres and so forth until sufficient cluster centres are selected. The remaining molecules are assembled around these cluster centres. The Pipeline Pilot cluster component is based on the Tanimoto coefficient and the default and employed fingerprint was FCFP_4.

2.7 The Program Pipeline Pilot

Pipeline Pilot is a scientific program which offers preimplemented computational modules in order to create work flows for automated processes such as analysing and reporting data. The program allows scientists to focus on innovation rather than programming tedious but known tasks. There are many possibilities to customise parameters of the components or to script in the program's own language Pipeline Pilot Script. All protocols presented in the following chapter are implemented with the professional edition of the program Pipeline Pilot from Accelrys. The major reason for using the professional edition was that it runs on a server

- resulting in much faster calculations. Furthermore, the professional edition has additional features such as the "diverse molecules" component. However, the aim was to generate protocols which also work with the free academic edition. Thus, there exists a simplified version of each protocol for the free academic version.

Chapter 3

Fragment Library Design

This chapter describes an approach used to identify a set of fragments that maximally represent a number of available compounds and therefore is a Fragment Set suitable for the SAR by catalogue approach. Input Libraries of compounds from suppliers are filtered to remove unwanted functionality, duplicates and salts and then split into Fragment and Non-Fragment Libraries. A number of different chemoinformatics approaches are used to identify Fragment Sets that contain compounds from the Fragment Library that maximally represent the chemical space of the compounds in the Non-Fragment Library. The characteristics of the compounds in the final Fragment Sets are compared to determine the best selection procedure. All protocols were implemented with the program Pipeline Pilot (Section 2.7 on page 60). Both the professional edition and the student edition of the program were used during implementation.

3.1 Input Libraries

Compounds were downloaded from the ZINC database (Irwin and Stoichet, 2005) (http://zinc.docking.org/) for three different suppliers to guarantee that the determination of the best performing Fragment Set selection is not biased in favour of a certain supplier. Single representation for each molecule of 600,220 compounds from Asinex (http://www.asinex.com), 79,228 from Maybridge (http://www.maybridge.com) and 321,371 from Specs (http://www.specs.net) were used as Input Libraries.

All of the following protocols were implemented with Pipeline Pilot. Clustering and similarity were performed using FCFP_4 fingerprints. (Refer to Chapter 2, page 51 for an introduction.)

3.2 Separation into Fragments and Non-Fragments

The Input Libraries downloaded from the websites named above were divided into Fragments and Non-Fragments as an initial preparation for Fragment Set selection. The Input Library corresponds to the chemical space and the Fragment Library to the fragment space.

Figure 3.1 summarises the process of the separation into Fragments and Non-Fragments. Duplicates were removed based on canonical SMILES (Section 2.2, page 52), with every molecule following the first occurrence being discarded. Further, all salts were removed. In addition, SMARTS strings (Section 2.3, page 53) listed in Table A.1 (on page 188) were used to identify compounds for removal that contain unwanted chemical functionality. The SMARTS filter contained 68 queries based on the papers of Baurin et al. (2004) and Verheij (2006).

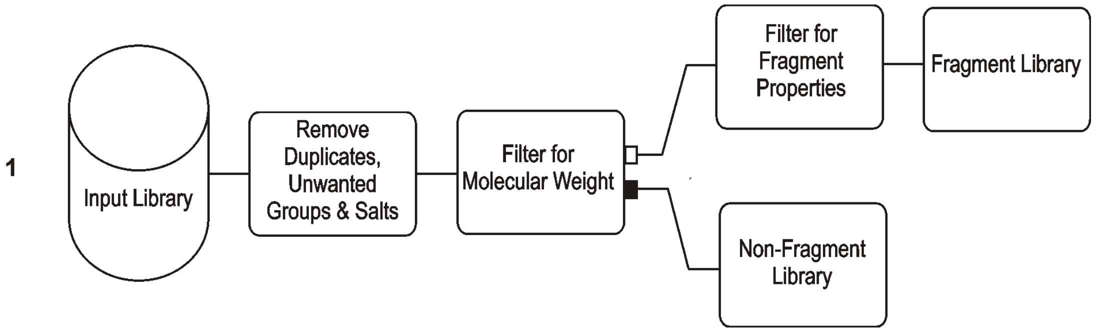

1
Input Library
Remove Duplicates, Unwanted Groups & Salts
Filter for Molecular Weight
Filter for Fragment Properties
Fragment Library
Non-Fragment Library

Figure 3.1: **Separation into libraries of Fragments and Non-Fragments.** After removal of duplicates and salts (not shown as separate step), the Input Library is filtered for unwanted groups defined in a SMARTS file. The compounds are further filtered for molecular weight; all compounds >250 Da (>300 Da if containing Sulphur) are written to the Non-Fragment Library. The smaller compounds are further filtered for other fragment-like properties and then written to the Fragment Library.

Several steps were applied to further select compounds of the Input Library after removing duplicates, salts and molecules containing unwanted groups. The number of compounds resulting from each of those steps is summarised in Table 3.1 with a detailed overview in Table 3.2. All compounds with a molecular weight MW $>$ 250 Da ($>$ 300 Da if sulphur is present) were assigned to a so-called Non-Fragment Library. The correspondingly named Fragment Library contains all compounds with an MW $\leq$ 250 Da ($\leq$ 300 Da if sulphur containing) excluding molecules which did not satisfy the criteria of an extended *Rule of Three* (Ro3) filter (page 38 for a definition). In an earlier test with the Asinex compounds, use of the original criteria of the Ro3 had removed a very high number of pre-fragments: Only 10% pre-fragments were selected for the Fragment Library. This is consistent with the criticism of the Ro3 mentioned by others (Köster et al., 2011, Section 1.3.6, page 47).

The extended Ro3 used to select fragments (with differences to the original Ro3 criteria in brackets if different) was:

- Atom count $>$ 0
- 100 $\leq$ molecular weight $\leq$ 250 Da (300 Da if sulphur) ($<$ 300 Da)
- N-count + O-count (hydrogen acceptors) $\leq$ 6 ($\leq$ 3)
- Hydrogen donors $\leq$ 3
- ALogP $\leq$ 3 (CLogP $\leq$ 3)
- Number rotatable bonds $\leq$ 5 ($\leq$ 3)
- Molecular polar surface area $\leq$ 80 Å^2 ($\leq$ 60 Å^2)
- Molecular solubility (calculated as LogS) $\geq$ -3.3 (not used)

3.2.1 Determination of the Solubility Threshold

The addition of the *Molecular Solubility* to the criteria is a strong selection filter which tends to remove many compounds (Section 2.4, page 57). The *Molecular*

Table 3.1: **Input and output size of Libraries** The numbers and percentages of compounds remaining during preparation of Fragment and Non-Fragment Libraries from the Input Libraries. *Percentage calculated relative to number passing the MW filter.

Input Library	**Asinex**	**Maybridge**	**Specs**
Initially from Zinc database	600,220	79,228	321,371
– duplicates removed	5,330 (1%)	2,322 (3%)	3,965 (1%)
– salts removed	0 (0%)	0 (0%)	0 (0%)
– unwanted functionality removed	317,755 (53%)	38,812 (49%)	196,331 (62%)
MW <250 Da (<300 Da with S)	24,914 (4%)	9,512 (12%)	17,620 (5%)
– removed with properties filter	13,450 (54%)*	5,050 (53%)*	8,829 (50%)*
– removed with solubility filter	3,768 (15%)*	1,434 (4%)*	3,150 (17%)*
Fragment Library	7,696 (1%)	6,484 (8%)	5,641 (2%)
Non-Fragment Library	252,221 (42%)	28,582 (36%)	103,455 (32%)

Table 3.2: **Failed compounds in numbers** Number of compounds in the low MW list that fail the properties filters to be included in the Fragment Library. MW: Molecular weight, HA: Hydrogen bond acceptors, HD: Hydrogen bond donors, ALogP, RB: Number of Rotatable Bonds, PSA: Polar surface area.

	MW<100	**HA>6**	**HD>3**	**ALogP>3**	**RB>5**	**PSA>80**
Specs (8,829)	38 (0.4%)	424 (5%)	57 (0.6%)	5,016 (57%)	640 (7%)	3,977 (45%)
Maybridge (5,050)	0 (0%)	213 (4%)	84 (2%)	2,482 (49%)	263 (5%)	2,954 (58%)
Asinex (13,450)	11 (0.1%)	1,183 (9%)	108 (0.8%)	4,870 (36%)	1,148 (8%)	9,004 (67%)

Solubility component implemented in Pipeline Pilot estimates the aqueous solubility of the compounds. In the section above, a solubility threshold of LogS $\geq$ -3.3 was chosen and here the reason for this choice will be explained. The aim was to find a balance between the likelihood of solubility and the number of remaining compounds. This was achieved with an empirical approach.

First, the aqueous solubility calculator of Pipeline Pilot was compared to that one of the program MOE (Molecular Operating Environment, Chemical Computing Group; done by Dr. Ijen Chen of Vernalis). Vernalis had good experience with the solubility prediction of MOE. In a comparison of estimated solubility calculated with both programs, Pipeline Pilot estimates stricter i.e. a lower solubility, than MOE. As a result, the Pipeline Pilot solubility estimation can be trusted in the same manner as MOE's.

The number of compounds removed on the basis of the threshold of aqueous solubility was determined for compounds of Asinex and Specs during initial tests. The percentage of compounds passing the individual solubility thresholds were comparable (Tables 3.3 and 3.4).

Table 3.3: **Number of Asinex compounds removed when adjusting solubility filter**

All compounds Asinex:		**600,220**		
After strict Ro3:		**8,750**	**100%**	
Solubility	**LogS**	**Passed**	**Passed**	**Failed**
100 µM	-4.0	8,027	92%	723
200 µM	-3.7	7,318	84%	1,432
300 µM	-3.5	6,656	76%	2,094
400 µM	-3.4	6,299	72%	2,451
500 µM	-3.3	5,940	68%	2,810
1 mM	-3.0	5,940	68%	2,810
2 mM	-2.7	3,696	42%	5,054

Table 3.4: **Number of Specs compounds removed when adjusting solubility filter**

All compounds Specs:		**321,371**		
After strict Ro3:		**7,212**	**100%**	
Solubility	**LogS**	**Passed**	**Passed**	**Failed**
100 µM	-4.0	6,546	91%	666
200 µM	-3.7	5,860	81%	1,352
300 µM	-3.5	5,306	74%	1,906
400 µM	-3.4	5,013	70%	2,199
500 µM	-3.3	4,665	65%	2,547
1 mM	-3.0	3,693	51%	3,519
2 mM	-2.7	2,813	39%	4,399

In the earlier PhD project of Dr. Kerrin Bright (Bright, 2009), the aqueous solubility threshold was set to 100 µM. 180 fragments were purchased from Specs. Thereof, 8 compounds were impure and a further 68 insoluble at 200 mM in

DMSO or at 100 µM in aqueous buffer. As 37% of the above compounds were insoluble under the given conditions, this indicates the necessity of a more constrained threshold. Tables 3.3 and 3.4 highlight that a 100 µM filter let 90% of the fragments pass the threshold. Conversely, this means if 37% less had passed the filter, it is more likely that all were soluble. 90% (the number of passed compounds) reduced by 37% corresponds to 66.7%. With regards to the Tables 3.3 and 3.4, this corresponds to approximately 500 µM aqueous solubility at which LogS is equal to -3.3, hence this was taken as the new threshold.

3.3 Selecting a Fragment Set

There are many different possibilities for how one can approach the generation of a subset of fragments which represents chemical space to the largest extent. This section describes five different fragment library design procedures. All the presented protocols start with the initial separation of the chemical space into Fragment Library and Non-Fragment Library (Figure 3.1). Subsequent steps vary depending on the individual protocol. They include clustering to guarantee diversity and are compared to superstructures to guarantee high coverage of the full chemical space. The final selection of fragments is written to a file called the Final Fragment Set.

The following protocols are shown as implemented in the Pipeline Pilot Professional Edition (version 8.0.1.500). In this project, the program was installed to run on a server for faster calculations. There is always an additional version implemented in the freely available Student Edition (version 6.1.5.0).

3.3.1 Cluster All

In the *Cluster All* protocol (Figure 3.2), the initial separation into two libraries is performed as described above (Figure 3.1). All compounds of the Fragment Library are compared to all compounds of the Non-Fragment Library. Each fragment is therefore augmented by a property of average similarity to all of the Non-Fragments. The similarity is calculated with the asymmetric similarity coefficient by Tversky (page 59) (Figure 3.2 2). Step 3 clusters both libraries into a pre-defined number of clusters (Section 2.6, page 60). Each cluster is written to a separate file (Figure 3.2 3). Finally step 4 reads every cluster sequentially and finds the fragments in each of them. The fragment with the highest average similarity per cluster, as calculated in step 2, is selected and written to the file Final Fragment Set (Figure 3.2 4). The computation time of this protocol is just a couple of minutes when Pipeline Pilot Professional is run on the server.

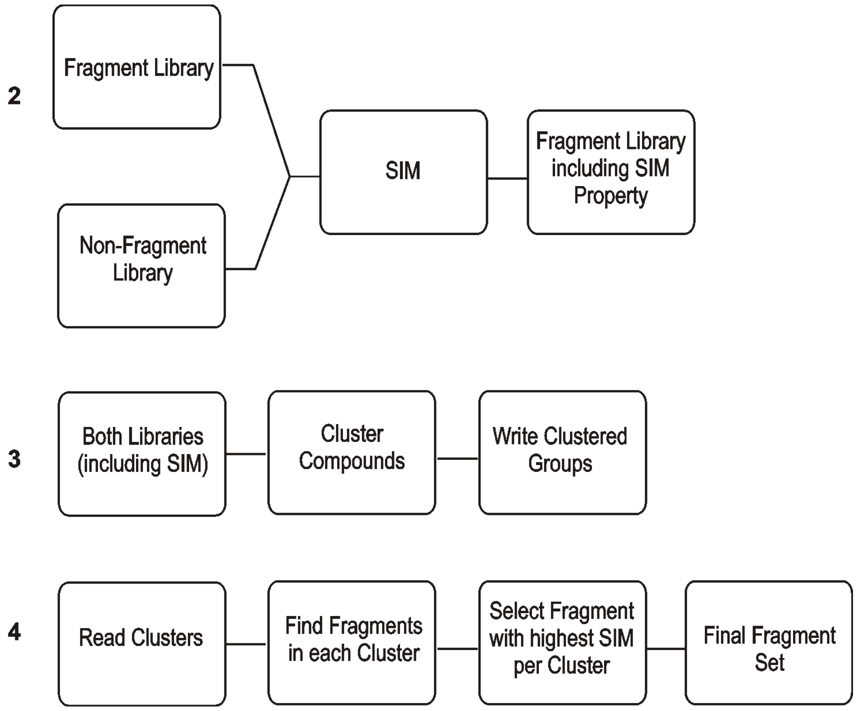

2
Fragment Library
Non-Fragment Library
SIM
Fragment Library including SIM Property
3
Both Libraries (including SIM)
Cluster Compounds
Write Clustered Groups
4
Read Clusters
Find Fragments in each Cluster
Select Fragment with highest SIM per Cluster
Final Fragment Set

Figure 3.2: **Cluster All** Selection of Final Fragment Set based of clustered libraries.

3.3.2 Cluster Fragments

The second protocol, *Cluster Fragments* (Figure 3.3), works in a similar way to the *Cluster All* protocol: All input compounds are split into two libraries (Figure 3.1). Secondly, each fragment is given a new property: average similarity to all the Non-Fragments (Figure 3.3 2). The difference is in the final step which clusters the Fragment Library only (not both Libraries as in *Cluster All*). Subsequently, that fragment with the highest average similarity in each cluster is selected and written to the Final Fragment Set file (Figure 3.3 3). The computation time of this protocol is just a couple of minutes on the server.

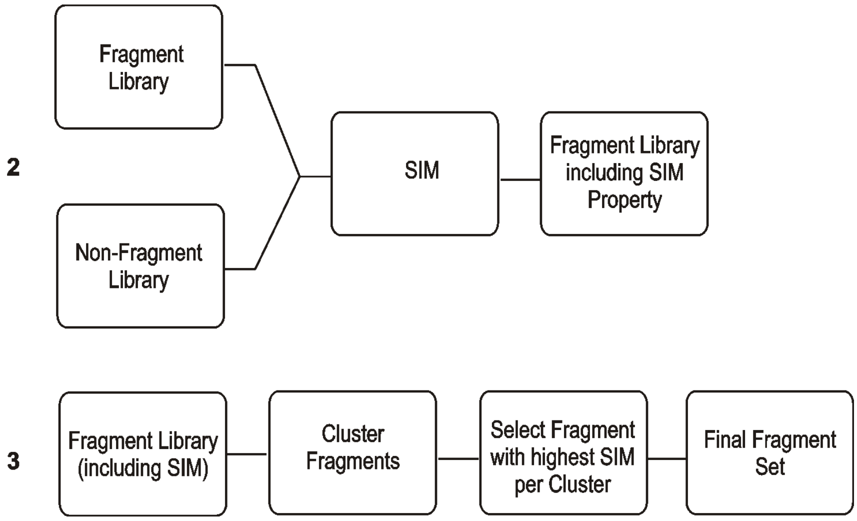

2
Fragment Library
Non-Fragment Library
SIM
Fragment Library including SIM Property
3
Fragment Library (including SIM)
Cluster Fragments
Select Fragment with highest SIM per Cluster
Final Fragment Set

Figure 3.3: **Cluster Fragments** Selection of Final Fragment Set based of clustered Fragments.

3.3.3 SIM within Cluster

This protocol, *SIM within Cluster*, (Figure 3.4) works like the previous ones using the *Cluster* component. The difference here is that the two Libraries are clustered first. Then the similarity between Fragments and Non-Fragments is calculated for each cluster separately (Figure 3.4 2). The fragment with the highest similarity to the Non-Fragments in its cluster is selected and written to the Final Fragment Set (Figure 3.4 3). The protocol runs within a couple of minutes with the Pipeline Pilot Professional version on the server.

3.3.4 Substructures

The following approach to fragment selection differs from the previous ones (Figure 3.5) and is not based on clustering. In the protocol *Substructures*, all fragments from the Fragment list are further "fragmented" into substructures. The presence of those substructures in the Non-Fragment list is rated. Those fragments whose substructures are most present in the Non-Fragments are selected (Figure 3.5 2).

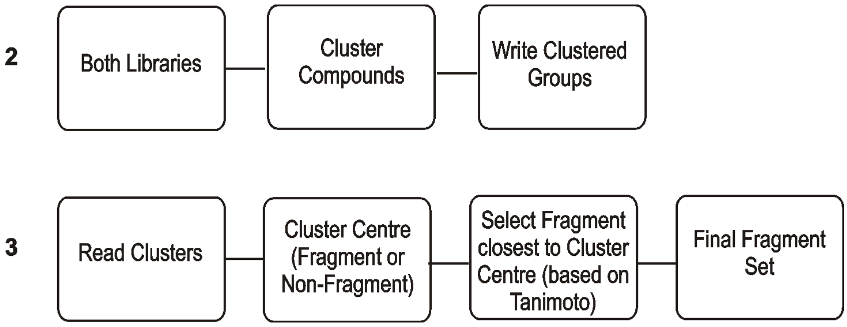

2
Both Libraries
Cluster Compounds
Write Clustered Groups
3
Read Clusters
Cluster Centre (Fragment or Non-Fragment)
Select Fragment closest to Cluster Centre (based on Tanimoto)
Final Fragment Set

Figure 3.4: **SIM within Cluster** Selection of Final Fragments Set based on the similarity within clustered Libraries.

Two rather similar modules used to count the presence of those substructures are available in Pipeline Pilot (*Substructure Count* and *Substructure Map*). Because they produce different results, they are treated separately. However, the logical procedure remains the same. In *Substructure Count* all occurrences of the substructures in all Non-Fragments are counted. Whereas if a query is mapped with the *Substructure Map* component, the count will increase by one independent of the number of mapped queries. Depending on the library size, these two protocols may take days to run on the server with the Pipeline Pilot Professional edition.

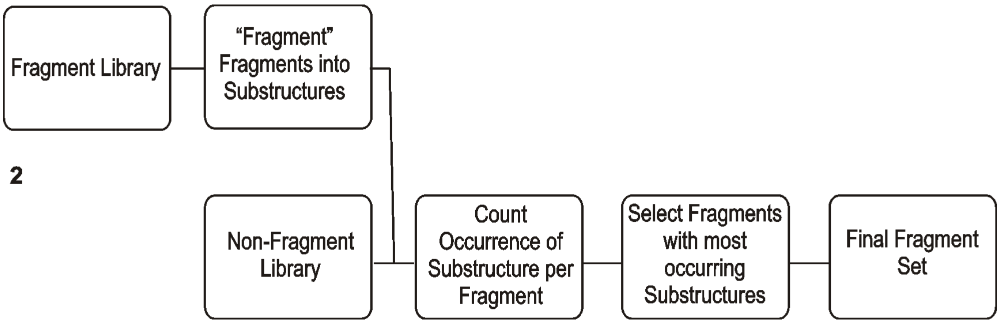

Fragment Library
"Fragment" Fragments into Substructures
2
Non-Fragment Library
Count Occurrence of Substructure per Fragment
Select Fragments with most occurring Substructures
Final Fragment Set

Figure 3.5: **Substructures** Selection of Final Fragment Set based on a substructure search.

3.3.5 Iterative Removal

The final and most complex method is called *Iterative Removal* and has been implemented with some help from Eddy Vande Water from Accelrys. Figure

3.6 (steps 2–4) shows the logic of this protocol. In step 2, the number of Non-Fragments which have a Tversky similarity > 60% to each fragment is counted. The 100 fragments with the highest number of such "nearest neighbours" (find more details below in Section 3.4.2) are then compared and the most diverse compound selected for the Fragment Set and removed from the Fragment Library. In step 3, the compounds with a Tversky similarity > 70% (Section 2.5.1, page 59) to this selected Fragment are removed from the Non-Fragment Library. The calculation is then repeated (back to step 2) with the reduced Fragment and Non-Fragment Libraries to identify the next most diverse Fragment with a high number of nearest neighbours and so on to eventually generate the Fragment Set of desired number of compounds (step 4). The similarity cut-offs were chosen to maximise the number of Non-Fragments for subsequent selections. Preliminary calculations showed that 50% Tversky similarity removes too many Non-Fragments and 80% too few (more in Section 3.4.2).

On the server with the Pipeline Pilot professional edition, the protocol will be finished a few hours after initiation. A slightly modified version without the *Diverse Molecules* component can run on the Pipeline Pilot student edition.

3.4 Additional Information to the Selection Protocols

3.4.1 Computation Time

The execution of the Pipeline Pilot protocols depends on the size of the input library, on which machine the calculation takes place and also on the number of fragments which must be selected for the Final Fragment Set. However as a rough guide, protocols based on clustering, *Iterative Removal* and the *Substructure* protocols take minutes, hours and days respectively for execution.

3.4.2 Nearest Neighbour Definition

The protocol *Iterative Removal* and some of the following analysis protocols require a definition of the term "nearest neighbour". "Neighbours" are com-

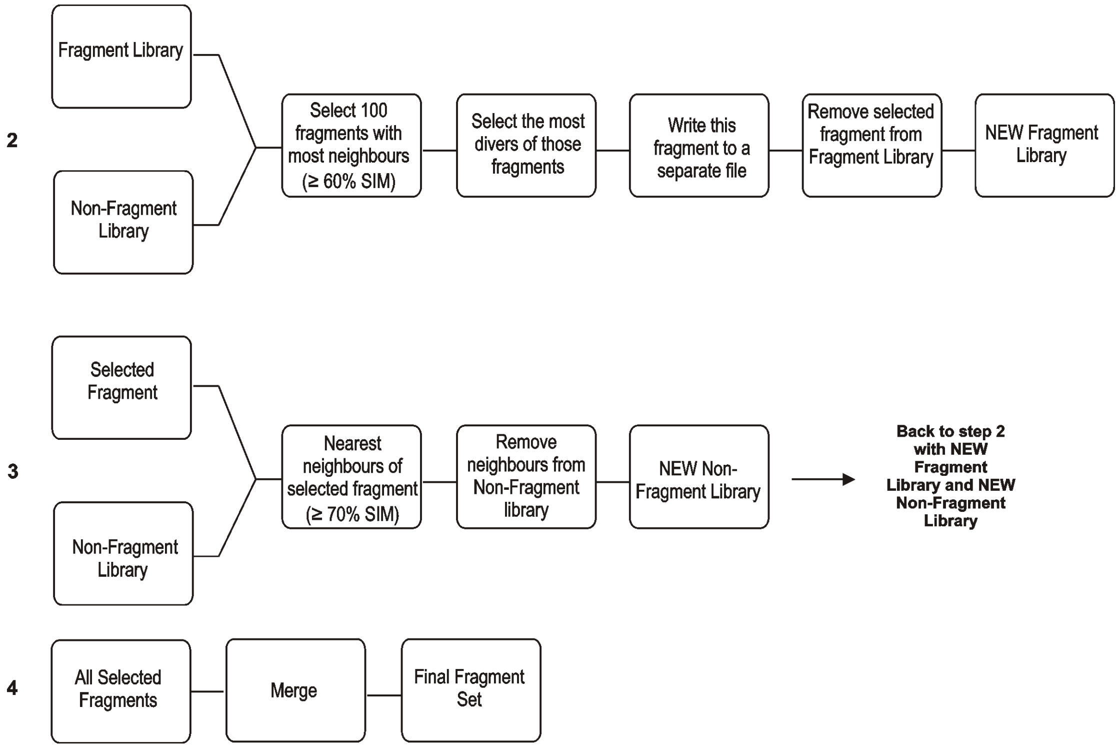
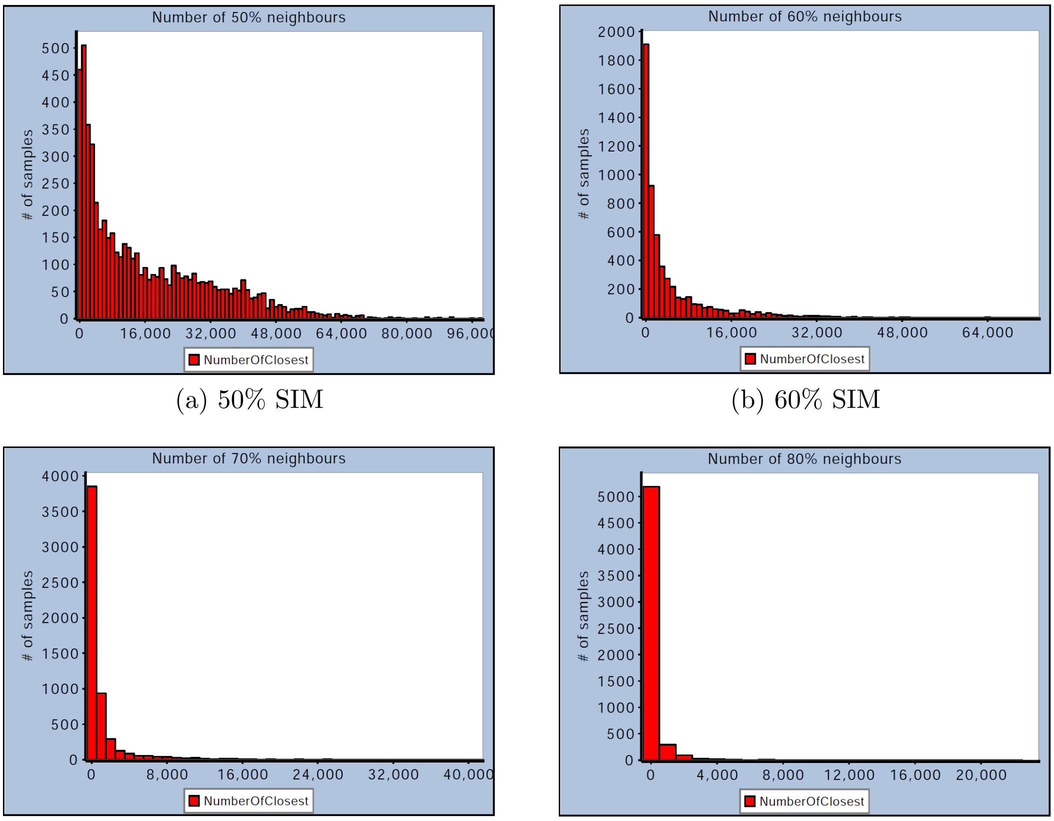

2
Fragment Library
Non-Fragment Library
Select 100 fragments with most neighbours (≥ 60% SIM)
Select the most divers of those fragments
Write this fragment to a separate file
Remove selected fragment from Fragment Library
NEW Fragment Library
3
Selected Fragment
Non-Fragment Library
Nearest neighbours of selected fragment (≥ 70% SIM)
Remove neighbours from Non-Fragment library
NEW Non-Fragment Library
Back to step 2 with NEW Fragment Library and NEW Non-Fragment Library
4
All Selected Fragments
Merge
Final Fragment Set

Figure 3.6: **Iterative Removal** Selection of Final Fragment Set based on iterative removal of top fragment and nearest neighbours.

pounds whose similarity to the fragment can be calculated with the Tanimoto or Tversky coefficient (Section 2.5, page 58). “Nearest neighbours” are compounds which have a similarity coefficient above a certain threshold between 0 and 1 (i.e. between 0 and 100%). Because here the compounds are fragments and the neighbours are larger compounds, the Tversky coefficient was used. Figure 3.7 illustrates, for the example of the Specs compounds, the nearest neighbour distribution. It is illustrated how many compounds have how many nearest neighbours for the thresholds 50–80%. The distribution is shown as histograms with a bin size of 1,000 compounds. One may note by the shape of the histograms that have a 50% similarity (Figure 3.7 a) as nearest neighbour definition includes far too many compounds whereas an 80% similarity (Figure 3.7 d) excludes too many compounds. Thus a nearest neighbour needs to be defined between 60 (Figure 3.7 b) and 70% Tversky similarity (Figure 3.7 c).

The *Iterative Removal* protocol was run using the 60% definition and the 70% definition, as well as with the combination of both. The run using 60% has



(a) 50% SIM

(b) 60% SIM

(c) 70% SIM

(d) 80% SIM

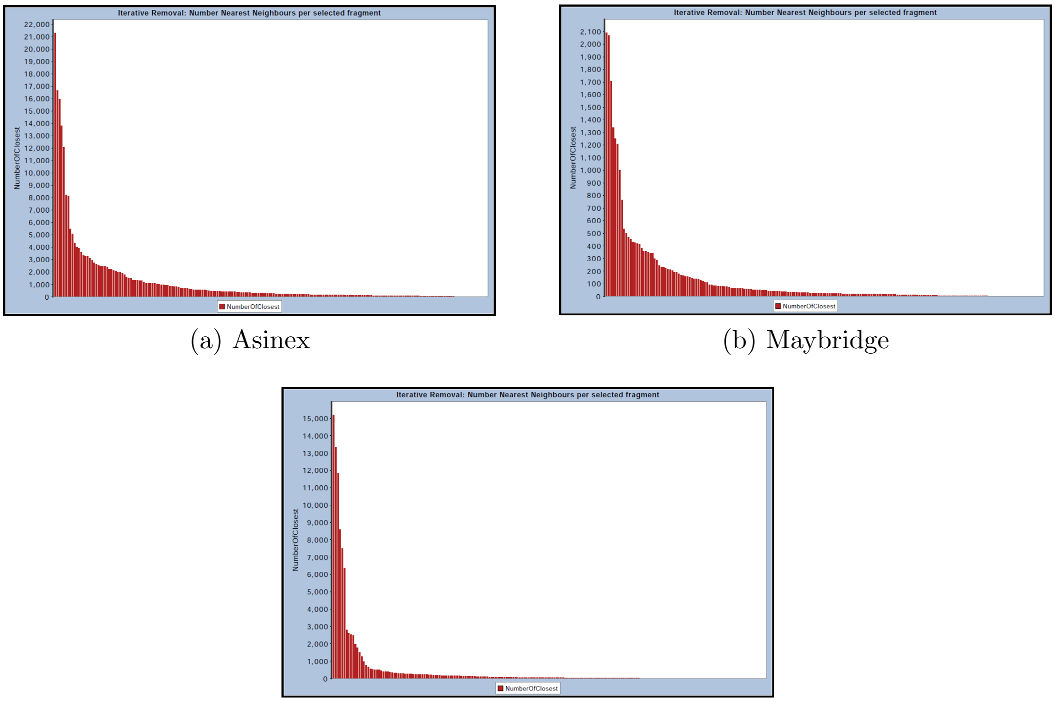
Figure 3.7: **Nearest neighbour distribution.** A screenshot from the Pipeline Pilot output shows the distribution of nearest neighbours depending on threshold set for definition of nearest neighbour. The similarity is calculated between the Fragments and the Non-Fragments of Specs.

been found to remove too many compounds, whereas using 70% removed too few. Nevertheless a combination of both values was found to perform best: The fragment with the highest number of nearest neighbours is selected based on 60% similarity. However, only those reference compounds being more than 70% similar to that selected fragment are removed from the Non-Fragments. Thus, the *Iterative Removal* is carried out based on two different definitions of “nearest neighbour”.

3.4.3 Determination of Fragment Set Size

For a suitable analysis of the protocols, the size of Final Fragment Sets needed to be decided. Table 3.1 on page 65 gives an overview on the size of the Fragment

Set during each step towards the final fragment selection. For example, the supplier Specs offers 321,371 input compounds. Those compounds were broken down into two Libraries consisting of 5,641 fragments and 103,455 non-fragments. For the analysis in this thesis 200 fragments were selected for the Final Sets. A bigger number was thought not to give a diverse and representative subset which can also be seen in Figure 3.8: The number of nearest neighbours decreases rapidly with the *Iterative Removal* procedure. A Fragment Set bigger than 200 would represent too high percentage of the Fragment Library, making 200 the optimal Fragment Set size. However, the *SIM within Cluster* and the *Cluster All* procedure generate a slightly smaller number. Due to the implementation of those procedures, there are possible clusters which do not contain any fragments.



(a) Asinex

(b) Maybridge

(c) Specs

Figure 3.8: **Decrease of nearest neighbours for the *Iterative Removal* Procedure.** A screenshot of the Pipeline Pilot output shows the available number of nearest neighbours decreases rapidly with this procedure. Pictured are the number of nearest neighbours removed by the procedure *Iterative Removal* for the selected fragments (in selection order) for all three test libraries – Asinex, Maybridge and Specs.

3.5 Profiling the Fragment Set

The aim of the selection protocols is to generate a Fragment Set which represents the maximum number of Nearest Neighbours in the Non-Fragment library.

The following analysis shown is applied to the Specs compounds as an example. Good experience with this supplier was made in previous buys (e.g. clean compounds, reasonable prices and quick delivery). However, complete analysis was done with compounds from Asinex and Maybridge to avoid any bias towards a particular supplier and delivered comparable results (Appendix B). Fragment Sets were profiled with their physicochemical and quality criteria such as diversity and drug-like properties. This section gives the results of the profiling process.

Table 3.5: **Physicochemical profile of the Fragment Sets of Specs - Part 1** Properties of the Fragment Sets generated by the different protocols for Specs and an overview of the original Fragment Libraries of all three supplier lists. An overview for the individual Fragment Sets of all suppliers can be found in the Appendix in Tables B.1 - B.3. MW: Molecular weight, AC: Number of heavy atoms (non hydrogen), FC: Formal charge, ALogP, HA: Number of hydrogen bond acceptors.

Library	MW	AC	FC	ALogP	HA
Asinex	216.5±33.35	15.3±2.30	-0.23±0.47	1.11±1.03	3.1 ±0.99
Maybridge	203.4±30.71	14.3±2.17	-0.18±0.42	1.36±0.89	2.9±1.03
Specs	205.4±37.69	14.5±2.63	-0.11±0.42	1.35±0.94	2.8±1.03
Cluster All	180.1±37.56	12.7±2.87	0.01±0.40	1.32±0.91	2.5±1.18
Cluster Fragments	179.5±36.87	12.7±2.80	0.01±0.38	1.33±0.86	2.5±1.12
SIM within Cluster	201.9±38.42	14.3±2.79	-0.04±0.43	1.48±0.92	2.8±1.18
Substructure Count	220.7±28.24	16.3±1.74	0.00±0.40	1.95±0.62	2.7±1.07
Substructure Map	237.7±26.34	17.3±1.46	-0.01±0.22	2.10±0.65	2.6±0.93
Iterative Removal	174.9±31.58	12.2±2.23	-0.11±0.32	1.38±0.96	2.1±0.85

3.5.1 Profiling for Physicochemical Properties

The fragment sets were profiled for physicochemical properties using the methods implemented in Pipeline Pilot. Tables 3.5–3.7 give an overview about the physicochemical properties of the generated Fragments sets. They are also graphically represented in Figure 3.9. A detailed overview can be found in the appendix in Tables B.1–B.3 .

Table 3.6: **Physicochemical profile of the Fragment Sets of Specs - Part 2** HD: Number of hydrogen bond donors, RB: Number of rotatable bonds, PSA: Polar surface area, LogS: Solubility, ArB: Number of aromatic bonds.

Library	HD	RB	PSA	LogS	ArB
Asinex	0.8±0.76	2.42±1.30	57.9±14.35	-2.30±0.76	7.09±3.57
Maybridge	0.7±0.75	1.94±1.27	56.6±14.78	-2.34±0.69	7.10±3.50
Specs	0.8±0.71	2.18±1.45	54.1±15.35	-2.33±0.74	6.90±3.31
Cluster All	0.7±0.69	1.44±1.30	49.8±17.84	-2.12±0.86	6.77±4.13
Cluster Fragments	0.7±0.71	1.59±1.35	48.9±17.60	-2.11±0.84	6.54±3.92
SIM within Cluster	0.7±0.67	1.90±1.43	52.3±16.13	-2.30±0.88	7.51±3.98
Substructure Count	0.8±0.78	2.21±1.11	49.3±16.67	-3.01±0.26	12.73±2.03
Substructure Map	0.7±0.81	1.77±0.93	47.2±17.05	-2.79±0.44	8.76±3.92
Iterative Removal	0.4±0.62	1.54±1.26	42.7±16.13	-2.17±0.77	6.02±2.98

Table 3.7: **Physicochemical profile of the Fragment Sets of Specs - Part 3** R: Number of rings, ArR: Number of aromatic rings, RA: Number of ring assemblies.

Library	R	ArR	RA
Asinex	1.9±0.63	1.3±0.69	1.5±0.52
Maybridge	1.8±0.73	1.3±0.67	1.4±0.55
Specs	1.8±0.71	1.2±0.63	1.4±0.52
Cluster All	1.7±0.78	1.2±0.77	1.3±0.51
Cluster Fragments	1.6±0.74	1.2±0.72	1.3±0.51
SIM within Cluster	1.9±0.73	1.4±0.75	1.4±0.55
Substructure Count	2.4±0.52	2.2±0.43	2.0±0.33
Substructure Map	2.9±0.68	1.5±0.71	2.1±0.44
Iterative Removal	1.5±0.70	1.0±0.54	1.2±0.45

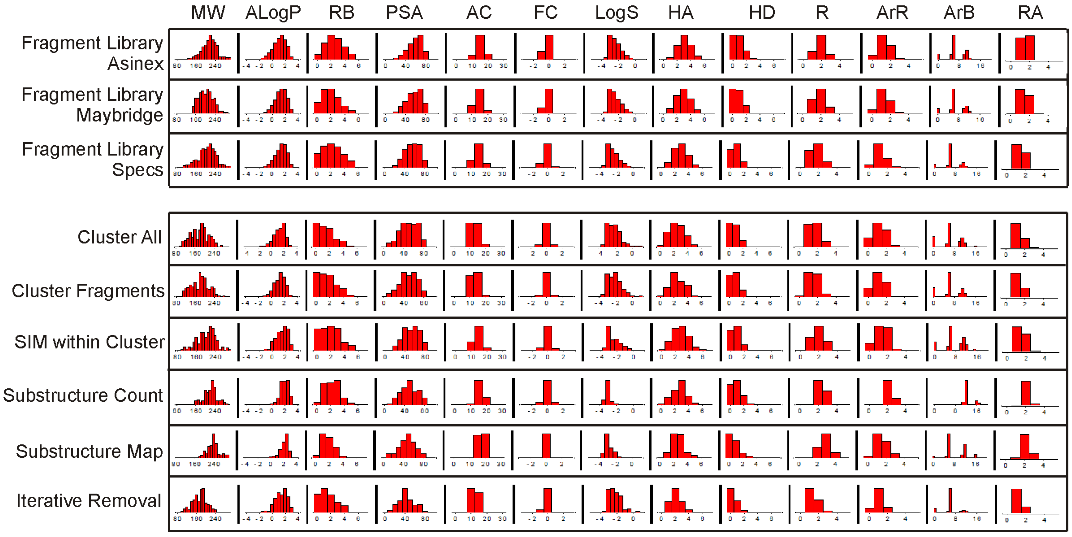


Figure 3.9: **Properties of Fragment Libraries and Specs Fragment Sets.** The first rows show the physicochemical property distribution of the three Fragment Libraries by supplier. The following rows show how these properties are distributed for the Fragment Sets generated with each of the different protocols on the example of Specs. MW: Molecular weight, ALogP, RB: Number of rotatable bonds, PSA: Polar surface area, AC: Number of heavy atoms (non hydrogen), FC: Formal charge, LogS: Solubility, HA: Number of hydrogen bond acceptors, HD: Number of hydrogen bond donors, R: Number of rings, ArR: Number of aromatic rings, ArB: Number of aromatic bonds, RA: Number of ring assemblies.

3.5.2 Quality Selection Criteria

In addition to the physicochemical properties the Fragment Sets were profiled as:

(a) The number of Nearest Neighbours with greater than 70% Tversky similarity (**SIMILARITY**)

(b) The average number of Non-Fragments that have less than 70% Tversky similarity for each fragment (**NON-SIMILAR**)

(c) The average similarity for each Fragment with drug-like molecules, calculated as (1) the average similarity with compounds from the Non-Fragments library that satisfy Lipinski's *Rule of Five* (**DRUG-LIKE 1**) and (2) the average similarity with compounds from the World Drug Bank (www.drugbank.ca) (**DRUG-LIKE 2**)

Table 3.8: **Quality criteria and analysis** The six quality criteria and their outcome for the different procedures. SIMILARITY is the number of nearest neighbours with > 70% Tversky similarity, NON-SIMILAR is the average number of Non-Fragments that have less than 70% Tversky similarity for each fragment, DRUG-LIKE 1 is the similarity to Non-Fragments filtered with the Lipinski filter, DRUG-LIKE 2 is the similarity with molecules from the World Drug Bank, DIVERSITY 1 is the average Tanimoto similarity within the Fragment Set and DIVERSITY 2 is the standard deviation of an equal cluster distribution within the Fragment Set. The asterix marks the best performing procedure where SIMILARITY is high, NON-SIMILARITY is low, DRUG-LIKE 1 is high and DIVERSITY 1 and 2 are low.

Protocol	**Supplier Library**	**SIMILARITY**	**NON-SIMILAR**	**DRUG-LIKE 1**	**DRUG-LIKE 2**	**DIVERSITY 1**	**DIVERSITY 2**
Cluster All	Asinex	217,156	246,820	0.421	0.307	0.201	6.34*
	Maybridge	23,446	28,190	0.373	0.295	0.172	7.62
	Specs	95,103	100,239	0.406	0.309	0.179	8.71
Cluster Fragments	Asinex	218,136	247,350	0.407	0.298	0.184	8.27
	Maybridge	23,866	28,183	0.367	0.291	0.162	9.44
	Specs	91,572	101,211	0.384	0.301	0.159	8.62
SIM within Cluster	Asinex	178,519	249,631	0.367	0.267	0.164*	8.40
	Maybridge	18,848	28,385	0.331	0.264	0.145*	8.50
	Specs	75,175	102,367	0.343	0.261	0.147*	10.28
Substructure Count	Asinex	108,589	249,439	0.390	0.267	0.256	10.03
	Maybridge	11,072	28,446	0.337	0.250	0.200	9.54
	Specs	45,757	102,508	0.367	0.265	0.235	11.42
Substructure Map	Asinex	121,667	250,317	0.367	0.276	0.201	11.42
	Maybridge	11,865	28,462	0.349	0.281	0.191	6.79*
	Specs	50,207	102,780	0.373	0.284	0.201	8.07*
Iterative Removal	Asinex	229,040*	245,541*	0.436*	0.319	0.198	10.07
	Maybridge	25,198*	27,976*	0.407*	0.312	0.198	8.60
	Specs	99,273*	98,749*	0.444*	0.335	0.193	8.18

(d) The chemical diversity of the Fragment Set, calculated as (1) **DIVERSITY 1**: the average Tanimoto similarity within the Fragment Set and (2) **DIVERSITY 2**: how equally distributed the Fragment Set is when clustered on fingerprint, calculated as the standard deviation of the number of compounds in each cluster. In the results reported here, 20 clusters were used as there were 200 fragments.

The protocols for this profiling analysis were implemented with Pipeline Pilot. The execution time of the analysis protocols is a few minutes per library with the Pipeline Professional edition on the server. Table 3.8 shows the result of the analysis for all three suppliers.

3.5.3 Library Overlap

The five implemented protocols generate different Final Fragment Sets. Table 3.9 illustrates the overlap between the final libraries generated with the input from Specs.

Table 3.9: **Library overlap on Specs example** Number of overlapping compounds between the different Fragment Sets.

Procedure	**Cluster All**	**Cluster Fragments**	**SIM within Cluster**	**Substructure Count**	**Substructure Map**	**Iterative Removal**
Cluster All	x	80	13	11	8	36
Cluster Fragments	80	x	15	12	8	35
SIM within Cluster	13	15	x	17	10	8
Substructure Count	11	12	17	x	69	10
Substructure Map	8	8	10	69	x	6
Iterative Removal	36	35	8	10	6	x

3.6 Discussion

Table 3.1 summarises the number of compounds from each supplier (Specs, Maybridge and Asinex) that remained after each stage of processing. In addition, Table A.1 also shows the number of compounds removed in generating the Input Library from each supplier for each of the unwanted chemical functionalities.

A large number of compounds are removed by the unwanted functionality filters. The filters for nitro, methylene, tertiary or quaternary amines, acrylates and atoms that are not C, N, F, Cl or S remove the most compounds. There is a similar profile of exclusions across the three suppliers, except the Maybridge Input Library has proportionately more acrylate and tertiary amine containing compounds. In addition, the Maybridge Input Library contains a proportionately larger number of compounds (12%) with low MW, compared to Specs (5%) and Asinex (4%), but about the same percentage (50%) of these low MW compounds from each supplier do not have the desired fragment properties. Table 3.2 lists the number of low MW compounds that do not pass the various property filters to be selected for the Fragment Library from the low MW list. The main failures are based on polarity (ALogP and PSA), with similar percentages of compounds failing for each of the suppliers.

The properties of the resulting Fragment Libraries are shown in Tables 3.5–3.7, and shown as bar charts in Figure 3.9. The Fragment Library from Specs has the expected distribution of properties given the criteria used for identifying the Library, with the most striking feature being the hard cut-off for solubility. The property distribution is similar for the other Fragment Libraries from Maybridge and Asinex (as would be expected given the strict criteria applied for selection), although the Asinex derived fragments are overall slightly larger compounds with more rotatable bonds, rings, H-bond acceptors and slightly more negatively charged.

Fragment Sets were derived and profiled for each of the six protocols for each Fragment Library from the three suppliers. In this section, the characteristics of the process and the properties of the different Fragment Sets are discussed in detail for one of the suppliers (Specs), with additional comments on any differences seen for the other two suppliers.

The properties of the six different Fragment Sets derived from the Specs Fragment Library are listed in Tables 3.5–3.7 and illustrated in Figure 3.9. Four main comments can be made about the properties of the different Fragment Sets compared to the parent Fragment Library. First, *Cluster All*, *Cluster Fragments* and *Itera-*

tive Removal select smaller compounds, as reflected in the molecular weight and heavy atom count. *SIM within Cluster* selects compounds with representative properties, whereas the two Substructure protocols select larger compounds with higher ALogP values. Secondly, *Iterative Removal* selects positively charged fragments, whereas the other protocols select some negatively charged ones, giving a small overall shift in the charged nature of the Fragment Sets. Also, compounds in the *Iterative Removal* Fragment Set have fewer hydrogen bond donors and acceptors. Thirdly, the Similarity within *Cluster Fragments* Set has the closest property profile to the Fragment Library. Finally, there are surprising differences in the properties of the compounds selected by the two *Substructure* protocols. The Fragment Set produced by the *Substructure Count* method has more hydrogen bond donors and acceptors, more rotatable bonds and more aromatic bonds, whereas the Fragment Set produced by the *Substructure Map* method has more rings and ring assemblies, a higher ALogP and larger compounds. The Fragment Sets derived by the six Protocols for the other suppliers show the same pattern of differences in characteristics (summarised in Tables B.1–B.3).

Table 3.10: **Final scoring of all procedures** Each time when a procedure performed best in one criterion (according to an asterix in Table 3.8), the relevant protocol gets one score. DRUG-LIKE 2 was not taken into account because it depends on the employed strategy (finding similar drugs or avoiding intellectual property). With nine scores, the *Iterative Removal* performs outstandingly better than the other protocols.

Protocol	**Score**
Cluster All	1
Cluster Fragments	
SIM within Cluster	3
Substructure Count	
Substructure Map	2
Iterative Removal	9

Some of these differences in properties can be rationalised by the nature of the protocols used. In *SIM within Cluster*, the compounds are clustered first and then the most representative fragment is selected. Assuming that clustering on FCFP_4 properties effectively clusters on the physicochemical properties, then it is not surprising this Fragment Set has the most similar properties to the Fragment Library. On the other hand, the *Substructure* protocols will select for larger fragments as more substructures will be present in such fragments and thus a higher score obtained for the presence of substructures in the Non-Fragments. The differences between *Substructure Count* and *Substructure Map* are due to the way bonds and ring features are counted. For example, if a molecule has seven more bonds and one more ring, then *Count* will give an increased score of eight, whereas *Map* would give an increased score of just two. The differences seen in the *Iterative Removal* set is probably because the smaller fragments will have a higher Tversky similarity to more Non-Fragments and will thus be selected in the first part of that Protocol. Such smaller fragments will also have a lower number of hydrogen bond donors and acceptors as well. However, the small differences seen in the average formal charge across the Fragment Sets is difficult to explain. It may not be significant, given the small number of compounds with charges that are present in the dataset.

Table 3.8 provides a profile of the overall characteristics of the Fragment Sets, considering the properties that are important for a screening collection. For these properties, the best library would be one that has a high SIMILARITY (and low NON-SIMILARITY) to the Non-Fragments, contains a diverse collection of fragments (low DIVERSITY scores), and is DRUG-LIKE. On these criteria, the Iterative Removal is overall the best performing protocol, with the exception of DIVERSITY scores, where it is average (Tables 3.8 and 3.10). As high SIMILARITY score is the primary aim of the library design, this has been analysed in more detail. Figure 3.10 a plots the number of fragments (y-axis) from each of the Fragment Sets that have more than a given number (x-axis) of compounds in the Non-Fragment Library from Specs which are more than 50% similar by Tversky. This metric reinforces the SIMILARITY calculation and shows that the Iterative Removal protocol is the most effective at generating a library that has the greatest coverage of the Non-Fragments. Figure 3.10 b shows how this Fragment Set covers the Non-Fragments at range of Tversky similarity from 20% to 90%.

3.7 Concluding Remarks

Different protocols for selecting Fragment Sets that are representative of a compound library were developed and investigated. The Iterative Removal protocol generates the best Fragment Set, judged by the similarity to the Non-Fragments and the overall characteristics of the Set. These protocols are relatively straightforward to implement and could be used to select fragments for screening with an increased probability that nearest neighbour compounds will be available for subsequent fragment evolution.

Two rounds of iteration led to the purchase of two new fragment sets in York - called "Michele_1" and "Michele_2". A preliminary version of the Iterative Removal was used to create "Michele_1". 201 fragments were purchased of Specs, whereof 95% were soluble at 200 mM in DMSO-d$_6$. The better solubility compares well to older libraries generated in York (63%) justifying the more sensible filter of 500 µM aqueous solubility (Section 3.2.1). "Michele_2" was created with the final version of the protocol and built on "Michele_1". Fragments were purchased from Specs, Sigma Aldrich, Asinex and Maybridge, and solved at 200 mM in DMSO-d$_6$. Both libraries were stored on plates in the dark after quality control, and then used to create a new database for York compounds. A full description can be found in Appendix C on page 194. The fragment sets were used in various screening campaigns in our institution and my results will be described in Chapters 6 and 7.

The library design approach presented in this chapter could prove useful to others for the design of fragment screening libraries that represent commercially available libraries, the compounds available in a proprietary in-house collection or a virtual library of compounds that could be rapidly synthesised given available resources. The protocols and a SMARTS file with unwanted functionalities can be downloaded from http://www.ysbl.york.ac.uk/fragments/. The results of this chapter were published as "Design of a Fragment Library that maximally represent available chemical space" (Schulz et al., 2011).

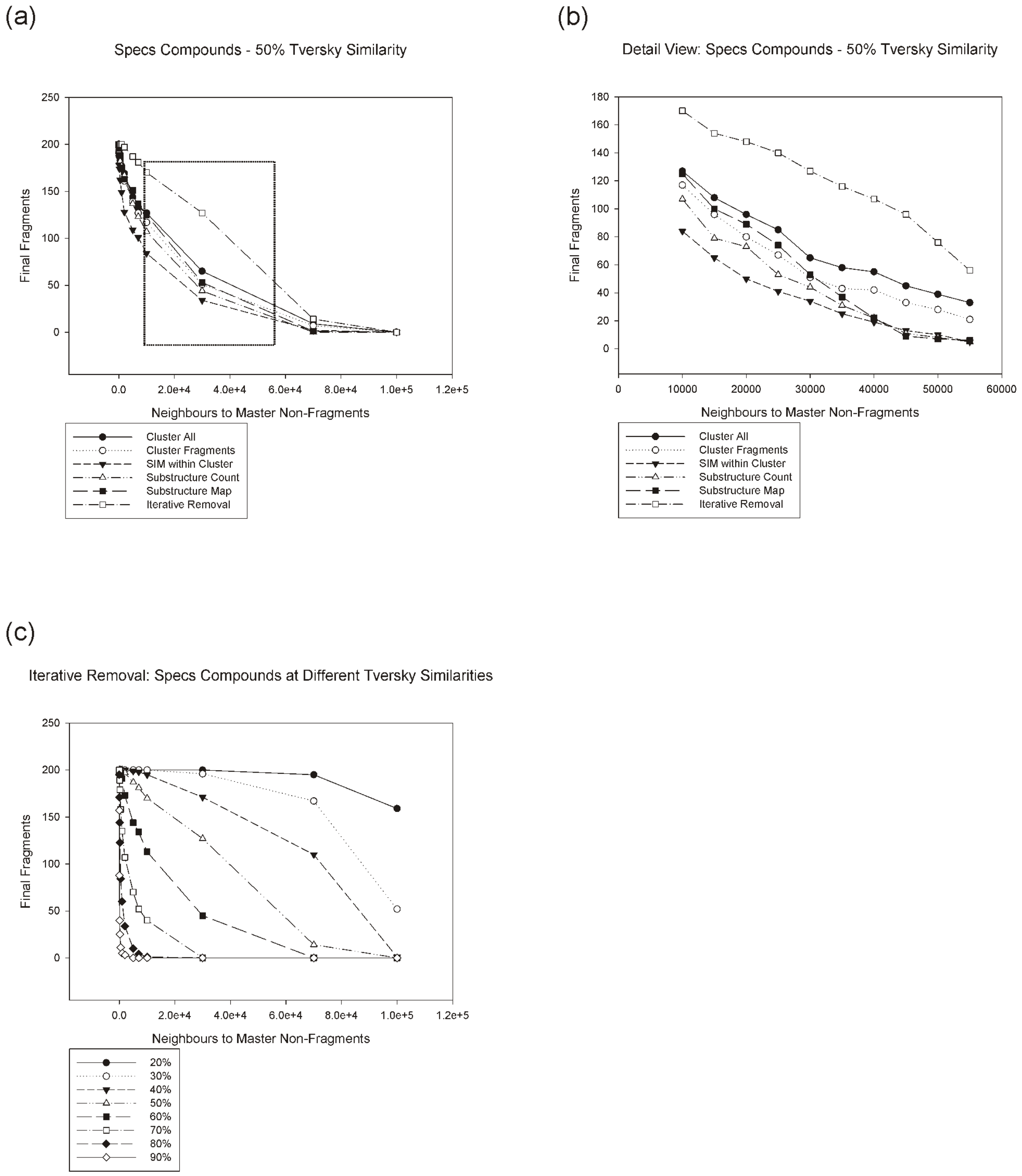


Figure 3.10: **Nearest neighbour plots for Specs derived Fragment Set.** Number of nearest neighbour compounds at 50% Tversky similarity for different Fragment Sets and number of nearest neighbours compounds for different Tversky similarity values for the Iterative Removal Fragment Set. (a) Specs compounds with 50% Tversky similarity. Number of compounds in the final fragment set per number of neighbours to the Non-Fragments library; section shown larger in (b). (c) Specs compounds at different Tversky similarities with the *Iterative Removal* procedure. Number of compounds in the final fragment set per number of neighbours to the Non-Fragments.

Chapter 4

MTSA

Thermal shift analysis is becoming widely used as a method for screening buffers and ligands to find stabilising conditions for proteins. For compound library screening, a large amount of raw fluorescence data has to be analysed. The data analysis software either provided by the equipment manufacturers or available in the public domain is cumbersome to use. The aim of this chapter is to devise a simple package that could be widely deployed for thermal shift assays. With this in mind the program MTSA with a graphical user interface (GUI) was developed within the statistical analysis package, Matlab (MathWorks) and its curve fitting toolbox, which is available in most institutions. For each experiment, the program outputs the melting temperature T_m, the deviation from a standard value ΔT_m (the thermal shift), and the quality of the fit R^2.

4.1 Introduction to Thermal Shift Analysis

Known for a long time (for example in Pace and McGrath, 1979), Thermal Shift Analysis (TSA) has recently become widely established as an efficient and effective method to screen a protein for ligand binding or conditions that improve stability. The method is based on heating a protein and recording the unfolding curve. This recording can be achieved by different methods such as differential scanning calorimetry or circular dichroism (Kranz and Schalk-Hihi, 2011), however fluorescence based experiments are usually carried out on plates in a qPCR (Quantitative Real Time Polymerase Chain Reaction) machine. To record the melting curve, an environmentally dependent fluorescence dye is added to the wells. When the protein starts unfolding ("melting"), the dye binds to the exposed hydrophobic groups giving a change in fluorescence. Although the idea that fluorescence increases when dyes bind to a protein's hydrophobic regions was first described nearly sixty years ago (Weber and Laurence, 1954), it has only recently gained popularity as a method for screening protein-ligand interactions (Kranz and Schalk-Hihi, 2011; Cummings et al., 2006; Sorrell et al., 2010), identifying ligands that can aid crystallisation for structural studies, or screening for buffers that can improve protein stability (Veddadi et al., 2006; Ericsson et al., 2006; Niesen et al., 2007). The technique is also known as Differential Scanning Fluorimetry (DSF), Temperature dependent Fluorescence (TdF) or "ThermoFluor". In some countries, ThermoFluor® is a registered trade name of a fully automated instrumentation developed by 3-Dimensional Pharmaceuticals Inc., later merged with Johnson and Johnson (Pantoliano et al., 2001; Matulis et al., 2004; Cummings et al., 2006). The experiment can be carried out in 96-well or even 384-well format, and the primary output is a plot of fluorescence against temperature for each well or sample. For well-behaved systems, this curve is sigmoidal in shape for temperatures around the melting temperature T_m.

4.2 Strategies of Data Analysis

A typical thermal shift experiment is pictured in Figure 4.1. The melting curves of a protein used with a concentration series of ligand are recorded. If the ligand binds to the protein, it is assumed to stabilise the protein, and the melting curves are shifted further to higher temperatures. In an ideal experiment, these curves are of sigmoid shape and are fully symmetric. The melting temperature T_m corresponds to the temperature where the protein is half folded and half denatured (Layton and Hellinga, 2010, Kranz and Schalk-Hihi, 2011). To analyse the experiment mathematically a curve model is needed. For thermal shift experiments derived by differential scanning fluorimetry (where the unfolding process is followed using the fluorescence of a dye), there exists three common ways of data analysis.

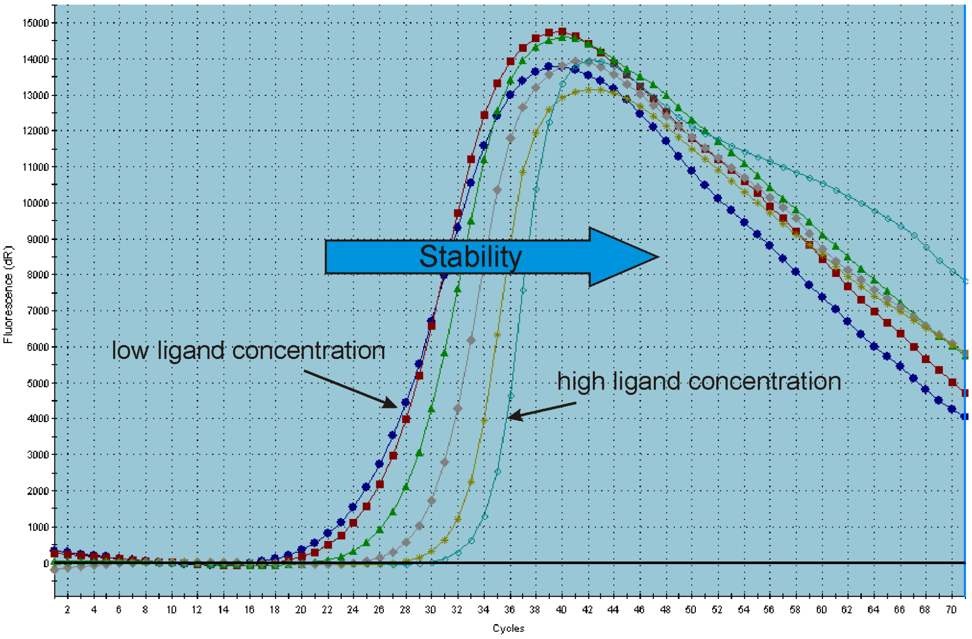


Figure 4.1: **Thermal shift experiment.** In an ideal experiment, the protein becomes stabilised by a ligand. This stabilisation leads to a shift of the thermal melting curve towards a higher temperature.

4.2.1 The Thermodynamic Model

The developers of the ThermoFluor method use a thermodynamic model to fit protein melting curves and to estimate the affinity of a protein-ligand interaction. A similar derivation was also described by Zubriene et. al. (2009). Because the ThermoFluor platform is widely used, this approach is discussed below. There is a lot of confusion about thermodynamic models in the literature due to typographic errors in various publications which have never been corrected. To address this issue, the following paragraphs go more into mathematical detail of the model.

Two slightly different ThermoFluor approaches are available in the literature. Both are based on an unfolding equilibrium model of protein in the native state N, unfolded protein U, free ligand L_f and bound ligand L_b (Matulis et al., 2005):

$$\mathrm{U} + \mathrm{L_f} \underset{unfolding}{\overset{K_U}{\rightleftharpoons}} N + L_f \underset{binding}{\overset{K_b}{\rightleftharpoons}} NL_b \tag{4.1}$$

The free energy is the total stability of the ligand-protein complex composed of protein stability energy $\Delta G_U(T)$ plus binding energy $\Delta G_b(T)$:

$$\Delta G(T) = \Delta G_U(T) + \Delta G_b(T) \tag{4.2}$$

The equilibrium constant for protein unfolding K_U is given by

$$K_U = \frac{[U]}{[N]} = e^{-(\Delta G_U(T)/RT)} \tag{4.3}$$

and the ligand binding constant is given by

$$K_b = \frac{[NL_b]}{[N][L_f]} = e^{-(\Delta G_b(T)/RT)} \tag{4.4}$$

First Approach

The inventors of the ThermoFluor first described an approach to fit the melting curves of a TSA assay (Pantoliano et al., 2001) specifies the fluorescence $\gamma(T)$ as:

$$\gamma(T) = \gamma_u + \frac{\gamma_f - \gamma_u}{1 + e^{-\Delta H_u/R(1/T - 1/T_m) + \Delta C_{pu}/R(ln(T/T_m) + Tm/T - 1)}} \tag{4.5}$$

The five parameters are melting temperature at the midpoint of the transition T_m, the unfolding enthalpy ΔH_u and heat capacity ΔC_{pu}, and the pre- and post-transitional fluorescence levels γ_f and γ_u, estimated with the Levenberg-Marquardt algorithm for minimizing the sum of the squares of the residuals.

In addition, the ligand binding affinity K_{L,T_m} at T_m can be described with the equation:

$$K_{L,T_m} = \frac{e^{-\Delta H_{u,T_0}/R(1/T_m - 1/T_0) + \Delta C_{pu,T_0}/R(ln(T_m/T_0 + T_0/T_m - 1)}}{[L_{T_m}]} \tag{4.6}$$

where K_{L,T_m} is the ligand dissociation constant at T_m (midpoint of protein unfolding in the presence of ligand), T_0 is the midpoint of protein unfolding in the absence of a ligand, $\Delta H_{u,T_0}$ the unfolding enthalpy in the absence of ligand, $\Delta C_{pu,T_0}$, the unfolding heat capacity in the absence of a ligand, $[L_{T_m}]$, the free ligand concentration at T_m ($[L_{T_m}] \simeq [L]_{local}$ if $[L]_{local} \gg [Protein]_{total}$) and R the universal gas constant (equation first described by Brandts and Lin, 1990).

Second Approach

The second ThermoFluor approach and its derivation were described by Matulis et al. (2005) and Kranz and Schalk-Hihi (2011). In contrast to the first approach above, the fluorescence intensity $\gamma(T)$ does not depend on the unfolding heat capacity ΔC_{pu}, but on two additional slope parameters m.

Matilus et al. (2005) and Kranz and Schalk-Hihi (2011) express the fluorescence of the dye $\gamma(T)$ in the assay with the equation

$$\gamma(T) = \gamma_F(T) + \frac{\gamma_U(T) - \gamma_F(T)}{1 + e^{\Delta G_U(T_m)/RT}} = \gamma_U(T) + \frac{\gamma_F(T) - \gamma_U(T)}{1 + e^{-\Delta G_U(T_m)/RT}} \tag{4.7}$$

Unlike in Pantoliano et al. (2001), pre- and post-translational fluorescences are described as linear functions depending on the temperature T. The baselines of folded (F) and unfolded protein (U) are expressed as

$$\gamma_F(T) = \gamma_{F,T_m} + m_F(T - T_m) \tag{4.8}$$

$$\gamma_U(T) = \gamma_{U,T_m} + m_U(T - T_m) \tag{4.9}$$

Two additional slope parameters m represent the slope of the temperature dependent fluorescence. The term $\Delta G_U(T)$ in Equation 4.7 may be replaced with the Gibbs-Helmholtz relationship (Equation 4.10). The parameters Gibbs free energy of protein unfolding $\Delta G_U(T)$, unfolding enthalpy $\Delta H_U(T)$ and unfolding entropy $T\Delta S_U(T)$ depend on the temperature.

$$\Delta G_U(T) = \Delta H_U(T) - T\Delta S_U(T) \tag{4.10}$$

The formula can be rewritten as a dependency from a reference temperature T_r, corresponding to the T_m of the protein in the absence of a ligand, and the heat capacity $\Delta C_{p,U}$:

$$\Delta G_U(T) = \Delta H_{U,T_r} + \Delta C_{p,U}(T - T_r) - T(\Delta S_{U,T_r} + \Delta C_{p,U} ln(T/T_r)) \tag{4.11}$$

According to Matulis et al. (2005), the Equations 4.7–4.11 can be summarised by the final model:

$$\gamma(T) = \gamma_{F,T_m} + m_F(T - T_m) + \frac{\gamma_{U,T_m} - \gamma_{F,T_m} + (m_U - m_F)(T - T_m)}{1 + e^{(\Delta H_{U,T_r} + \Delta C_{p,U}(T-T_r) - T(\Delta S_{U,T_r} + \Delta C_{p,U} ln(T/T_r)))/RT}} \tag{4.12}$$

Equation 4.12 is fitted with a non-linear least squares algorithm to estimate the six parameters γ_{F,T_m}, γ_{U,T_m}, m_F, m_U, $\Delta H_{U,T_r}$ and T_m. The heat capacity $\Delta C_{p,U}$ is considered to be temperature independent and kept constant. The value for $\Delta C_{p,U}$ is either measured with other methods or estimated based on protein composition.

At this stage, there are some discrepancies in the literature between Matulis et al. (2005) and Kranz and Schalk-Hihi (2011). After inspection, the derivation of Equation 4.12 by Matulis et al. (2005) is correct, and the derivation by Kranz and Schalk-Hihi (2011) contains some typographical errors. Their final equation misses the term m_F in the numerator of the fraction in Equation 4.12 and feature a wrong bracket for RT in the exponential term of the denominator which results in wrong units. These misleading typographical mistakes are really unfortunate because the manuscript by Kranz and Schalk-Hihi (2011) is published in a special issue of *Methods in Enzymology* about FBLD, which is likely to become a guiding review article for scientists working on fragments (Kuo, 2011).

Furthermore, this second approach resulting in Equation 4.12 differs from the derivation by Pantoliano et al. (2001) (Equation 4.5). More parameters (six instead of five) are used for fitting. The heat capacity is not included in the fit and kept fixed, whereas two new slope parameters m_U and m_F are introduced.

After deriving the fitting equation, Matilus et al. (2005) and Kranz and Schalk-Hihi (2011) explain how the affinity of the ligand can be estimated using the relationship in Equation 4.1. The total ligand concentration L_t is defined by

$$L_t = (1 - K_U)(\frac{P_t}{2} + \frac{1}{K_U K_b}) \tag{4.13}$$

where P_t the total protein concentration and K_U and K_b and the two equilibrium constants. The derivation of L_t is described in Cimmperman et al. (2008).

Summarised, the ligand concentration needed to raise the protein T$_\mathrm{m}$ to a given value can be expressed as:

$$\begin{aligned}L_t =&(1 - e^{-(\Delta H_{U,T_r} + \Delta C_{p,U}(T_m - T_r) - T_m(\Delta S_{U,T-r} + \Delta C_{p,U} ln(T_m/T_r))/RT_m)})\\ &\times (\frac{P_t}{2} + 1/e^{-(\Delta H_{U,T_r} + \Delta Cp,U(T_m - T_r) - T_m(\Delta S_{U,T_r} + \Delta C_{p,U} ln(T_m/T_r)))/RT_m})\\ &\times (e^{(\Delta H_b(T_0) + \Delta G_{p,b}(T - T_0) - T(\Delta S_b(T_0) + \Delta C_{p,b} ln(T/T_0))}/RT)\end{aligned} \tag{4.14}$$

After consultation of one of the authors, Dr. James Kranz (personal communication, 2012), he states that there is an additional typographical error with a ± sign in Pantoliano et al. (2001) and Matulis at al. (2005) leading to wrong ligand affinity K_{L,T_m} and ligand concentration terms L_t. This error was also discovered by Zhang and Monsma (2010) who explained the mistake originating in a confusion of models: instead of the ligand binding to the native protein, the model describes ligand binding to the unfolded protein. The model was corrected in Zubriene et al. (2009), leading to the equations:

$$L_t = (K_{U,T_m} - 1)(\frac{P_t}{2K_{U,T_m}} + \frac{1}{K_{b,T_m}}) \tag{4.15}$$

and

$$\begin{aligned}L_t =&(e^{-(\Delta H_{U,T_r} + \Delta C_{p,U}(T_m - T_r) - T_m(\Delta S_{U,T-r} + \Delta C_{p,U} ln(T_m/T_r))/RT_m)} - 1)\\ &\times (\frac{P_t}{2} + 1/e^{-(\Delta H_{U,T_r} + \Delta Cp,U(T_m - T_r) - T_m(\Delta S_{U,T_r} + \Delta C_{p,U} ln(T_m/T_r)))/RT_m}\\ &+ 1/e^{-(\Delta H_{b,T_0} + \Delta C_{p,b}(T_m - T_0) - T_m(\Delta S_{b,T_0} + \Delta C_{p,b} ln(T_m/T_0))}/RT_m)\end{aligned} \tag{4.16}$$

To summarise, there are many calculation and typographical errors in the literature on TSA. A model for the ligand concentration is not used in this thesis.

4.2.2 The Boltzmann Model

Other groups, mainly those working on qPCR machines, use the Boltzmann equation (Equation 4.17) to fit the transition part of their data (Ericsson et al., 2006;

Niesen et al., 2007; Sorrell et al. 2010). The equation describes the Boltzmann distribution for the states of a system. Despite its name, the parameters here do not strictly conform to the thermodynamic equation. General variables are used instead of thermodynamic parameters. The parameter T' corresponds to the point of inflection and midpoint between the asymptotes. The equation is a four-parameter logistic model:

$$\gamma(T)_{Boltzmann} = min + \frac{max - min}{1 + e^{\frac{T' - T}{a}}} \tag{4.17}$$

where $\gamma(T)$ is the fluorescence, min and max are the temperatures at the fluorescence intensity before the transition and at the end of the transition respectively. In respect to the above named models, these variables correspond to γ_f and γ_u in Equation 4.5. The melting temperature T_m is at T':

$$T_{m,Boltzmann} = T' \tag{4.18}$$

and corresponds equally to the midpoint and to the point of inflection of the transition curve.

4.2.3 Higher Order Polynomial Equations

There are also reports where researchers fit the fluorescence data with higher order polynomial equations (Yeh et al., 2006; Niesen et al., 2007; Crowther et al., 2009; Wang et al. 2011). In this case, the melting temperature T_m is defined as the inflection point of the data. These models are not strictly based upon thermodynamic equations.

4.3 The Program MTSA

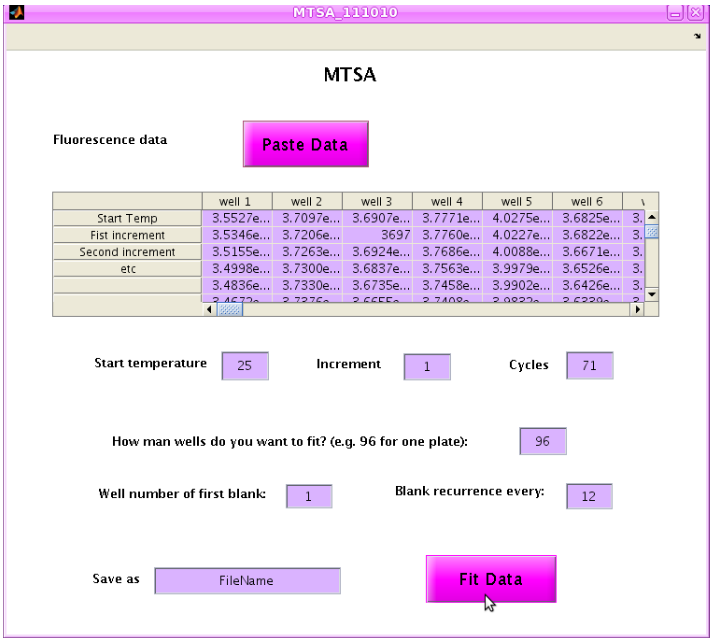
While the ThermoFluor system provides analysis software with the device, the current software for analysing melting curves produced on qPCR machines requires rather tedious data preparation and manipulation. The fluorescence curves need to be exported from the instrument, cut (which is often performed with an

Excel tool from the Niesen group (Niesen et al., 2007)) and further exported to separate fitting software to obtain the T_m. Some groups have developed software in-house (Vedadi et al., 2006) for this analysis; others use a series of protocols and scripts. In this chapter, a simple program was devised that could be widely deployed. MTSA with a graphical user interface (GUI) was developed within the statistical analysis package, Matlab (MathWorks) and its curve fitting toolbox, which is available in most institutions.

The raw fluorescence data from qPCR machines is usually available in a table where the rows signify the fluorescence at a particular temperature and the columns the different wells or experiments when the data is exported as horizontally grouped. It is good practise to include blanks (usually the protein in a standard buffer with no ligand) at regular intervals to account for instrument variability, such as plate edge effects, or for sample and pipetting errors.

The analysis begins with a "cut and paste" procedure of the table of fluorescence data into the MTSA interface (Figure 4.2). The user enters the starting temperature, temperature increment, and number of increments to define the columns in the data table. In addition, the number of wells to be fitted and the position of blank samples (e.g. ligand-free protein) are required.

The work in this thesis applied the thermal shift method to initial ligand screening, and thus does not require determination of thermodynamic parameters to estimate the binding constant (Equations 4.6 and 4.14). Thus the more complicated thermodynamic model (Equations 4.5 and 4.12), derived by the inventors of ThermoFluor requiring knowledge of heat capacity, was discarded. First, the heat capacity either has to be estimated based on protein sequence, or derived by calorimetric characterisation, which is not convenient for straightforward assay development. Second, the heat capacity is not temperature independent as the authors state (Gomez et al., 1995). The Boltzmann equation has the same shape as the thermodynamic model. Therefore, following the instructions in a Nature protocol (Niesen et al., 2007), the Boltzmann equation was used for first fitting experiments. However, its initial usage did not give satisfactory fits around the asymptotes (Figure 4.3 a) so an additional parameter c was introduced. The five-parameter logistic model is referred to as the Sigmoid-5 equation:



MTSA_111010

MTSA

Fluorescence data

	well 1	well 2	well 3	well 4	well 5	well 6	\
Start Temp	3.5527e...	3.7097e...	3.6907e...	3.7771e...	4.0275e...	3.6825e...	3.
Fist increment	3.5346e...	3.7206e...	3697	3.7760e...	4.0227e...	3.6822e...	3.
Second increment	3.5155e...	3.7263e...	3.6924e...	3.7686e...	4.0088e...	3.6671e...	3.
etc	3.4998e...	3.7300e...	3.6837e...	3.7563e...	3.9979e...	3.6526e...	3.
	3.4836e...	3.7330e...	3.6735e...	3.7458e...	3.9902e...	3.6426e...	3.
	3.4672e	3.7376e	3.6655e	3.7408e	3.9832e	3.6339e	3.

Start temperature 25

Increment 1

Cycles 71

How man wells do you want to fit? (e.g. 96 for one plate): 96

Well number of first blank: 1

Blank recurrence every: 12

Save as FileName

Figure 4.2: **GUI of MTSA.** To use MTSA, the horizontally grouped data need to be present in one block (the Stratagene machine for example exports in two blocks, so data need to be pasted together first). Use the computer's clipboard to copy the block of fluorescence data. In MTSA, using the button "Paste Data" to paste from the clipboard. Enter the start temperature, increment and number of cycles, as well as how many wells you want to have fitted. If you have blanks on the plate, enter the well number of the first blank and the interval between recurring blanks.

$$\gamma(T)_{Sigmoid-5} = min + \frac{max - min}{(1 + e^{\frac{T'-T}{a}})^c} \tag{4.19}$$

where a is the Hill slope (the steepness) and c is asymmetric factor introduced here to account for the asymmetric shape of the curves in order to improve fitting. T' is a parameter approximately in the middle of the transition area. For $c = 1$ (i.e. when the curve is fully symmetric), Sigmoid-5 becomes the Boltzmann equation (Equation 4.17). Then $T_m = T'$ and equal to the point of inflection and the midpoint.

Because of the asymmetric factor c, this model is not strictly based upon thermodynamic equations. However, the non-thermodynamic models have been used before to fit the melting curves (Sections 4.2.2 and 4.2.3).

When $c \neq 1$, the T_m can be defined as either the midpoint (for further discussion refer to Chapter 8 on page 151):

$$T_{m,Midpoint} = T' - a \times ln(2^{\frac{1}{c}} - 1) \tag{4.20}$$

or as the point of inflection:

$$T_{m,Inflection} = T' - a \times ln(\frac{1}{c}) \tag{4.21}$$

The program first identifies the temperatures T_{max} and T_{min}, where T_{max} is the temperature where the fluorescence is at its maximum, while T_{min} represents temperature of the minimum of the fluorescence data found on the left hand side (lower temperature) of T_{max}. The points for temperatures less than T_{min} and greater than T_{max} are then discarded. The parameter T' is approximately the melting temperature and set to 60 ± 50 ℃. The remaining data points are fitted with the Sigmoid-5 model.

It was found to be important to include boundary conditions for the fitting parameters: $0 < a < 10$ and $0 < c < 10$. The maximum and minimum fluorescence must be within ±2% of the *max* and *min* values respectively. The fit is made with a non-linear least squares method with a starting point of 5 for a, 1 for c, maximum value at *max* and minimum value at *min*. These boundary conditions were empirically determined to help the fit to be found. If *min* and *max* are forced to be within ±2% of the experimental values, the fits become better around the asymptotes

The default algorithm "Trust Region" was used which allows a less complex function to reasonably reflect the function in the neighbourhood around a certain point. Without these boundary conditions, the fitted data were exceeding the *max* value resulting in lower quality fits. The start values are necessary to find the actual fit.

Using the Sigmoid-5 equation, better fits were obtained compared to the more widely used four-parameter Boltzmann equation (missing the asymmetric factor c). This can be seen in Figure 4.3, which illustrates how the additional asymmetric factor c of the five-parameter equation helps to fit the curves a bit more smoothly, especially around the asymptotes.

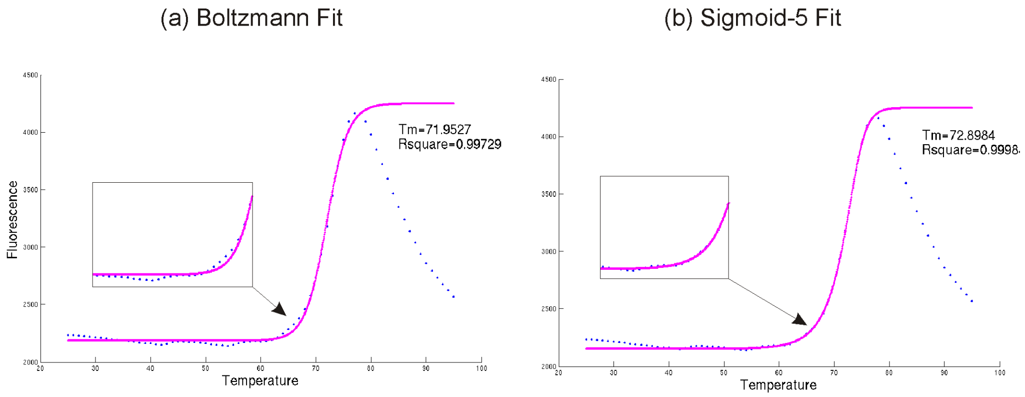


Figure 4.3: **Comparison of four- and five-parameter fitting equation.** The data points of the same experiment were fitted with the standard four-parameter Boltzmann equation (a) and with the five-parameter fit (b). The fit is better with the five-parameter equation which is visible around the asymptotes (the blown up section) and also in the coefficient of determination R^2. The T_m differs in almost 1 degree.

The program outputs a png file (an example is shown in Figure 4.3 (b)) for each column of data (each experiment) showing how well the sigmoid curve fits the original data. In addition, there is an output file (to be opened with a text editor) which provides a log and summary of the calculation for all the experiments. It contains a log of the fitting of each experiment, including the final equation of fit, number of iterations, coefficients (including the range of their 95% confidence interval) and the coefficient of determination, R^2. Furthermore, the average T_m ($\overline{T_m}$) value is determined from the blanks (blank T_m) and the standard deviation of $\overline{T_m}$. The output file concludes with a summary of the experiment number, T_m, ΔT_m and coefficient of determination R^2 (which expresses the correlation between data points and fit).

The program has been designed to operate smoothly with simple cut and paste input of data from different makes of qPCR machines (tested devices were AB 7300, AB 7500, Bio-Rad CFX96 and Agilent Stratagene MX3005), and to work across different versions of different operating systems (please note that Java is required; for Mac the correct version might have to be downloaded and the path set). This means that there were not implemented some desirable features, such as importing annotation of experimental detail (variation across manufacturers) or outputting Excel spreadsheets (variation with different operating systems).

However, the program does include some simple error checking to identify when experiments have failed or did not produce optimal fits. If the standard deviation of the blank T_m is greater than 0.2 or the coefficient of determination R^2 is less than 0.999, then the program issues a warning. In our experience, a problem with the blanks usually signifies a general problem with the experimental setup (reagents or pipetting errors) and the plate should be repeated. Other problems we have experienced include non-flat baselines, double-humped curves and wave-like curves. The program will always try to fit input data if it can identify a suitable *max* and *min*. However, odd-shaped curves will result in poor R^2 values and aberrations in the T_m values. These can be identified within the output file so the user can then examine the raw fluorescence data by eye. In the final analysis, the user needs to judge if the fit and the obtained T_m and ΔT_m is reliable or not. To estimate ligand binding, a significant ΔT_m is often considered to be above three times the standard deviation of the blank (Sorrell et al., 2010).

4.4 Concluding Remarks

In summary, a freely available tool was developed that significantly improves and accelerates the analysis of thermal shift data, which is especially helpful for a high throughput of plates. The program code and GUI (*.m and *.fig file) for the three different approaches (1. Boltzmann fit, 2. Inflection point of Sigmoid-5 fit, 3. Midpoint of Sigmoid-5 fit) can be downloaded from http://www.ysbl.york.ac.uk/fragments/MTSA.

The program requires the curve fitting toolbox and a compatible version of Java. Mac users might experience problems as on this operating system, Matlab uses the Java version provided by the operating system. Refer to http://www.mathworks.co.uk/help/matlab for further details.

A paper on the program MTSA was published in Analytical Biochemistry in 2012 (Schulz et al., 2012).

Chapter 5

N-Myristoyl Transferase

This chapter describes attempts to produce the protein N-myristoyl transferase (NMT) as a target for fragment screening. In summary, subcloning was performed to obtain double His-tagged protein for immobilisation on a surface Plasmon resonance (SPR) chip for fragment screening. The protein was assessed with mass spectrometry and tested in initial SPR experiments. However, it proved challenging to scale up expression and generate stable protein suitable for reliable fragment screening. For these reasons the focus moved onto other proteins.

5.1 Background

This chapter deals with the NMT proteins from *Leishmania donovani* (ld), *Leishmania major* (lm) and *Trypanosoma brucei* (tb). *Leishmania major* and *donovani* are parasites which are responsible for visceral and cutaneous Leishmania. *Trypanosoma brucei* on the other hand is a parasite responsible for sleeping sickness. N-myristoyl transferase catalyses the transfer of myristate from myristoyl-coenzyme A (MCoA) to the N-terminal glycine of a substrate protein. This process is important for the survival of the cells. While the binding site of MCoA is conserved within the different organisms, the peptide binding site varies, thus this site will be a good starting point to find selective inhibitors for drugs against Leishmania (Bowyer et al., 2008).

The project was originally planned as a collaboration between Pfizer, University of Dundee, University of York, NIMR and Imperial College. My part was the fragment screen in collaboration with Prof. Debbie Smith of the Centre for Immunology and Infection at York.

5.2 Producing Double His Tagged ldNMT

5.2.1 Motivation

Surface Plasmon Resonance (SPR) is a powerful technique for identifying and characterising the binding of small molecules and fragments to a protein (more in Section 10.5). In a direct binding experiment, the protein is linked to the SPR chip and the different small molecules flowed past. A number of different attachment strategies are possible. An attractive approach is to attach the protein through a His-tag to an NTA chip, exploiting the tag which is often present for purification. This should ensure a consistent attachment through a defined point, compared to methods such as biotinylation where any free NH2 group (such as lysines) can be attachment points. However, often the protein does not stick very well to the chip and bleeds off during the experiment. To prevent that, a second His-tag can be added to the protein in order to make it stick more to the NTA groups (Fischer et al., 2011). In the case of ldNMT, the double His_6-tag building was achieved via subcloning because a clone with a single His_6-tag was already available. In this thesis, "His-tag" refers to a hexa histidine tag, and "double His-tag" to two hexa histidine tags separated by a spacer respectively.

5.2.2 Subcloning

The template used was a single N-terminal His-tagged ldNMT construct in pSKB2 aka pET28-PPX (from Dr. Jim Brannigan) containing a protease cleavage site.

It was ligated into a pET28a vector using NheI and SacI enzyme restriction sites (according to the methods described in Section 10.2 on page 172). The primers were designed to amplify the template sequence via PCR.

5.2.3 Transformation

The ligation mixtures were transformed into XL10 Gold cells. The plasmids of the five different cultures were extracted, digested with restriction endonucleases and finally run on a 1% agarose gel (Figure 5.1). The concentration of the plasmids is shown in Table 5.1.

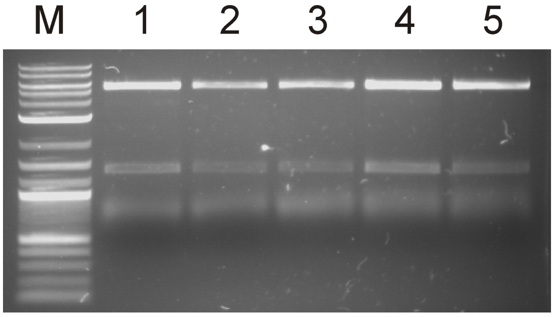


Figure 5.1: **Gel of Double Digest ldNMT in pET28a.** The extracted and digested plasmids of five selected colonies of the transformation procedure were run on a 1% agarose gel.

Table 5.1: **DNA concentration**

1	2	3	4	5
53.3 ng/µl	69.6 ng/µl	45.7 ng/µl	54.6 ng/µl	55.4 ng/µl

5.2.4 Sequencing

Samples 2 and 5 contained the highest DNA concentration and were selected for sequencing. Both plasmids showed a silent mutation on position 6 of the template as shown in Table 5.2. It is known that this mutation was already present in the initial template received from Dr. Jim Brannigan and was therefore not considered to be disadvantageous. Plasmid 2 showed some errors in the added second His-tag (data not shown). It was not obvious if that was due to an error in sequencing or a mutation. Since plasmid 5 did not show any mutations and contained the expected sequence, it was selected for further experiments and plasmid 2 was discarded.

Table 5.2: **Silent Mutation**

position	1	2	3	4	5	6
original gene	A	T	G	T	C	C
template	A	T	G	T	C	T
amino acid	Met			Ser		

5.2.5 The New ldNMT Construct

Figure 5.2 illustrates the new ldNMT construct schematically. Table 5.3 shows the sequence of the new ldNMT clone. The expressed protein has a mass of 53,128.5 Da and a theoretical pI of 6.68.

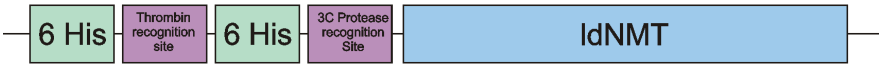


Figure 5.2: **Double His_6-tagged ldNMT construct.** The construct consists of two His_6-tags on the 5' end interrupted by two protease recognition sites.

5.2.6 Expression Tests

Plasmid stock 5 was used to transform into competent *Escherichia coli* Rosetta 2 (DE3) expression cells. Figure 5.3 illustrates that the protein is well expressed during the first expression tests where the cells were lysed in Milli-Q. However, the soluble fraction is not very big.

Several lysis buffers were tested (sugar, tris, tris glycerol, phosphate, details in Section 10.2), but none of them improved the fraction of soluble protein. However, these issues were known from experience with the single His-tagged protein (Dr. Jim Brannigan, personal communication). Therefore, scale-up experiments were performed straight away.

5.2.7 Scaling up Protein Production

To overcome the solubility issues, the protein was purified with a His trap crude 1 ml column (nickel-NTA) which contains larger pore sizes.

Table 5.3: **Sequence of double His_6-tagged ldNMT**

start of original sequence numbering:				1
10’	20’	30’	40’	50’
MGSSHHHHHH	SSGLVPRGSH	MASHHHHHHS	SGLEVLFQGP	HMSRNPSNSD
60’	70’	80’	90’	100’
AAHAFWSTQP	VPQTEDETEK	IVFAGPMDEP	KTVADIPEEP	YPIASTFEWW
110’	120’	130’	140’	150’
TPNMEAADDI	HAIYELLRDN	YVEDDDSMFR	FNYSEEFLQW	ALCPPSYIPD
160’	170’	180’	190’	200’
WHVAVRRKAD	KKLLAFIAGV	PVTLRMGTPK	YMKVKAQEKG	QEEEAAKYDA
210’	220’	230’	240’	250’
PRHICEINFL	CVHKQLREKR	LAPILIKEVT	RRVNRTNVWQ	AVYTAGVLLP
260’	270’	280’	290’	300’
TPYASGQYFH	RSLNPEKLVE	IRFSGIPAQY	QKFQNPMAML	KRNYQLPNAP
310’	320’	330’	340’	350’
KNSGLREMKP	SDVPQVRRIL	MNYLDNFDVG	PVFSDAEISH	YLLPRDGVVF
360’	370’	380’	390’	400’
TYVVENDKKV	TDFFSFYRIP	STVIGNSNYN	ILNAAYVHYY	AATSMPLHQL
410’	420’	430’	440’	450’
ILDLLIVAHS	RGFDVCNMVE	ILDNRSFVEQ	LKFGAGDGHL	RYYFYNWAYP
460’				
KIKPSQVALV	ML*			

The crude lysate containing DNAse and protease inhibitor cocktail tablets was applied to the column and eluted as fractions with a 20 mM to 1 M imidazole gradient (Figure 5.4). The fractions were assayed by SDS-PAGE. Those found to contain ldNMT were pooled and diluted in anion exchange buffer in preparation for the second purification step. The protein solution was loaded on a HiTrap Q FF 5 ml column and eluted in fractions with a 20 mM to 500 mM salt gradient. Fractions found to contain ldNMT protein were pooled together and concentrated (for details see Section 10.3 on page 176). Several expression and purification tests were performed. Although the protein expresses well, it is not very soluble (Figure 5.3). The final yield from 500 ml of cell culture is about 0.7 mg. One major issue during these tests was protein appearing triple-banded

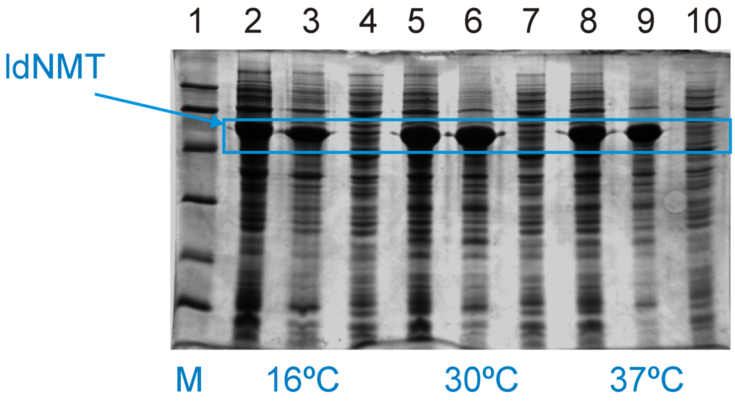
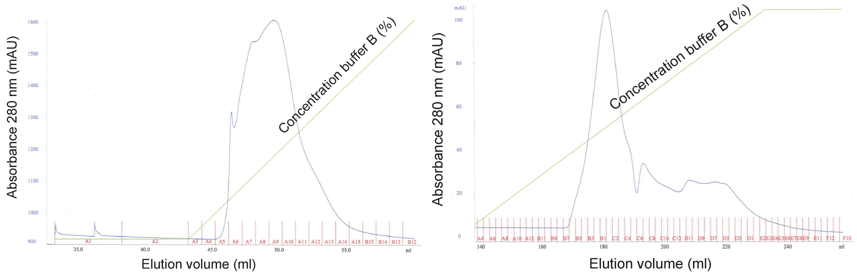


Figure 5.3: **SDS gel of initial expression trial.** ldNMT is expressed well as one can see from lanes 1, 5 and 8 (expression at 16, 30 and 37℃ respectively) containing the total lysate. However, most of the protein is insoluble which show lanes 3 and 4, 6 and 7 and 9 and 10, where the first of each pair represents the insoluble fraction and the second the soluble fraction (expression at 16, 30 and 37℃ respectively). (1) marker, (2)–(4) total lysate, insoluble fraction and soluble fraction when expressed at 16℃, (5)–(7) total lysate, insoluble fraction and soluble fraction when expressed at 30℃, (8)–(10) total lysate, insoluble fraction and soluble fraction when expressed at 37℃. Cells were lysed in Milli-Q water.

on the gels (Figure 5.5, lanes 5–8) giving strong indication of degradation.



(a) Nickel-NTA

(b) Anion Exchange

Figure 5.4: **Chromatograms of ldNMT.** Two chromatograms show the UV absorbance (blue line), as ldNMT is eluted from the nickel-NTA column (a) and eluted from the anion exchange column (b). The green line shows the gradient of buffer B (imidazole and salt respectively).

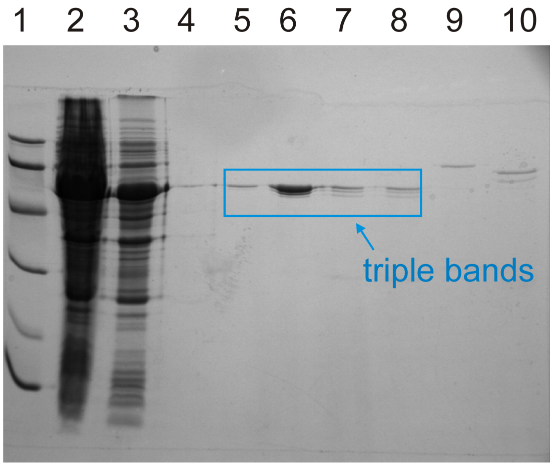


Figure 5.5: **Gel of double His-tagged ldNMT showing triple bands.** On the gel, double His-tagged ldNMT is shown during the purification. One can see that the protein still seems full length after the nickel chromatography (lane 2, no degradation visible). However, the eluted fractions after the second purification step, ion exchange chromatography, indicate triple bands on the gel (lanes 5–8) assuming degradation of the protein. (1) marker (2) lysate (3) flow through after nickel chromatography (4) flow through after anion exchange (5)–(10) selected fraction after anion exchange chromatography

5.3 Mass Spectrometry

Mass spectrometry (MS) can be used to analyse proteins by measuring the mass by charge ratio after ionising the sample. To confirm ldNMT degradation, a sample was handed to Simon Grist who kindly performed the experiments on the electrospray mass spectrometer (Section 10.4). It turned out that the double His-tagged NMT is degrading from both ends.

The MS experiment shows three well-defined peaks at 53,209 Da, 51,712 Da, 50,834 Da (Figure 5.6) which could likely explain the triple band on the gel (Figure 5.5, lanes 5–8). The calculated mass of the protein is 53,128.5 Da. The masses of the three peaks can be explained with the sequence of double His-tagged ldNMT when one assume one species corresponds to the phosphorylated full length protein, one species misses the sequence MGSSHHHHHHSSG and another species misses MGSSHHHHHHSSGLVPRGSH. Knowing the protein solution also contains full length protein, some tests with SPR were performed (Section 5.5). The aim was to capture the double His-tagged protein on the chip while washing off the degrading species containing only one His-tag.

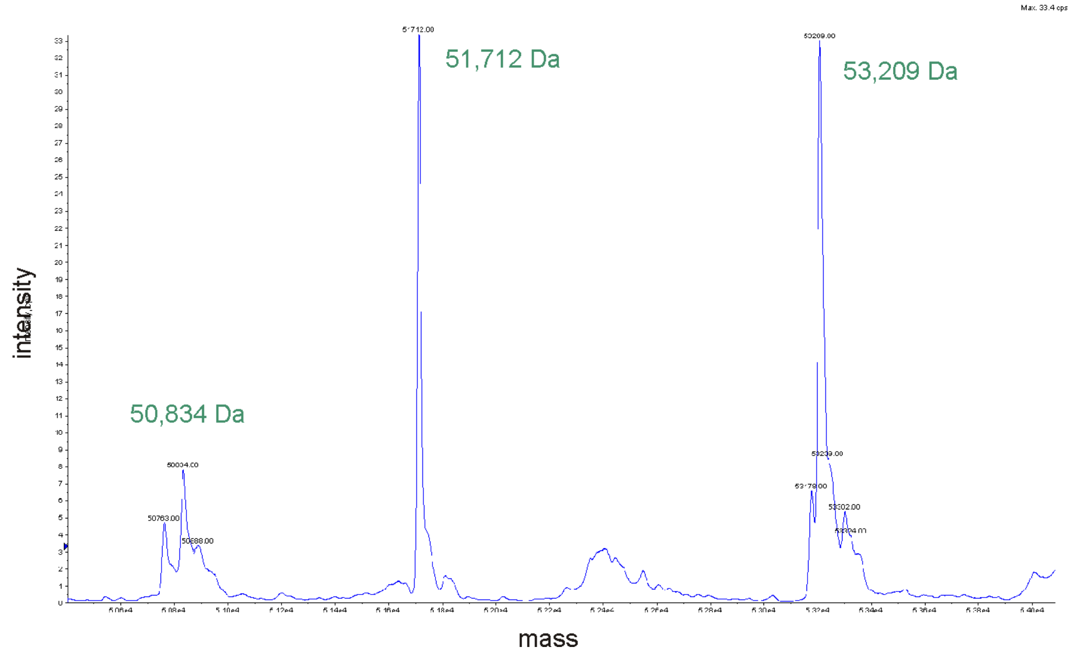


Figure 5.6: **Mass Spectrometry results of ldNMT.** Three clear peaks appeared at 50,834 Da, 51,712 Da and 53,209 Da which can be related to the triple bands on the gels after ion exchange chromatography.

5.4 lmNMT and tbNMT

Expression cells containing lmNMT and purified tbNMT (both N-terminal His$_6$-tagged) were kindly provided by Dr. Jim Brannigan (Section 10.3). Together with double His$_6$-tagged ldNMT, all three NMT proteins were tested by SPR experiments for their suitability for fragment screening.

5.5 SPR with NMT

Double His$_6$-tagged ldNMT, single His$_6$-tagged lmNMT and tbNMT were assayed with SPR on a Biacore T100 machine. Attachment of the proteins via their His-tags to the NTA of the chips was attempted. However, the proteins were bleeding off the chip very quickly. To prevent this, the proteins were covalently immobilised by activating the carboxymethyl groups of the dextran coating with EDC/NHS, assembling the protein via its His-tag and covalently binding it via its lysine residues, followed by deactivation with ethanolamine. A typical immobilisation sensorgram is shown in Figure 5.7 a. However several challenges appeared when performing the experiments including inexplicable buffer effects, baseline drift and detection of cofactor binding (Figure 5.7 b–f for a summary of representative examples with tbNMT).

The following points summarise the experiments performed with NMT:

- Although knowing that ldNMT degraded, it was tested in HBS-P buffer (0.01 M HEPES pH 7.4, 0.15 M NaCl, 50 µM EDTA, 0.05% Surfactant P20) with the aim to catch the full length constructs on the NTA chip and wash the degraded species away. However, the protein did not stick to the chip and the double His$_6$-tag did not show improved binding.
- LmNMT was tested in HBS-P buffer. It showed a high drop off rate and also the response of MCoA was < 2 RU (response units).
- TbNMT was covalently bound to an NTA chip using HBS-P buffer. This way, the proteins should all be oriented the same way without being washed away during the experiment. After immobilisation, the buffer was changed to Tris (200 mM NaCl, 50 mM tris-HCl, pH 7; following a protocol obtained from collaborators in Dundee). Simple running buffer injections showed unexpected buffer effects on the sensorgrams. The control MCoA did not show any binding. When the protein was immobilised de novo, a response of 36 RU could be observed for MCoA, however protein bleed-off was still experienced. Since the bleed-off is rather linear, several inhibitors were tested. However these tests proved to be challenging because the baseline did not return to the original level and the ligands seemed to accumulate on the chip.

5.6 Concluding Remarks

NMT proved to be a very challenging target. An SPR assay could not be established with any of the three different NMT proteins. Whilst working with NMT, the test protein hen egg white lysozyme (HEWL) was used to establish parts of the screen and test some of the fragments. Initial experiments with HEWL looked very promising, thus the NMT project was postponed until a later stage.

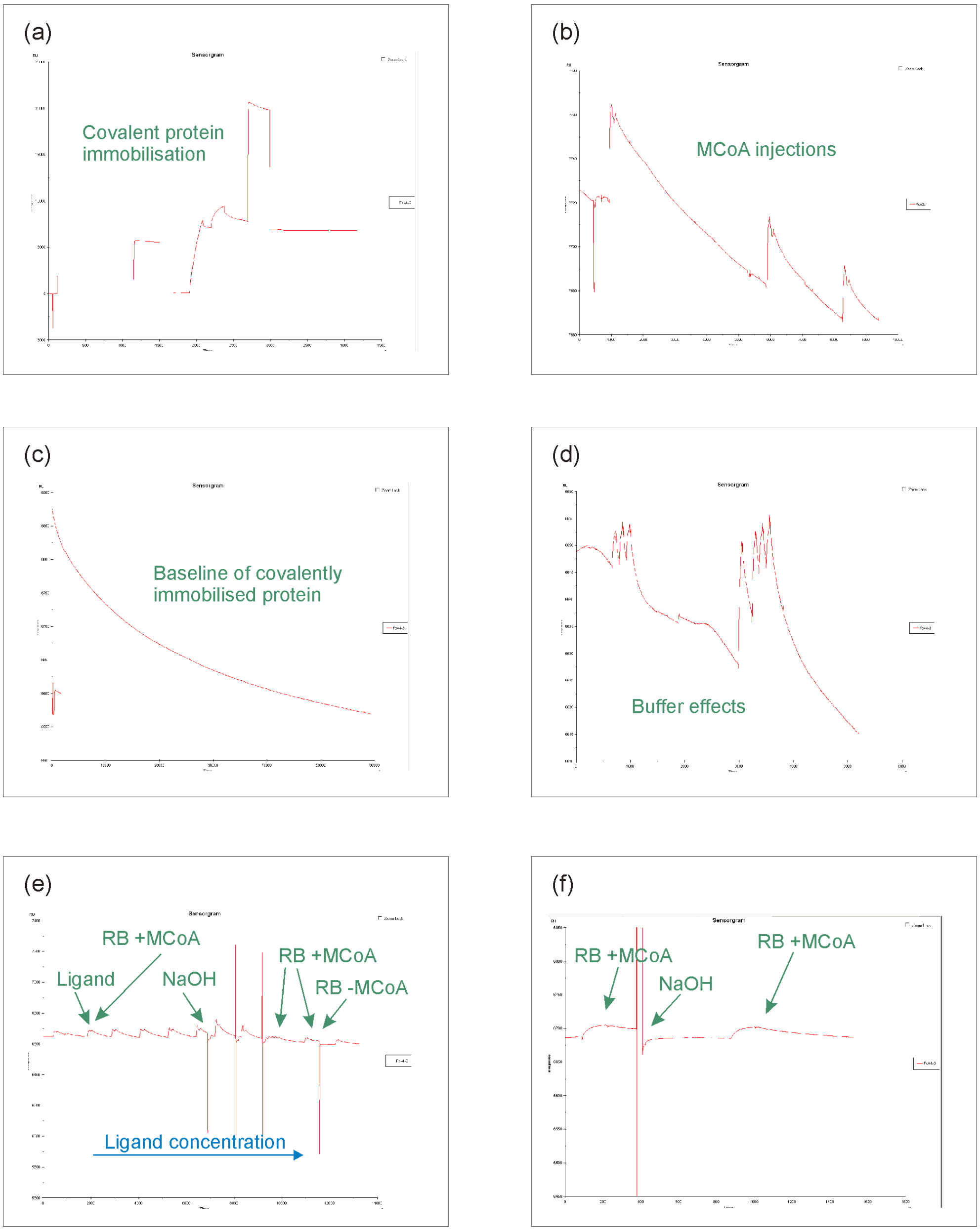


Figure 5.7: **Examples of SPR experiments with tbNMT.** (a) A typical covalent immobilisation curve. tbNMT was immobilised to 8,000 RU. (b) MCoA injected three times to a chip with covalently immobilised protein. Although the baseline drifts severely a response of 40–50 RU can be observed. (c) Covalently attached protein is bleeding of the chip. Overnight loss of about 3,000 RU. (d) Unexpected buffer effects. (e) Test ligand with expected potency against tbNMT of 2 µM builds up on the chip. Buffer injections (RB) with (+) and without (-) MCoA and 5 mM sodiumhydroxide as regeneration to clean the chip show unexpected responses. Running buffer injections exhibit binding curves. (f) Running buffer injections show a binding effect. Binding compounds seems to accumulate on the chip. A wash with 5 mM NaOH brings the curve below the original baseline.

Chapter 6

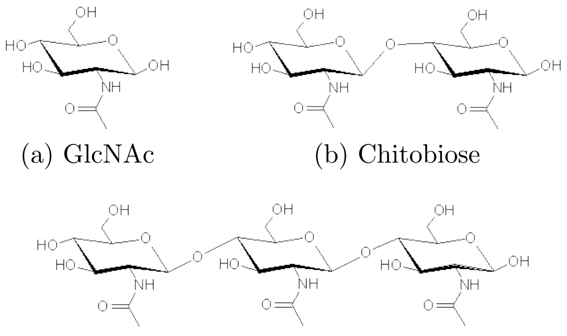
Fragment Experiments with HEWL

This chapter describes the experiments in fragment screening that were performed for the protein Hen Egg White Lysozyme (HEWL). The work fell into four main parts. A novel Biacore assay was developed where HEWL binds to the dextran surface of a CM5 Biacore chip. After optimisation, the assay was used to screen 50 fragments from the "Michele_1" library, and potential hits were validated with an established cell based activity assay. Subsequently, some fragments from the "Michele_2" library were screened with thermal shift analysis. The issues with processing the TSA data led to the development of the MTSA program described in Chapter 4. Initial crystallisation trials have been performed when further work was halted as a more promising target (BtGH84, Chapter 7) became available.

6.1 Background

Hen Egg White Lysozyme (HEWL) is a glycanhydrolase of 14 kDa. The protein is often used as a model system because large quantities of stable, pure protein are easily available and the protein readily crystallises. HEWL lyses cells by cleaving the $\beta-(1 \rightarrow 4)$ glycosidic bonds of the murein cell wall (Mörsky, 1983). The enzyme was first discovered by Sir Alexander Fleming in 1922, when some drops from his nose fell on an agar plate and killed the bacteria colonies on it. It has become one of the most studied proteins in history and thus it is an excellent model system. There are a lot of known inhibitors, but most of them are carbohydrate like. These properties make HEWL a suitable target for a fragment screen to test the properties of the newly generated libraries described in Chapter 3.

For the following assays, the known inhibitors N,N'-diacetylchitobiose (chitobiose) (Figure 6.1 b) and N,N',N"-triacetylchitotriose (chitotriose) (Figure 6.1 c) were used as control compounds. Chitotriose has an expected K_D of 7–10 µM. Chitobiose with a K_D of 170 µM lies in the range of a good fragment hit. Another sugar which binds very weakly to HEWL is N-Acetylglucosamine (GlcNAc) (Figure 6.1 a) that has a literature K_D of 40–60 mM. All of the mentioned K_Ds were obtained at pH 5.0 (Dahlquist et al., 1966). Glucose itself binds so weakly to HEWL that it can be considered as a negative control and is used as such.



(a) GlcNAc

(b) Chitobiose

(c) Chitotriose

Figure 6.1: **GlcNAc, chitobiose and chitotriose.** Structures of (a) GlcNAc (N-Acetylglucosamine), (b) chitobiose (N,N'-diacetylchitobiose) and (c) chitotriose (N,N',N"-triacetylchitotriose) in Howarth representation.

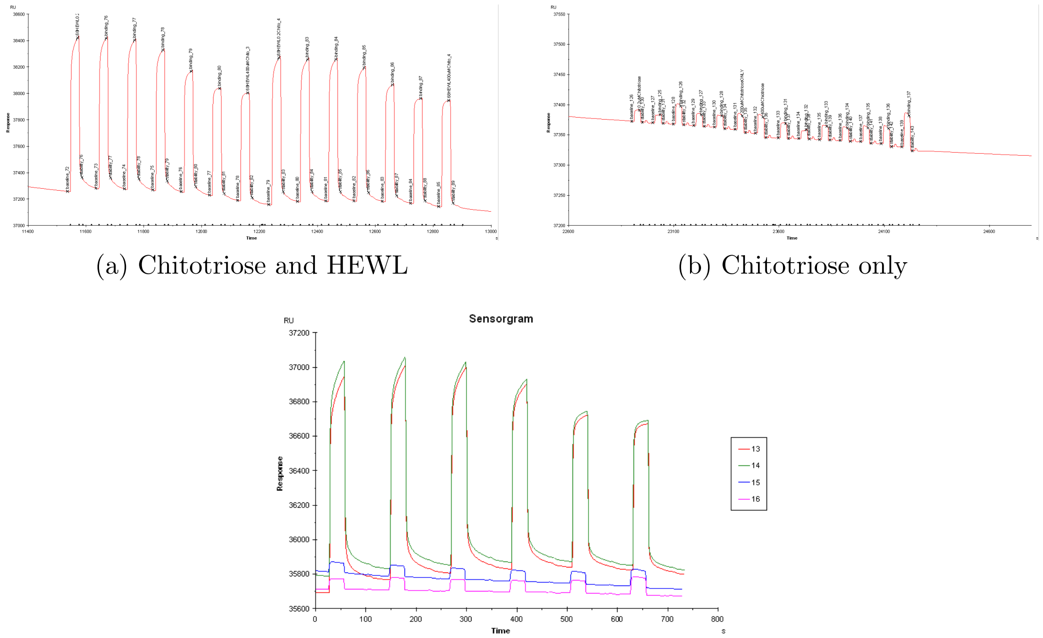
6.2 The Alternative Biacore Screen with HEWL

CM5 chips are carboxymethylated dextran chips which are widely used for Biacore experiments (for an introduction to SPR see Section 10.5 on page 180). In a conventional Biacore experiment, molecules are coupled to the surface via NH_2, -SH, -CHO, -OH or -COOH after special preparation. However, when HEWL is injected onto an unprepared CM5 chip with the Biacore technology, the protein binds to the dextran surface of the chip. This phenomenon was used to develop this alternative Biacore screen, where the binding of a competitive ligand will be seen by reducing the quantity of HEWL that binds to the plain surface of the chip. 1–2 mg/ml protein generates a signal of about 1000 response units (RU), so a competitive ligand will give a much larger change in the signal compared to the response of about 10 RU from binding of a fragment to HEWL conventionally immobilized on the chip. The new screen described here is able to give responses of about 200 RUs or more. Chitotriose with an expected K_D of 7–10 µM reduces the protein signal by about 250 RU. Since its K_D of 170 µM lies in the range of a good fragment hit, chitobiose was used as a positive control in the following experiments. The stronger inhibitor chitotriose was used for initial tests. Relating to the "conventional Biacore screen", the new screen described here is referred to as "alternative screen". Details of the method are described in Section 10.5.3 on page 182.

6.2.1 Screen Using the Alternative Biacore Assay

HEWL was screened with the Biacore T100 in PBS buffer with 5% DMSO in 96-well plates. The setup needs only one of four available flow cells compared to the two flow cells required for the conventional screen. Therefore, the old reference cell of an used CM5 chip can be recycled. The plate was set up with alternating wells of fragment only and fragment plus protein. Figure 6.2 illustrates the sensorgrams of a typical experiment: A concentration series of ligand with a fixed amount of protein is injected onto one flow cell of the sensor chip. The peak height indicates how much protein is binding to the surface of the chip. However, the curved shape of the response peak shows that the binding is not saturating, which could be due to additional non-specific binding. However, NSB (carboxymethyl dextran sodium salt in 0.15 M NaCl containing 0.02% NaN_2) was found not to reduce the non-specific binding or accumulation on the surface. When the compound is competing for the active site of the protein, less protein can bind to the chip and the response is alternated, i.e. reduced for inhibitors and increased for activators.

Several HEWL concentrations were tested to estimate which concentration gives a reliable response, but also does not overload the chip with protein. A HEWL concentration of 1 mg/ml HEWL (68 µM) was found to be most suitable. When the response peaks are plotted against the ligand concentration, dose-binding



(a) Chitotriose and HEWL

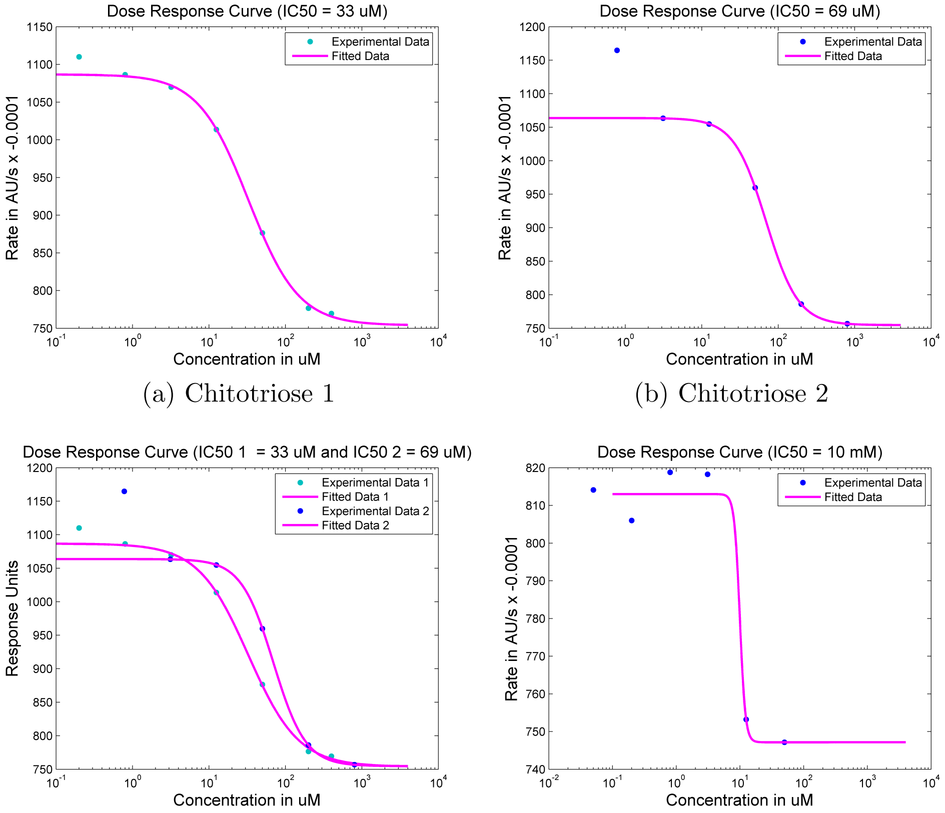
(b) Chitotriose only

(c) Chitotriose and HEWL

Figure 6.2: **Sensorgrams of the alternative Biacore screen.** HEWL concentration at 68 µM (a) Chitotriose and HEWL in buffer without DMSO. This is a screenshot of one of the initial tests. (b) Another screenshot of the initial test with chitotriose only in buffer without DMSO. These peaks are subtracted from the curves with HEWL and chitotriose to eliminate the ligand response. (c) HEWL with chitotriose in buffer without DMSO from an automated experiment. Cycle 13 (red) and cycle 14 (green) show the concentration series of chitotriose together with HEWL, cycle 15 (blue) and cycle 16 (pink) show the concentration series of chitotriose only.

curves can be deduced. All points plotted and fitted in the following figures are the average of two separate measurements. The positive controls chitobiose and chitotriose tested in DMSO free buffer produce IC_{50} in the expected range (Figure 6.3). For chitotriose, data fitted with the method described in Section 10.6 (page 183) generate an IC_{50} of 33 µM in Figure 6.3 a, and in a later experiment an IC_{50} of 69 µM in Figure 6.3 b. The averaged IC_{50} for chitotriose found in the assay is thus at 51 µM and for GlcNAc at 10 mM. The literature K_D is 7–10 µM for chitotriose and 40–60 mM for GlcNAc. The corresponding sensorgrams for these plots are shown in Figure 6.2. The point of the lowest concentration was considered to be an outlier and had to be excluded from the fit.

Figure 6.4 shows some examples of the difficulties that can arise. It is possible



(a) Chitotriose 1

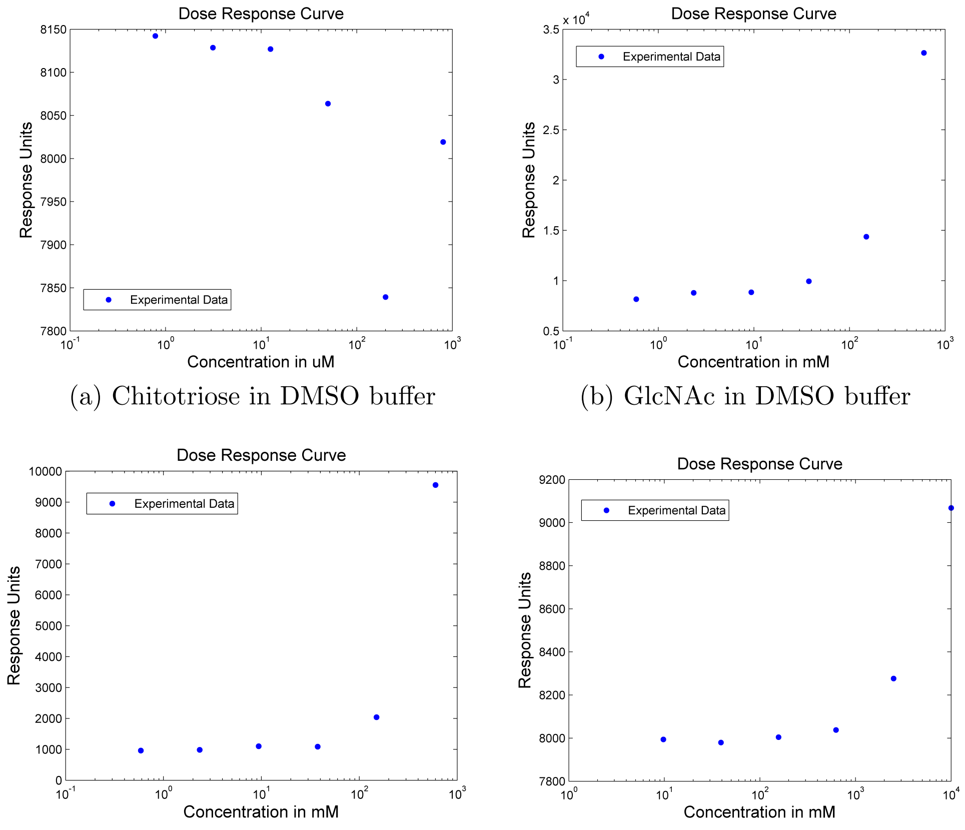
(b) Chitotriose 2

(c) Chitotriose 1 and 2 overlaid

(d) GlcNAc

Figure 6.3: **Dose-response curves that can be fitted.** The examples are of chitotriose and chitobiose in DMSO free buffer. The data points derive from the top of the peaks shown in the sensorgrams shown in Figure 6.2. Plotted are the averages of duplicate measurements of which the response of the ligand only was subtracted. (a) IC_{50} of 33 µM for chitotriose. (b) IC_{50} of 69 µM for chitotriose in a second experiment. (c) Overlay of the two chitotriose dose-response data. (d) GlcNAc with half concentration of HEWL (34 µM) and an IC_{50} of 10 mM. (n=2)

to fit the dose-response curves for chitotriose when in regular buffer (Figure 6.3 c), but if 5% DMSO were present, the data could not be fitted (Figure 6.4 a). Nevertheless a reduction of the signal can be observed. Chitobiose which is expected to bind in the affinity range of a fragment did not produce data that could be fitted (Figure 6.4 d). At the highest concentration, data points of the weakly binding GlcNAc and with medium affinity binding chitobiose show an increase in signal (Figure 6.4 b and d). The same happens for the signal of glucose at high concentration although this was tested in DMSO free buffer. For none of these data sets could curves be fitted.



(a) Chitotriose in DMSO buffer

(b) GlcNAc in DMSO buffer

(c) Glucose

(d) Chitobiose in DMSO buffer

Figure 6.4: **Non-fittable dose-response curves.** (a) Chitotriose in 5% DMSO buffer. (b) GlcNAc in 5% DMSO buffer (c) Negative control glucose. (d) Chitobiose in 5% DMSO buffer; the 10 mM is only plotted as a single data point because the duplicate was an outlier. (n=2)

The tests reveal that an IC$_{50}$ determination with the alternative Biacore screen can be challenging. However, binding of a ligand is clearly indicated when the ligand is injected in a moderate concentration. An initial plate with 48 fragments from the "Michele_1" fragment library was tested for further assessment of the method (structures are shown in Appendix D). Figure 6.5 is an example of the typical output from the assay, for the example of ysbl000266.

The following empirically derived classification was developed to analyse the binding of the individual compounds in a spreadsheet:

- **Inhibiting** $R(P : Chito) < R(P : F) - R(F) < R(P) - [R(P) - R(P : Chito)) * 0.3]$

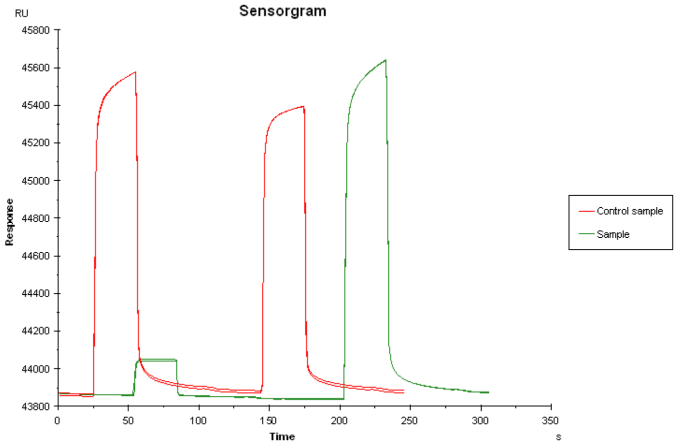


Figure 6.5: **Fragment screen example sensorgram.** For compound ysbl000266 the sensorgram of a full cycle is shown. The injections follow in the order: 1. fragment (1st peak green), 2. fragment and protein (2nd peak green), 3. fragment (1st peak green), 4. fragment and protein (2nd peak green), 5. protein (1st peak red), 6. chitotriose and protein (2nd peak red). If the fragment plus protein (big green) peak minus the fragment only (small green) peak is significantly shorter than the protein only (big red) peak, then the compound is likely to bind. (n=2)

- **Activating** $R(P) + [R(P) - R(P : Chito]) < R(P : F)$
- **Otherwise interesting** $R(P : F) - R(F) < R(P : Chito) * 0.9$

with $R(P : F)$:= response of protein with fragment, $R(F)$:= response of only fragment, $R(P : Chito)$:= response of protein with chitotriose , $R(P)$:= response of protein only.

A number of compounds gave some response in this assay – both inhibiting and increasing the binding of HEWL to the Biacore chip surface. The compound numbers are summarised in Table 6.1 (structures in Appendix D). These compounds were investigated further.

6.2.2 Binding Affinity Assay with HEWL

To confirm the binding of the fragments detected with the alternative Biacore binding screen, a follow-up assay was established to obtain the binding affinity data for seven selected compounds (ysbl000266, ysbl000268, ysbl000273,

Table 6.1: **Fragments hits with the alternative Biacore assay on HEWL**

YSBL database entry		
Inhibiting	**Activating**	**Interesting**
ysbl000266	ysbl000265	ysbl000267
ysbl000268	ysbl000279	ysbl000303
ysbl000273	ysbl000294	ysbl000304
ysbl000276	ysbl000297	
ysbl000277	ysbl000298	
ysbl000281		
ysbl000291		
ysbl000293		
ysbl000295		
ysbl000300		

ysbl000276, ysbl000277 and ysbl000281). ysbl000261 was used as a negative control because it did not show any binding in the initial screen. Chitobiose served as positive control. Figure 6.6 pictures a bar chart of the responses of the concentration dependent screen. The first two sets show the fragment response, the second two sets the response of protein and fragments. From the left to the right, the ligand concentration rises. The tops of the bars of the protein and fragment set indicate a sigmoid shaped curve.

However, after correction (subtracting the fragment response as background from the fragment plus protein response, some of the dose-response data are plotted in Figure 6.7), the following results became obvious: For chitobiose (Figure 6.7 a) and for the compounds ysbl000268 (Figure 6.7 c), ysbl000273 and ysbl000281 (Figure 6.7 d) the beginning of a downward pointing dose-response curve is indicated. However, the responses do not saturate. For the negative control ysbl000261 and also for ysbl000266 (Figure 6.7 b), the curves had an upward trend at higher concentrations. For ysbl000276 and ysbl000277 the curves were relatively flat and did not confirm any binding. Thus no dose-response curve using equation 10.1 on page 183 could be fitted.

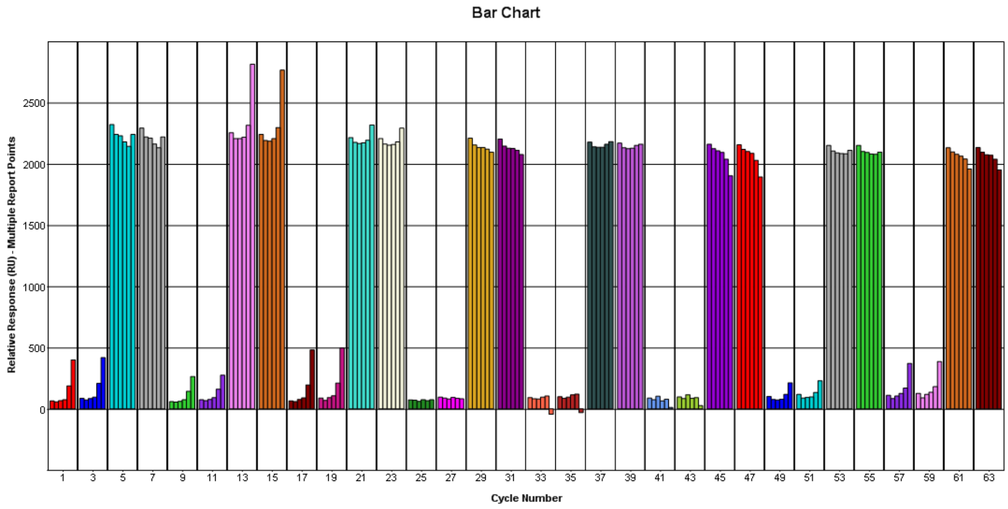
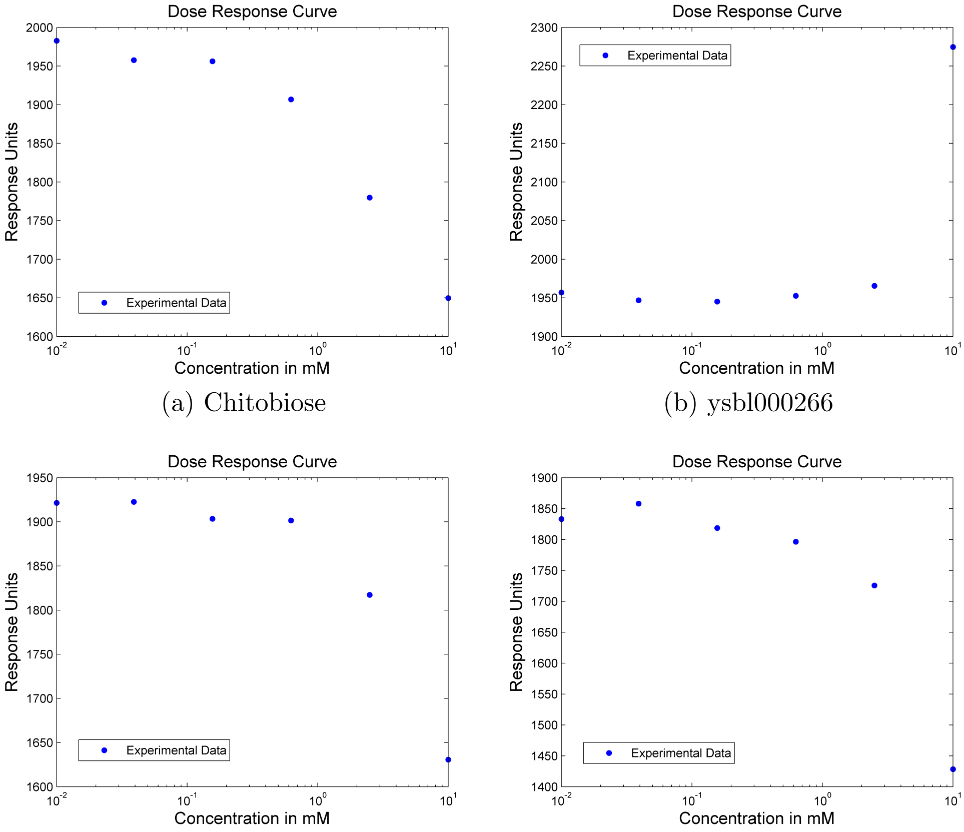


Figure 6.6: **Compound binding in the alternative HEWL screen** The bars represent the peak height (i.e. response unit in the sensorgrams) for each injection. The lower bars are the responses for the fragment alone (used as a blank subtraction); the higher bars are protein with fragment. In the event of binding, the shape of the tops of the protein-ligand bars (minus the fragment bars) should form a sigmoid curve. The values of the subtracted bars (height) can also be plotted to obtain IC_{50} curves such as in the following Figure 6.7. Cycles 1–8 positive control chitobiose, 9–16 ysbl000266, 17–24 ysbl000268, 25–32 ysbl000273, 33–40 negative control ysbl000261, 41–48 ysbl000276, 49–56 ysbl000277 and 57–64 ysbl000281.

6.2.3 Discussion of the Alternative Biacore Screen

The alternative Biacore screen is an interesting way of screening carbohydrate binding proteins. The major advantages are the relatively high responses compared to the conventional Biacore screen and the need of only one flow cell on a Biacore CM5 chip. Old chips can be recycled and less consumables are needed. Nevertheless the assay is challenging. Many side effects occur such as protein and ligand building up on the surface as a consequence of the high concentrations. Some of the data sets fit very nicely and give good dose-response curves (Figure 6.3), others are not fittable and show outliers (Figure 6.4). Most problems can be explained by the high concentrations of protein and ligand used. The molecules can be too sticky or precipitate resulting in odd signals. Figure 6.4 b and c illustrate how the small molecules glucose and GlcNAc accumulate heavily on the surface at higher concentrations. Also the fragments ysbl000266 and the negative



(a) Chitobiose

(b) ysbl000266

(c) ysbl000268

(d) ysbl000281

Figure 6.7: **Binding affinity screen.** (a) The responses demonstrate chitobiose binding, but do not indicate saturation. Compounds ysbl000266 (b), and ysbl000261 (negative control) accumulate on the surface and give a higher response with higher concentration. ysbl000268 (c) and ysbl000281 (d) show the beginning of an inhibition curve, but no saturation occurs like with chitobiose. (n=2)

control ysbl000261 give a higher response with higher concentration (Figure 6.7 b). Figure 6.3 c displays how chitotriose gave respectable IC_{50} curves in regular buffer, but produced curves with outliers when DMSO was present (Figure 6.4 a). Although the fit for GlcNAc in regular buffer was possible and produced affinity data in an expected range, the shape of the curve is not of desired quality (Figure 6.3) and the determined IC_{50} can only be taken as a guide. In the binding affinity screen where the maximal concentration of chitobiose (K_D of 170 µM) was 10 mM the concentration should give a full IC_{50} curve for chitobiose (Figure 6.7 a). However the curve does not saturate. Considering these results, it is also important to recall that the IC_{50} is not the same as the K_D. The IC_{50} is not an absolute value and depends on the assay conditions. In the alternative Biacore screen, the IC_{50} depends on the affinity and concentration of dextran on the chip surface as well as on the protein concentration. In summary the higher the protein concentration, the higher the IC_{50}. That fact can explain why the curves do not reach saturation. In conclusion, the alternative Biacore screen is an interesting new approach to screen carbohydrate binding proteins. However, the screen requires more optimisation. At the present it is difficult to obtain reliable IC_{50} values (Figure 6.4). However, the alternative screen gave indication of fragments binding to HEWL. The next section describes cross-validation of this fragment binding using an enzyme activity assay using whole cells as substrate.

6.3 The Enzyme Activity Assay

HEWL is able to lyse some bacteria by cleaving $\beta-(1 \rightarrow 4)$ glycosidic bonds of N-acetylmuramic acid and N-acetyl-D-glucosamine residues in mucopolysaccharide cell walls. *Micrococcus luteus* – formerly known as *Micrococcus lysodeikticus* – can be used as the substrate for a HEWL activity assay (first suggested by Shugar 1952). The turbidity of the solution with cells and enzyme can be recorded at a 450 nm wavelength. As HEWL cleaves the cell walls the turbidity decreases. When a ligand is present the HEWL activity should be either inhibited or activated and therefore the cleavage rate - measured in absorption units per second (AU/s) - slows down or increases respectively.

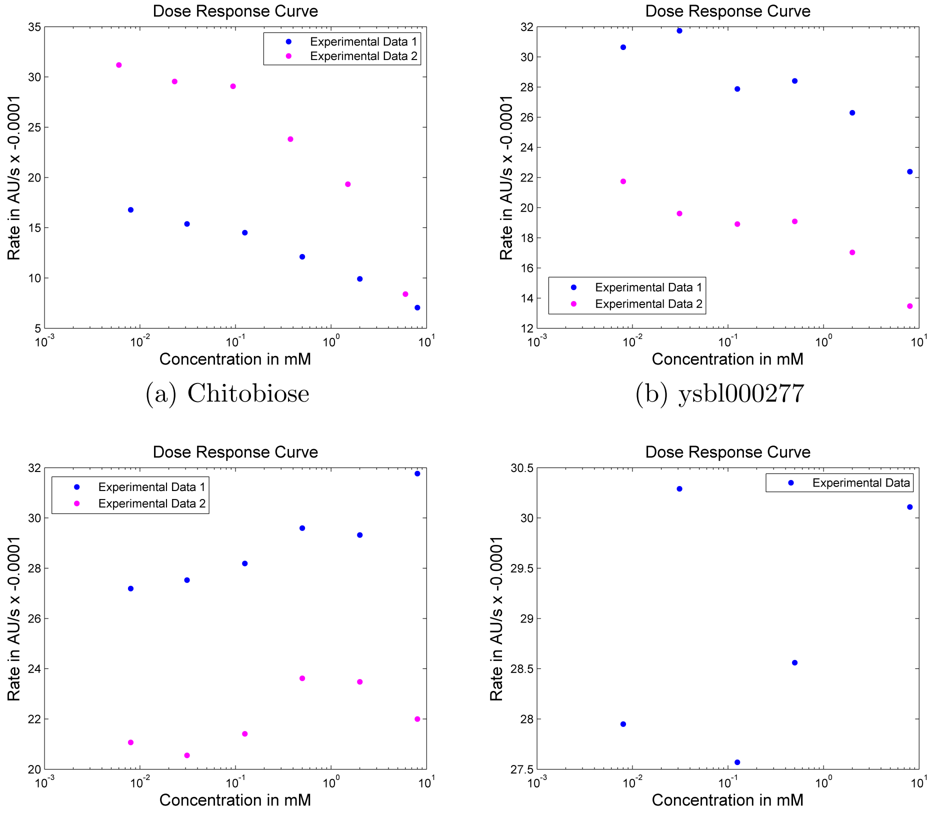
The assay was set up in 96-well format on a plate reader with 2.3–4.6 µM HEWL and 0.33–0.5 mg/ml *Micrococcus luteus*. The rate of cell lysis was used to generate dose-response curves for the different ligands. The rate AU/s (alias the slope) was determined with Excel. The points of the first 30 seconds were excluded to allow the system to stabilise. In most cases the slope was determined for periods of 30–230 seconds to ensure a linear slope. The stronger an inhibitor binds to HEWL, the less negative the slope should be in theory. The slope was altered by multiplication with the factor $-10,000$ in order to give the dose-response curves the classical sigmoidal shape (low concentration corresponds to high response and vice versa, similar to the previous Biacore assays). If the lowest concentration was 0, then a value of 0.01 was used instead to plot on the logarithmic scale (methods in Section 10.7 on page 183).

In Figure 6.8 some representative results of the enzyme activity assay are illustrated. To demonstrate reproducibility, all data were recorded as duplicates. Chitobiose shows a hint of inhibition (Figure 6.8 a) with the response falling for about 10–20 × -0.00001 absorption units at maximum ligand concentration. However, the curve does not reach saturation. Also compound ysbl000277 (Figure 6.8 b) indicates some inhibition with the response decreasing for about 7 × -0.00001 absorption units. Like chitobiose the inhibition curve does not reach saturation. However, the point at the highest concentration was found to be an outlier and therefore the compound was not considered as a potential inhibitor. Compounds ysbl000281 and ysbl000294 (Figure 6.8 c and d) show an oscillating response within the range of about 5 × -0.00001 absorption units. For fragment ysbl000281, the data points indicate a weak activation rather than an inhibition. However, considering the change of response being much smaller than for chitobiose, the compound unlikely shows activity. Fragment ysbl000294 does not show activity.

The plots in Figure 6.9 illustrate the most interesting fragments of the screen: N-acetylglycine (ysbl000265), o-toluic acid (ysbl000267) and benzothiazole (ysbl000 297) (Figure 6.10). All three show activating activity. The range of the response, 20 × -0.00001 absorption units, is in almost the same range as for chitobiose. More solid of the compounds was purchased and they were tested in further enzyme activity and thermal shift experiments (Sections 6.3.1 and 6.4.3).

Figure 6.11 shows the results of control experiments. Initially, the cells were incubated with each of the three fragments (N-acetylglycine, o-toluic acid or benzothiazole) in the absence of HEWL. No cell lysis was observed. Subsequent addition of 4.6 µM HEWL to the same sample, the cells are lysed as previously.

The pH of the solution itself remains in the normal range after the lysis: pH 7.4 for N-acetylglycine, for o-toluic acid and benzothiazole, and pH 8.0 if chitobiose is present. *Micrococcus luteus* cells without any additives in Milli-Q water have a pH of 6.3.



(a) Chitobiose

(b) ysbl000277

(c) ysbl000281

(d) ysbl000294

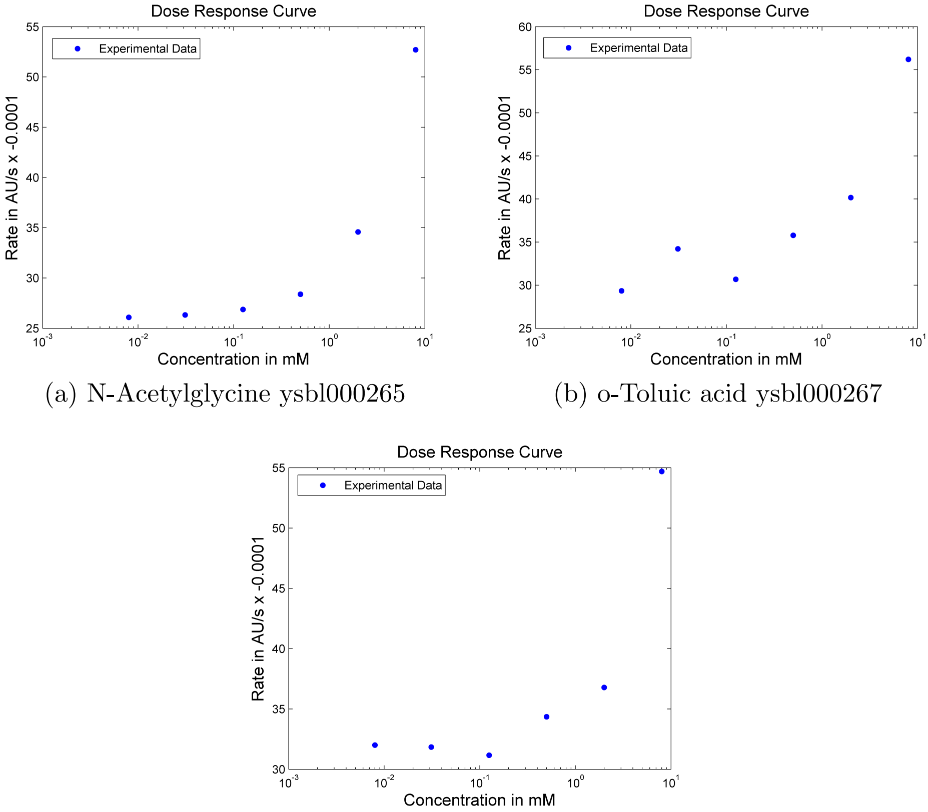
Figure 6.8: **Results of the enzyme activity assay.** (a) Chitobiose shows inhibition, but no saturation. (b) ysbl000277 also demonstrates inhibition, but no saturation. The point as the highest concentration has to be considered as an outlier. (c) The response of ysbl000281 is a weak activation rather than an inhibition. (d) Fragment ysbl000294 does not indicate neither inhibition or activation. (n=1)

6.3.1 Determination of the Michaelis Menten Constant

Enzyme kinetics can be modelled with the popular but simple Michaelis-Menten model. In this specific experiment, the Michaelis-Menten constant K_M of the assay is the cell concentration at which the HEWL activity is half maximal (respectively the substrate concentration where the enzyme activity is half maximal).

The K_M is obtained with the following equation:

$$Rate = V = \frac{V_{max}[S]}{[S] + K_M} \tag{6.1}$$



(a) N-Acetylglycine ysbl000265

(b) o-Toluic acid ysbl000267

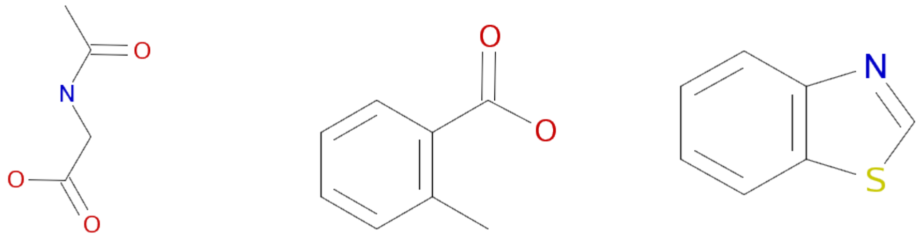
(c) Benzothiazole ysbl000297

Figure 6.9: **Interesting dose-response curves of the enzyme activity assay.** (a)–(b) Dose-response curves of ysbl000265, ysbl000267 and ysbl000297.

Knowing K_M the inhibition constant K_i which is independent of the assay conditions, unlike the IC_{50}. The K_i represents the competing ligand concentration which would bind to 50% of the receptor at equilibrium in the absence of other competitors. Thus K_i and K_D are in most cases identical. Exceptions will occur when the ligand binds to a form of the receptor prior modified by the substrate or when the receptor itself modifies the substrate. The Cheng-Prusoff-equation describes the relationship between K_i and IC_{50} (Cheng and Prusoff, 1973):

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_M}} \tag{6.2}$$

It was not possible to determine the K_M as the enzyme activity does not reach a saturating maximum. Instead activity decreases after reaching a certain concen-



(a) N-Acetylglycine (b) o-Toluic acid (c) Benzothiazole

Figure 6.10: **Structures of potential activators.** (a)–(c) Structures of ysbl000265, ysbl000267 and ysbl000297.

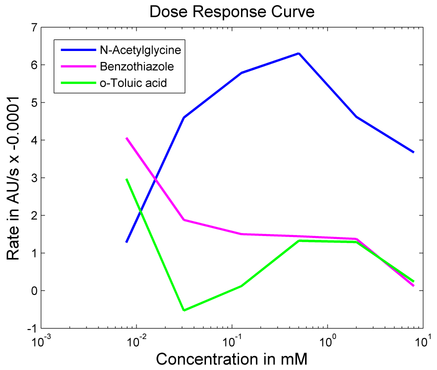


Figure 6.11: **Validation of the potential activators.** When HEWL is not present, no concentration dependant change of rate can be observed.

tration and does not stay at the asymptote. Also a minimum asymptote cannot be found. This is illustrated in Figure 6.12 which shows a substrate velocity curve for the substrate *Micrococcus luteus*. Instead of reaching a maximum and saturation, the enzyme activity decreases again after 0.23 mg/ml substrate concentration. These data cannot be fitted and thus were not used to determine K_M. The experiment was repeated without ligand as well as with 10 mM of the activators N-acetylglycine, o-toluic acid and benzothiazole. The pH of the solutions was found to be 7.4 for no ligand present, 7.4 for N-acetylglycine, 6.6 for o-toluic acid and 7.0 for benzothiazole. The results of this test are illustrated in Figure 6.13. This time the curves neither saturate nor reach a maximum. The shape of the curves in Figure 6.12 could not be reproduced.

Since these three ligands are expected to be activators, it was attempted to find the point of maximum activation at the substrate concentration of 0.17 mg/ml,

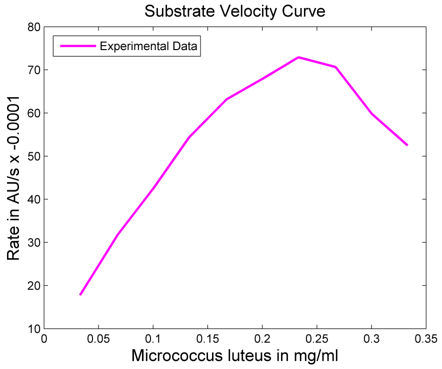


Figure 6.12: **Cell concentration series for K_M determination.** Cell concentration was varied to find the K_M of the assay where the cell concentration allows a half maximal enzyme activity. However, instead of saturating at the maximal activity, the curve goes down again.

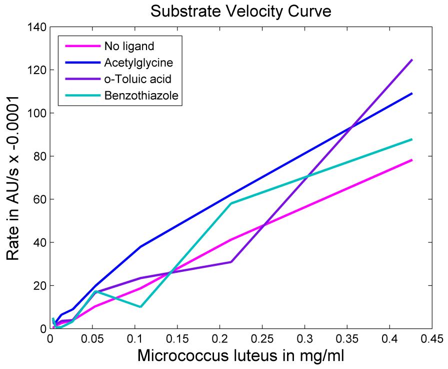


Figure 6.13: **Cell concentration series for K_M determination with ligands.** Cell concentration was varied to find the K_M of the assay where the cell concentration allows a half maximal enzyme activity. However the curves do not reach a maximum rate. The ligands were present at 10 mM concentration.

which lies in the range of the maximum enzyme activity (see above). The compounds were tested up to a concentration of 25 mM. However, benzothiazole precipitated at 25 mM. The dose-response curves did not reach maximum activation (Figure 6.14).

Due to the initial failure to obtain the full range IC_{50} of chitobiose and the non-consistent results with a lower substrate concentration (Figure 6.12), another IC_{50} determination using half the substrate concentration was attempted (Figure 6.15). When a substrate concentration of 177 µg/ml was used, the obtained dose-response curve can be fitted after excluding one data point. The fit is suboptimal,

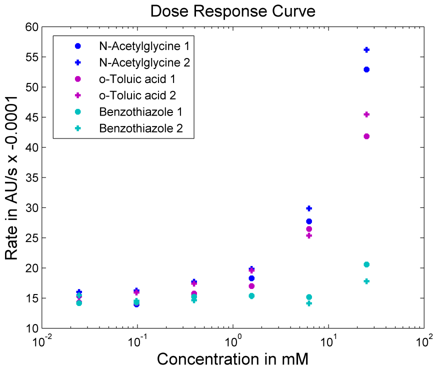


Figure 6.14: **Dose-response curves for activators to find maxima.** All activators were tested up to a concentration of 25 mM to find the maximal activity.

but indicates an IC_{50} in the expected range (K_D in literature is 170 µM).

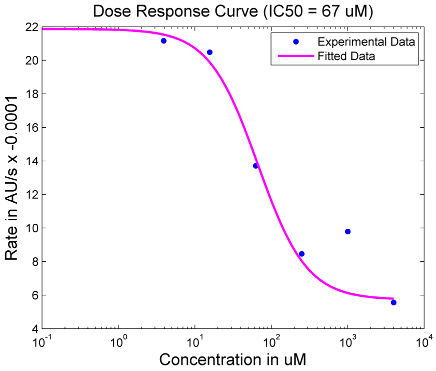


Figure 6.15: **IC_{50} estimation for chitobiose with the enzyme activity assay.** Although the fit is not optimal, the calculated IC_{50} lies within the expected range. One data point was excluded as an outlier.

6.3.2 Discussion of the Enzyme Activity Assay with HEWL

The enzyme activity assay is a cost-effective and easily implemented assay. The 50 mM phosphate buffer pH 7.4 was found to buffer the acidic compounds well. However, the assay faces major reproducibility issues. In this section, 14 fragments were initially tested. Subsequently mainly three different presumably activating compounds, N-acetylglycine, o-toluic acid and benzothiazole, were tested. The dose-response curves for the potential activators do not reach their maxima.

The reasons for this remain uncertain. The difficulties could either derive from the validity of the assay itself or because of the very weak binding constants of these compounds. The major problematic aspect is the unknown working concentration of substrate. Cells can be prepared as a mass-by-volume suspension. However, since the cells are heavier than the solvent they will sink to the bottom of the tubes. Each pipetted aliquot will always have a different number of cells. Further, it can be assumed that cells sink to the bottom of the plate during the actual experiment and might not be available as a substrate for HEWL during the experiment. It must also be considered that not all cell walls have exactly the same composition which likewise may result in an inaccurate determination of substrate concentration. Concluding, one can say that an assay with known substrate concentration would be a big advantage to test potential activators.

6.4 Thermal Shift Analysis

Thermal shift analysis (TSA) is a straightforward assay based on the thermal denaturation of proteins. The central assumption is that binding of a ligand will stabilise a protein, yielding improved thermal stability of the protein. The experiment is performed with a real time PCR machine (qPCR) and environmentally sensitive fluorescent dyes. The fluorescence of the dye is quenched in an aqueous environment. In the presence of hydrophobic residues which will become exposed when the heated protein unfolds, fluorescence increases and a protein melting curve can be recorded. The temperature where the amount of folded and unfolded protein is equal is named the melting temperature ($\mathrm{T_m}$). The program MTSA was written to aid fitting and analysis of the data (details of this program are provided in Chapter 4 of this thesis). The following results were produced by fitting the 5-parameter logistic equation Sigmoid-5 and taking the inflection point of the curves as $\mathrm{T_m}$. The experimental setup is described in Section 10.8.

6.4.1 Buffer Screen and Initial Tests

The first step was finding the buffer conditions which best stabilised the protein for good melting curves. The following buffers were screened at 50 mM concentration including 150 mM NaCl: HEPES pH 7.5, MES pH 6.5, bicine pH 9.0, Na/K PO_4 pH 7.0, MOPS pH 7.0, CAPS pH 10.0, CHES pH 9.5, PIPES pH 6.5.

Initially 100 µg/ml protein in 200 mM CAPS buffer at pH 10.0 with 150 mM NaCl was used to test some ligands such as the possible activators N-acetylglycine, o-toluic acid, benzothiazole and the positive control chitobiose. Ligands were screened at a maximum concentration of 5 mM for the fragments and 0.5 mM for chitobiose. However these conditions often produced double-humped curves (Figure 6.16) which does not allow accurate determination of the T_m.

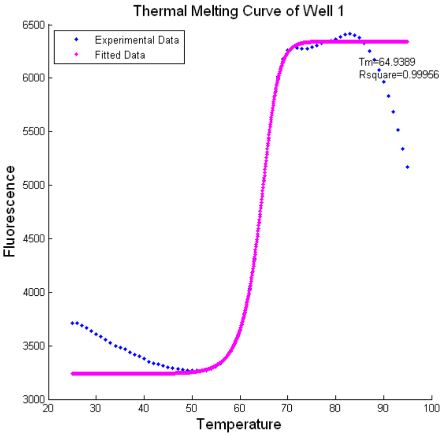


Figure 6.16: **Double humped curves in high pH buffer.** The initially chosen condition 200 mM CAPS pH 10, 150 mM NaCl, 100 µM HEWL (DMSO free) produced double humped curves in experiments.

According to other reports, HEWL forms dimers at pH 5–9 and higher order oligomers at pH 10–11 (Sophianopoulos and van Holde 1961, Sophianopoulos and van Holde 1964). Kumar et al. (2009) states that the self-aggregation at pH 12.2 can be inhibited by the addition of chitotriose. At pH 4.0 the protein is almost in its native state. Only heating at 80℃ for many days can make HEWL form aggregates at that pH (Arnaudov and de Vries 2005). Thus two lower pH buffers - sodium acetate pH 4.5 and citric acid pH 3.8 - were tested at 50 and 200 µg/ml HEWL concentration with and without positive control chitobiose and 2.5% final DMSO concentration.

The melting curves at higher protein concentrations (200 µg/ml) are more regularly shaped and produce better fits. The baseline of the curves at low protein concentration (50 µg/ml) are rather noisy. Thus, only the curves at higher protein concentration are considered in the following sections. A good buffer should feature a high protein melting temperature without ligand present (ergo being a stabilising buffer), a low standard deviation (meaning low variability and high reproducibility) and a large temperature shift for the positive control (chitobiose). The melting temperature for HEWL in sodium acetate buffer was slightly higher (73.53℃) than in citric acid (73.13℃). However, the standard deviation in acetate buffer was higher (0.21℃) than in citric acid (0.11℃). The temperature shift produced by chitobiose is slightly higher in citric acid buffer (0.19℃ for the 200 µM and 0.67℃ for the 2 mM chitobiose concentration) than in sodium acetate buffer (0.18℃ for the 200 µM and 0.65℃ for the 2 mM chitobiose concentration). For this reason, citric acid was chosen as the new screening buffer.

6.4.2 Sensitivity Test of Different qPCR Machines

At the time TSA was established in YSBL different qPCR machines were available (AB 7300, AB 7500, Bio-Rad CFX96 and Agilent Stratagene MX3005). It turned out that the data is not consistent between them. In particular, the Bio-Rad machine seemed to produce different results. To obtain a comparison, a sensitivity check was performed on all available machines to select the best one for purchase. The established assay conditions of HEWL were used to choose between the three available devices from Bio-Rad, AB (AB 7500) and Stratagene. A HEWL concentration of 500 µM in 12 twofold dilution steps was aliquoted three times and tested in octuplets on each of the three machines. The results were analysed for the maximum dilution at which a fit of the melting curve was still possible (sensitivity) and which fit showed the lowest standard deviation (reproducibility). The Bio-Rad device was found to detect a HEWL melting curve to a lower concentration limit of 250 µg/ml which made it the device with the by far lowest sensitivity. The AB and the Stratagene machine were comparable in sensitivity and reproducibility. Curves could be detected relatively confidently down to 62 µg/ml and in some cases even down to 31 µg/ml. The Agilent Stratagene MX3005 machine was purchased for the lab because it offers the most user-friendly interface and was better value. All the following experiments were performed on that model.

6.4.3 Small Fragment Screen with TSA

A small number of fragments were screened against HEWL with the thermal shift assay including the three activators discovered in Section 6.3 on page 119 and a selection of fragments from "Michele_2 ". The following two paragraphs summarise the results of the experiments.

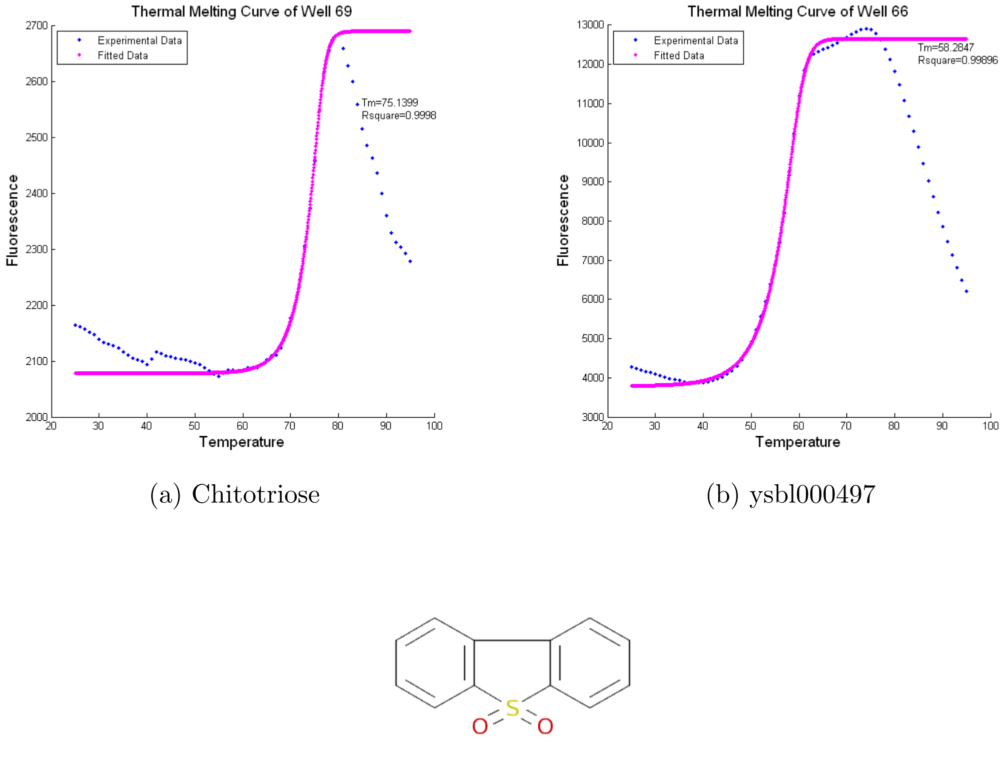
Potential Activators and Chitotriose

For the TSA experiments with HEWL every third well was used as a blank resulting in 36 blanks per 96-well plate. The standard deviation of the HEWL plates was relatively high (0.2–0.5℃). A significant thermal shift was defined as more than three times the standard deviation (according to Sorrell et al., 2010). As expected the negative control glucose did not indicate any temperature shift. Chitotriose showed a significant shift of 1.62℃ for the high concentration (5 mM), but not at the 500 µM concentration (0.12℃). The three potential activators N-acetylglycine, o-toluic acid and benzothiazole did not demonstrate a significant thermal shift.

Other Fragments

The buffer 50 mM citric acid pH 3.8 with 150 mM NaCl was used to estimate fragment screening suitability of the assay. Nineteen compounds from "Michele_2" (Sigma Aldrich) were tested on three plates. The standard deviations of the blanks were again undesirably high: plate 1 was 0.24℃, plate 2 0.26℃ and plate 3 0.51℃. No significant shift was found. However compound ysbl000461 indicated a hint of a thermal shift with an average of 0.2℃ for the low concentration 500 µM (lower than standard deviation) and an average of 0.4℃ for the high concentration 5 mM.

Fragment ysbl000497 alters the shape of the plateau of the melting curve. In the other cases the HEWL melting curves in citric acid buffer produces a very pointed narrow plateau. However for ysbl000497 the curves of the 5 mM concentration have a longer, slightly double-humped plateau (Figure 6.17). Interestingly, according to NMR this compound has a very low concentration and is not 100% soluble at 2 mM in D_2O.



(a) Chitotriose

(b) ysbl000497

(c) ysbl000497

Figure 6.17: **Different shaped melting curve with fragment ysbl000497.** (a) The shape of a melting curve with 5 mM chitotriose is a typical melting curve for HEWL in citric acid buffer at pH 3.8. (b) The shape of the melting curve with the fragment ysbl000497 displays a double-humped region at the maximum asymptote. (c) Structure of ysbl000497.

6.4.4 Discussion of TSA with HEWL

The initial occurrence of double-humped curves in the HEWL TSA experiments can be explained with the literature: HEWL tends to form aggregates and fibrils under many pH and heat conditions. HEWL fibrils occur at pH 2.0 and 3.0 when the protein is heated up to 57°C for several days. However, HEWL at pH 4.0 does not aggregate at all even when observed for 42 days. For the pH 4.0 protein, only incubation at 80°C formed small spherical aggregates. These observations suggest different unfolding pathways for HEWL at pH 3.0 and pH 4.0 (Arnaudov and de Vries, 2005). Also, the addition of organic solvents although followed by a long incubation time will lead to fibrils (Krebs et al., 2000). Although the heating period of the experiments in this chapter only lasted one to two hours, the definite influence of the pH and the possible influence of DMSO must be considered. The unfolding curve of the TSA experiments in the pH 3.8 buffer confirms the observations made by Arnaudov and de Vries (2005) using static light scattering that the protein does not aggregate at pH 4.

TSA as a cost-effective high throughput method has a relatively high rate of false negatives. This fact makes the method more attractive in primary screening rather than as a cross-validation method. In the TSA screen with HEWL the binding of the potential activators N-acetylglycine (ysbl000265), o-toluic acid (ysbl000267) and benzothiazole (ysbl000297) could not be confirmed. Since the standard deviation of the blanks in the plates was relatively high and the expected temperature shift would be relatively low, it is likely that the potential activators were missed out as false negatives. One should also consider the very high melting temperature of HEWL above 70℃. It is also possible that some compounds are not stable at that temperature, and will not result in a thermal shift.

The fragment ysbl000497 seems to influence the unfolding behaviour of HEWL. This is particularly interesting because the compound is not very soluble according to the NMR quality control. However it remains to say that the quality control took place with another compound batch than that used in the screens. After the three activators appeared as hits, more solid of them was purchased in order to have ample material for follow-up experiments.

The plates demonstrated a high standard deviation although using 32 blanks. The high standard deviations most likely result from pipetting errors.

It would be worth additionally analysing the TSA data obtained with HEWL with the midpoint method and with the Boltzmann equation. Chapter 8 on page 151 goes into more detail about how the slightly different $\mathrm{T_m}$ determination methods can produce different results.

It remains to question whether the 14 kDa protein is simply too small for TSA. The dye SYPRO orange binds to the hydrophobic regions of the unfolding protein. A small protein like HEWL will not have a big hydrophobic core. It might be interesting to see if TSA is more suitable for bigger proteins with a higher number of folded units.

6.5 Crystallisation

Some initial crystal trays with HEWL were set up in order to confirm the binding of the potential activators crystallographically. As reported many times they grew easily and quickly. The best conditions came from the CSS 1 & 2 screen (tray 2, Dr. Marek Brzozowski's conditions): 2 M sodium formate, HEPES pH 7.5, 30 mg/ml HEWL and 0.8 M sodium formate, 10% PEG 8K, 10% PEG 1K, HEPES pH 7.5, 30 mg/ml HEWL (Section 10.9.1, page 186) and follow-up trays were set up. First trials to assess appropriate soaking conditions were performed on tray 1 (Section 10.9.1, page 186) at several highly concentrated fragment solutions (10–200 mM). However, crystals cracked at the highest concentration. The lowest tested concentration (10 mM with 2.7% DMSO) did not break the crystals, although crystals looked damaged.

6.6 Final Remarks about HEWL

HEWL turned out to be a challenging project. While the protein can be purchased at low costs and quickly forms crystals, HEWL does not behave straightforwardly in other assays.

A novel Biacore assay was developed which can be used for carbohydrate binding proteins on the reference cell of a recycled CM5 chip: The protein binds with its sugar-binding site to the dextran surface of the chip and added ligands will compete for the binding site of the protein. The set-up allows the protein to be the reporting signal, which results in a high response. This is a big advantage towards immobilised protein where binding ligands give far lower responses. Nevertheless the assay has its drawbacks. Materials may accumulate at the surface and make it difficult to obtain reproducible results.

Similar experiences were made with the subsequent enzyme activity assay. HEWL cleaves the cell walls of *Micrococcus lysodeicticus* and the turbidity of the solution can be followed at 450 nm absorption. Chitobiose and 14 fragments considered to influence HEWL activity were tested with the novel Biacore assay. However, dose-response curves did not give clear results. The major issue with the assay is the unknown substrate concentration. For strong inhibitors this might not be an issue, but it seems to be for weakly binding fragments.

As a third assay thermal shift analysis was applied to HEWL. The conditions were optimised because under certain conditions, HEWL tends to form aggregates and fibrils. Controls, the three potential activators and 19 compounds of "Michele_2" were tested. Only the positive control chitotriose demonstrated a significant thermal shift. None of the fragment test sets showed any binding neither the binding. Binding could also not be confirmed for the three potential fragment activators.

Although issues were encountered when attempting to confirm fragment binding, initial crystallisation trials were started. As expected, HEWL crystals grew easily and rapidly. The first soaking tests led to cracking of the crystals - an effect of too harsh soaking conditions such as high DMSO and ligand concentrations or too rapid soaking. However at the time the initial crystallisation experiments were set up with HEWL, the attention moved to the protein BtGH84 (a glycoside hydrolase from Bacteroides thetaiotaomicron) which turned out to be a much simpler target (Chapter 7 on page 134). Therefore HEWL was put to the side for potential later studies.

Chapter 7

The Target BtGH84

After two challenging protein targets (NMT in Chapter 5 and HEWL in Chapter 6), fragment screening with the glycoside hydrolase BtGH84 turned out to be very successful. Production of stable protein was straightforward and about 500 compounds were tested with a thermal shift assay. The analysis of the large amount of data was made possible with the – in Chapter 4 devised – program MTSA. Interesting hits were assayed with surface plasmon resonance to determine affinity. A number of attempts were made to determine the structure of hit fragments complexed to BtGH84 – however no bound fragment was observed. The project is very promising for follow-up experiments.

7.1 Background of BtGH84

The common post-translational modification of GlcNAc addition (saccharide 2-acetamido-2-deoxy-D-glucopyranose) in cells of higher eukaryotes was first described by Torres and Hart, 1984. The carbohydrate becomes attached to serine and threonine residues via a beta-glycosidic linkage (O-GlcNAc) and is commonly found at phosphorylation sites (Kamemura et al., 2002; Cheng et al., 2000) with a reciprocal relationship between the presence of O-phosphate and O-GlcNAc. Similar to phosphorylation, O-GlcNAc modification is happening several times during the lifetime of a protein (Chou et al., 1992). The O-GlcNAc modification is thought play to a relevant role in Alzheimer's disease (Griffith and Schmitz, 1995), cancer (Chou and Hart, 2001) and diabetes type II. More than 500 kinases are responsible for phosphorylation and more than 140 phosphatases for dephosphorylation (Manning et al., 2002). However for the O-GlcNAc modification there are only two enzymes in mammals: one to add the Glc-NAc (glycosyltransferase OGTase) (Kreppel et al., 1997) and one to remove it (O-GlcNAcase, glycoside hydrolase) (Gao et al., 2001). O-GlcNAcase is very important in cellular signalling and stress response.

The structure contains a C-terminal acyltransferase and a N-terminal glycoside hydrolase domain. The N-terminal domains of eukaryotic O-GlcNAcases are highly similar to some bacterial enzymes. The function of these bacterial enzymes is not very clear although they have a high sequence similarity. The bacterial and the eukaryotic O-GlcNAcases have both been grouped into the glycoside hydrolase family GH84.

A close homologue to the human enzyme is BtGH84 - the glycosid hydrolase from *Bacteroides thetaiotaomicron VPI-5482*. It also cleaves O-GlcNAc from post-translationally modified proteins using a mechanism involving substrate-assisted catalysis. Nearly the entire catalytic centre is conserved (except one amino acid) within the active site architecture of BtGH84 compared to the human O-GlcNAcase (Dennis et al., 2006). BtGH84 consists of four domains whereof the fourth domain can bind carbohydrates (Dennis et al., 2006). PUGNAc (O-(2-acetoamido-2-deoxy-D-glucopyra-nosylidene)amino-N-phenylcarbamate) (Mohan and Vasalla, 2000) (Figure 7.1) inhibits BtGH84, although the literature binding constant varies between nanomolar to micromolar range with the most often mentioned binding constant in low micromolar range (Dennis et al., 2006; He et al., 2011; PhD thesis of Dr. Yuan He, 2011; Dr. Jens Landström, unpublished data).

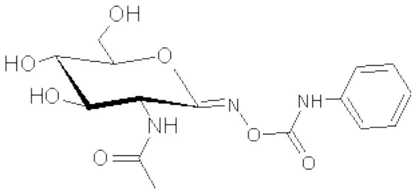
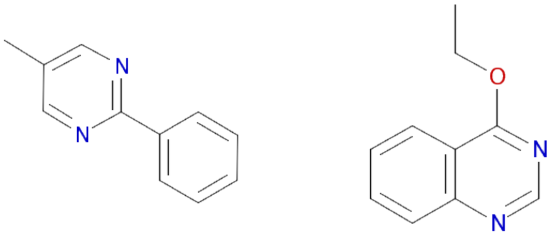


Figure 7.1: **PUGNAc.** The 353 Da compound PUGNAc binds to BtGH84 although values for the inhibition constant vary starkly in the literature.

Dr. Jens Landström screened BtGH84 against the first half of "Michele_1" (about

100 fragments) and found the fragment ysbl000293 to be the best inhibitor with an IC_{50} of 600 µM and ysbl000310 to be the best activator with an EC_{50} of 4 mM (unpublished data).



(a) ysbl000293 (b) ysbl000310

Figure 7.2: **ysbl000293 and ysbl000310.** According to Dr. Jens Landström's fragment screen ysbl000293 inhibits BtGH84 with an IC_{50} of 600 µM whereas it is activated by ysbl000310 with an EC_{50} of 4 mM.

In this chapter, the TSA screening results of almost 500 compounds against BtGH84 are presented, namely fragments from "Michele_1" and "Michele_2" and neighbours of Dr. Jens Landström's hits (library named "Jens"). Interesting hits were cross validated with Biacore experiments. Some compounds were soaked into BtGH94 crystals. However none of the structures showed ligand binding.

7.2 Protein Production

Protein stocks of N-terminal His_6-tagged protein from Dr. Jens Landström and Dr. Yuan He as well as self-produced protein were used in this chapter (Dennis et al. 2006).

The protein with an N-terminal His_6-tag was produced in *E. coli* BL21(DE3). Cells were grown overnight at 37°C and expression was induced with 1 mM IPTG at an OD600 of 0.8. Cells were lysed by sonication and the protein was purified via affinity chromatography using a 5 ml nickel-NTA column. It was eluted in fractions with 0 to 100% buffer containing 500 mM imidazole. Fractions were assessed with SDS-PAGE (Figure 7.3) and the protein containing fractions were pooled together and concentrated. The single step purified BtGH84 to a suitable purity. (A more detailed description can be found in Chapter 10. Protein production was published by Dennis et al., 2006.)

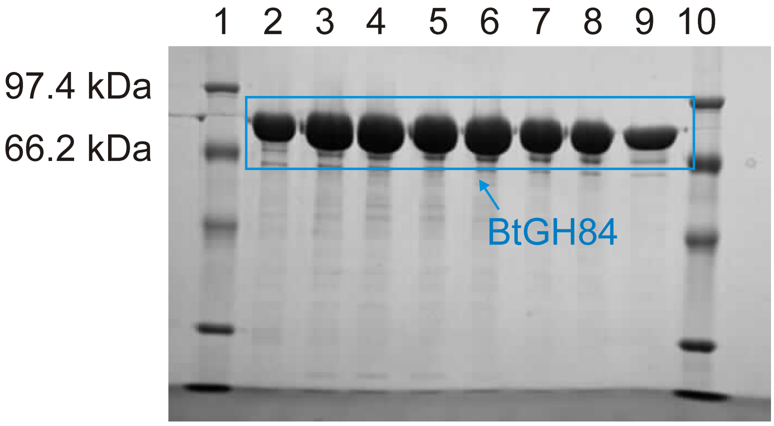


Figure 7.3: **Purity of BtGH84.** SDS-PAGE analysis of the used fractions of BtGH84 (82 kDa) after affinity chromatography. (1) and (10): protein marker; (2)–(9): BtGH84 eluted fractions.

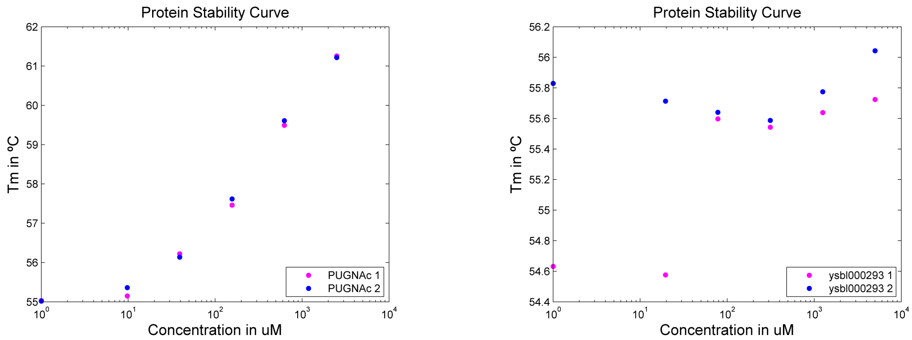
7.3 Thermal Shift Analysis

The main screening method for BtGH84 in this thesis was thermal shift analysis (for more details about TSA, refer to Section 6.4 on page 126). About 500 fragments from the prior generated fragment libraries "Michele_1" and "Michele_2" (Chapter 3) and neighbours (compounds similar to hits previously found by Dr. Jens Landström, thus called library "Jens") were screened using this method. The handling of the resulting large data sets would not have been possible without the analysis program MTSA (Chapter 4, page 85).

7.3.1 Initial Tests, Buffer and Concentration Effects

BtGH84 itself behaved exemplarily for thermal shift analysis. Using the fluorescent dye SYPRO orange melting curves fit nicely and are reasonably reproducible. According to a student's project from Bailey Massa who tested several buffers, Na/K PO_4 buffer pH 7.7 was found to be the most suitable to screen BtGH84 with TSA. According to Dennis et al., 2006, the optimum pH for BtGH84 is pH 6.0 (Dennis et al. 2006) and for the human enzyme 6.5 (Cetinbas et al. 2006). Thus the buffer was changed to sodium phosphate pH 6.0 after initial tests. (More details about materials and methods in Section 10.8, page 184.)

The thermal shift experiments were performed in 50 mM sodium phosphate buffer at pH 6.0 with 150 mM NaCl and 1x SYPRO orange. The fragment hit ysbl000293 with an IC_{50} of 600 µM (Dr. Jens Landström, unpublished work) and PUGNAc were used for tests and as positive controls. Two concentration series were measured for each compound. PUGNAc stabilises BtGH84 significantly whereas the thermal shift for the fragment is less obvious but unambiguous: ysbl000293 shifts the T_m about 0.8°C at 5 mM concentration, PUGNAc exhibits a shift of 6°C at 2.5 mM concentration. The T_m for PUGNAc increases almost linearly with each concentration step while the correlation for ysbl000293 is ambiguous (Figure 7.4).



(a) PUGNAc

(b) ysbl000293

Figure 7.4: **Thermal shift of the positive controls.** Both positive controls PUGNAc and ysbl000293 stabilise BtGH84 significantly in the thermal shift experiments. The zero concentration was adjusted to 1 µM to enable a logarithmic plot.

The effects of protein concentration and of DMSO in the assay were assessed to estimate their influence on subsequent TSA screening experiments. Different protein concentrations at 0% DMSO were tested which revealed that the T_m increases with protein concentration. Table 7.1 illustrates the different protein concentrations and their shifts (each measurement was recorded as quadruplicate). Unsurprisingly, higher protein concentrations feature lower standard deviations. The standard deviation becomes comparable for protein concentrations from 0.5 µM. The higher the concentration, the better the quality of the fits.

Unexpectedly the melting temperature increases with protein concentration which may suggest cooperative stabilisation effects of the protein. Screening a range of DMSO concentrations at 1.4 µM BtGH84 reveals a stabilising effect of DMSO on BtGH84. The melting temperature increases with DMSO concentration up to

Table 7.1: **Working concentration BtGH84** Estimating the best working TSA concentration for BtGH84. (n=4)

For 0% DMSO		
BtGH84 (µM)	$\overline{T_m}$ **(℃)**	σ **(℃)**
0.27	50.07	0.37
0.54	52.21	0.16
0.95	53.00	0.06
1.35	54.38	0.14

6% DMSO where the T_m reaches its maximum (Figure 7.5). These findings led to the final screening conditions of 50 mM sodium phosphate buffer pH 6.0 with 150 mM NaCl, 1.34 µM BtGH84, 5% DMSO and 1x SYPRO orange.

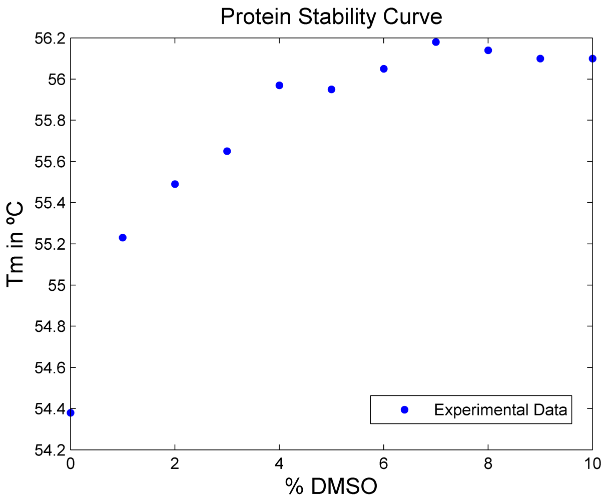


Figure 7.5: **Influence of DMSO on BtGH84 stability.** BtGH84 becomes more stable with DMSO in the buffer. Data for 1.4 µM protein.

Interestingly the stabilisation effects of higher DMSO concentrations were not confirmed by a later experiment using the fluorescent dye Deep Purple (Section 7.3.3 on 141). Here the T_m for wells with 6.5% DMSO was 57.22℃, but one degree higher when only 1.5% DMSO was present in the buffer. Yet that experiment was performed with a different batch of protein.

7.3.2 TSA Screen of Michele_1 and Michele_2

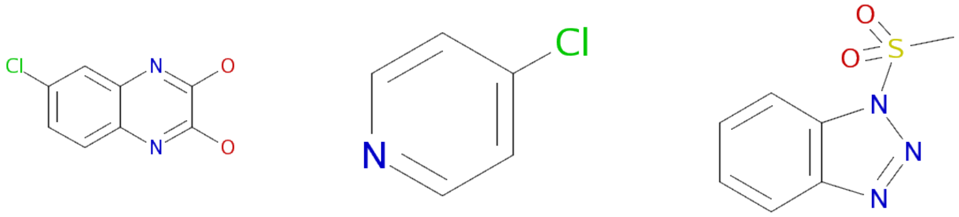
In total about 300 fragments were screened against BtGH84 with TSA. All mentioned structures in this chapter are pictured in Appendix D. The compounds were screened at low and high concentration (1 and 10 mM) as quadruplicates in 50 mM sodium phosphate buffer pH 6.0/150 mM NaCl with 1.4 µM BtGH84 protein, 1x SYPRO orange and a final DMSO concentration of 5%. Each plate was composed of eight blanks with protein in buffer containing 5% DMSO and of eleven fragments. The data was analysed with MTSA (Chapter 4, page 85) using the inflection point of the Sigmoid-5 equation (Section 4.3). The results of the screen are summarised in Table 7.2. Some of the fragment hits stabilised the protein at low concentration, but destabilised it at high concentration. This effect was considered to be caused of high concentrated ligands interfering with the assay rather than deriving from false positives. Some structures are shown in Figure 7.6.

Table 7.2: **Fragment hits for BtGH84 when using the inflection point** (n=4; 8 blank controls per plate)

	Soluble in assay	**Blanks**		$\overline{\Delta T_m}$ **at**	
Fragment	**(yes/no)**	$\overline{T_m}$ **(°C)**	σ **(°C)**	**1 mM (°C)**	**10 mM (°C)**
ysbl000298	y	56.69	0.10	0.71	-2.98
ysbl000299	y	56.69	0.10	0.51	-6.63
ysbl000317	n	56.91	0.11	0.68	-0.52
ysbl000416	y	56.66	0.11	0.29	-0.10
ysbl000438	y	56.52	0.22	0.86	1.41
ysbl000486	y	56.34	0.08	0.22	0.14
ysbl000507	y	55.94	0.78	3.11	-11.88
ysbl000509	y	56.35	0.27	2.09	1.68

7.3.3 TSA Screen of Neighbours Jens

A further 150 compounds from the library “Jens” were assayed. This library contains nearest neighbours to hits of an initial fragment screen (Dr. Jens Landström). The compounds were screened and analysed the same way as the frag-



(a) ysbl000438 (b) ysbl000507 (c) ysbl000509

Figure 7.6: **Fragment hits of the TSA screen.**

ments above with the difference of 32 controls and eight compounds per plate. The hits and their thermal shifts derived with the inflection point method are summarised in Table 7.3. The table contains further information about the solubility of the compounds in the screen. All data derived from compounds with poor solubility should be handled with care because precipitation may interfere with the assay.

Compounds ysbl000673, ysbl000683, ysbl000708, ysbl000730, ysbl000733, ysbl000749, ysbl000752 and ysbl000779 were followed up in control experiments to confirm their binding.

Controls

The neighbours ysbl000673, ysbl000683, ysbl000730, ysbl000733, ysbl000749 and ysbl000752 were tested in three further controls:

1. without dye
2. without protein
3. with a different dye

1. None of the compounds had intrinsic fluorescence, i.e. when there was no dye present, no response could be recorded.

2. The tested neighbour ysbl000749 could be ruled out for further tests as it showed a response while no protein was present. It featured two differently shaped curves depending on the concentration (1 mM and 10 mM). The compound also

Table 7.3: **Hits from library "Jens" for BtGH84 when using the inflection point method** Solubility was assayed for two steps. The first step is a pre dilution of the compound on a separate ligand plate, the second step is the dilution to the final assay conditions. (n=4; 32 blank controls per plate)

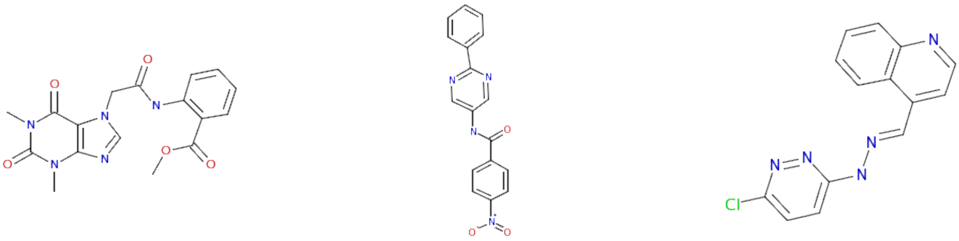
	Soluble (yes/no)		**Blanks**		$\overline{\Delta T_m}$ **at**	
Fragment	**1st step**	**2nd step**	$\overline{T_m}$ **(°C)**	σ **(°C)**	**1 mM (°C)**	**10 mM (°C)**
ysbl000673	y	y	56.16	0.24	0.79	1.00
ysbl000683	y	y	56.49	0.15	0.36	1.08
ysbl000684	n	n	56.49	0.15	0.12	0.89
ysbl000705	n	y	56.67	0.20	0.32	0.73
ysbl000720	y	y	55.13	1.05	0.41	4.87
ysbl000730	n	n	55.87	0.09	1.36	2.69
ysbl000733	y	n	55.87	0.09	0.52	1.36
ysbl000734	y	n	55.87	0.09	1.58	2.70
ysbl000737	n	n	55.87	0.09	0.08	0.63
ysbl000749	n	n	54.92	1.68	9.82	11.24
ysbl000752	y	y	54.92	1.68	0.98	1.50
ysbl000767	n	n	55.26	0.47	0.83	2.49
ysbl000770	n	y	56.00	0.13	0.18	0.92
ysbl000775	n	n	56.00	0.13	0.21	0.52

changed colour in wells of different concentrations (yellow at low concentration and red at high concentration on the pre diluted ligand plate) which may indicate tautomerisation.

3. Deep purple was used as an alternative dye to SYPRO orange. It was found to work best at 6x concentration which results in a final concentration of 6.5% DMSO and 1.5% acetonitrile. The SYPRO orange filter and the Nile red filter were both used to record the melting curves, but the Nile Red filter seemed to produce better curves.

A thermal shift introduced by all compounds could be confirmed. Although the absolute number differs, the quality of the shifts was comparable to the results obtained with SYPRO orange. The only exception was ysbl000733 which caused a negative shift at high concentration.

The fragments ysbl000673 ($\overline{\Delta T}_{m(1mM)} = 0.11\,°\mathrm{C}$, $\overline{\Delta T}_{m(10mM)} = 0.82\,°\mathrm{C}$), ysbl000 730 ($\overline{\Delta T}_{m(1mM)} = 0.18\,°\mathrm{C}$, $\overline{\Delta T}_{m(10mM)} = 1.69\,°\mathrm{C}$) and ysbl000752 ($\overline{\Delta T}_{m(1mM)} = 0.23\,°\mathrm{C}$, $\overline{\Delta T}_{m(10mM)} = 2.07\,°\mathrm{C}$) progressed to the next evaluation step (Biacore) (Figure 7.7). In accordance with Dr. Jens Landström these compounds were considered as the most interesting and the most promising binders.



(a) ysbl000673 (b) ysbl000730 (c) ysbl000752

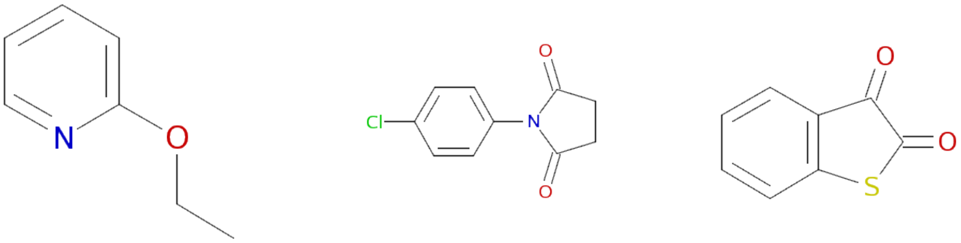
Figure 7.7: **Neighbour compounds progressed to Biacore.**

7.3.4 Discussion and Remarks

BtGH84 behaves exemplary with TSA. Almost 500 compounds were tested in a straightforward assay. Surprisingly BtGH84 is thermally more stable when DMSO is present in the buffer. Although this finding is dispelled by DMSO destabilising the protein when using the alternative fluorescent dye Deep Purple. It may be noted that the experiments were performed with different batches of protein and it would be interesting to follow up that contradiction with further experiments. Deep purple was found to work best with the implemented Nile red filter and at a 6x concentration which results in a final concentration of 6.5% DMSO and 1.5% acetonitrile. The neighbour compound ysbl000749 looks like it can tautomerise or react easily and was thus excluded from further assays.

7.4 Biacore with Fragments and Neighbours

In order to cross-validate the hits and to obtain K_D of the fragments and neighbours, they were tested with Biacore (some structures of assessed compounds in Figure 7.8). After catching all protein in the same orientation via its His_6-tag, BtGH84 was covalently attached to an NTA chip in order to avoid it being washed off the chip during the experiment. The protein remained stable in the time between the two experiments when the chip was stored in buffer at 4°C. Biacore experiments were performed with duplicate concentration series in quadruplicating steps. The initial experiment with the maximum concentration of 10 mM exhibited solubility issues, thus the experiment was repeated with a maximum concentration of 1 mM. The K_Ds were determined with the Biacore software using affinity steady state analysis. (Methods in Section 10.5.2 on page 181.)



(a) Ysbl000299 (b) Ysbl000423 (c) Ysbl000548

Figure 7.8: **Compounds assessed with Biacore.**

The concentration of the relevant stocks of compounds in the "Jens" library was unknown since they were saturated solutions only. Thus the calculation was performed assuming the highest concentration is 200 mM like the fragments. That means the actual derived K_Ds must be significantly lower since the actual concentrations are far below 200 mM. Table 7.4 gives an overview about the K_Ds. However the table also illustrates that points needed frequently to be excluded from the analysis to obtain K_D values. In most of the cases the ligand precipitated at the high concentration, therefore the experiment was repeated with a maximum concentration of 1 mM. Table 7.5 suggests that although the fragments were mostly soluble within the lower maximum concentration series (1 mM), the response in RU was often too low to fit the data. To analyse the data many points were excluded as they were outliers. Nevertheless the positive control PUGNAc (Figure 7.1) proves that the measurement of a binding event with that assay is possible.

The software has two components to analyse the data. Either steady state affinity which uses the height of the curves or kinetics analysis which determines the slope for the kinetic rate constants k_a and k_d. Fragment binding data were usually

Table 7.4: **Hits for BtGH84 cross validated with Biacore at 10 mM maximal concentration** All binding data was obtained by binding affinity module using steady state analysis. Several high concentration data points needed to be excluded in order to allow the software fit the curves. They were considered as outliers. (n=2)

Fragment	Soluble on plate	K_D (mM)	χ^2	Excluded points (mM)
ysbl000298	n	73.0	0.087	2x 10
ysbl000299	n	39.5	0.026	2x 10
ysbl000317	n	34.4	0.827	-
ysbl000438	y	n/a	n/a	n/a
ysbl000486	y	22.2	0.143	2x 10
ysbl000509	y	2.1	0.566	2x 2.5, 2x 10
ysbl000673	y	0.170	0.086	2x 10
or	y	0.489	0.129	2x 2.5, 2x 10
ysbl000730	n	2.9	0.416	2x 2.5, 2x 10
ysbl000752	y	n/a	n/a	n/a

not good enough for a kinetics analysis. However, kinetics analysis could be performed for PUGNAc and delivered a k_a of 0.08 (µMs)$^{-1}$, a k_d of 0.57 s^{-1} and a K_D of 7.3 µM (Figure 7.9 shows the sensorgram of the experiment). When the K_D of PUGNAc is determined with steady state affinity analysis a value of 5.3 µM is extracted. Table 7.5 gives an overview of the K_Ds of PUGNAc, fragments and neighbours tested with Biacore.

7.4.1 Discussion

The determination of the K_D of PUGNAc delivered 5–7 µM which is in the same range as values from He et al. (2011) who found it to be 2.5 µM. The finding enforces the belief that the binding affinity is indeed in micromolar range and not in low nanomolar as stated by Dennis et al. (2006). For the remaining compounds, definition of an affinity constant remained problematic. Compounds were not soluble up to a concentration of 10 mM, however when using a maximum concentration of 1 mM data were not reliable enough and K_D seem to be above that concentration. The response produced with the Biacore experiments was

Table 7.5: **Hits for BtGH84 cross validated with Biacore at 1 mM maximal concentration** PUGNAc was analysed with both the steady state (SS) and the kinetics (kin) module. All other compounds were analysed with steady state binding analysis. Several high concentration data points needed to be excluded in order to allow the software fit the curves. They were considered as outliers. (n=2)

Fragment	**Soluble on plate**	**RU range**	**K_D (mM)**	χ^2	**Excluded points (mM)**
PUGNAc (SS)	y	14	0.005	1.49	-
PUGNAc (kin)	y	14	0.007	0.458	-
ysbl000293	y	<1	0.622	0.103	1x series, 1x 1
ysbl000298	y	12	7.4	0.054	-
ysbl000299	y	13	6.1	0.015	0.004
ysbl000317	y	20	8.8	0.996	-
ysbl000438	n	2.5	0.110	0.064	0.25, 1
ysbl000486	y	<1	0.660	0.108	-
ysbl000509	y	2.2	6.5	0.019	1
ysbl000673	y	5	3.3	0.0244	1
ysbl000730	y	-2	n/a	n/a	
ysbl000752	y	<1	0.844	0.0142	-

often not high enough (third column in Table 7.5). K_D greater than half the highest ligand concentration mean that the data has to be handled with care. It could mean that the data plots do not have sufficient curvature to generate a reliable fit. In many cases the K_D are even above the maximum concentration. Furthermore, many data points had to excluded to make the fit possible for the software to generate (Tables 7.4 and 7.5).

The binding constant for PUGNAc is interesting as it enforces the binding constant to be in the micromolar range and speaks against a nanomolar binding constant.

In general, binding constants derived with Biacore experiments should be lower (stronger binding) than with other methods because the protein is attached to the chip instead of being free in solution. The reduced mobility makes it more likely for the ligand to find the protein.

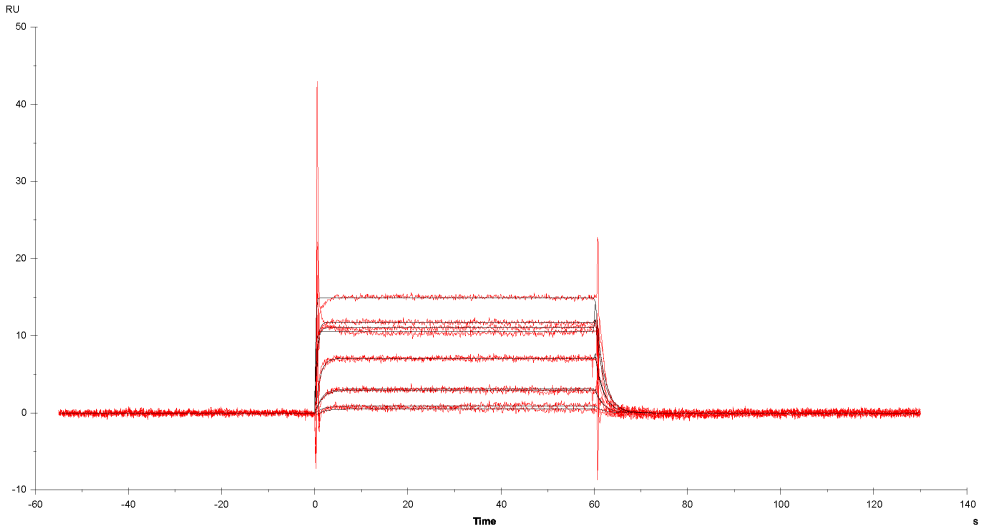


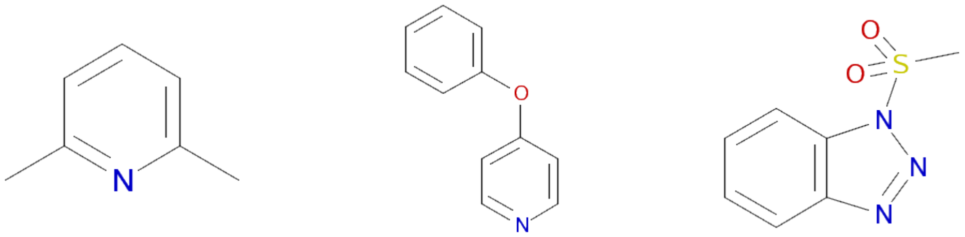
Figure 7.9: **PUGNAc sensorgram.** The known inhibitor of BtGH84 binds with an inhibition constant of K_D 7.3 µM to the enzyme according to a Biacore experiment. The experiment was performed with concentration series ranging from 0.4 µM to 100 µM in 4-fold concentration steps. (n=2)

Affinity constants are often reported with biochemical assays such as enzyme inhibition assays where a competitive ligand competes for binding to the enzyme. One major drawback is the time needed to develop such an assay. They can report IC_{50} values which can be converted into the K_i values which are usually identical to K_D (more in Section 6.3.1). However, it is difficult to detect weak binding compounds (high micromolar) with biochemical assays.

Other direct and more sensitive methods to obtain K_D values are biophysical, and include ITC (Isothermal calorimetric) and NMR (Nuclear Magnetic Resonance). ITC measures quantitatively the thermodynamic parameters of a protein-ligand interaction. Labelling of the protein is not needed. Drawbacks of ITC are the use of high quantities of protein and that compounds need to be very soluble which mostly becomes a problem for testing mM binders. Similarly, titration by NMR would be limited by compound solubility. NMR becomes difficult for proteins above 30 kDa when measuring labelled protein and HSQC (BtGH84 weighs more than 80 kDa). Other NMR methods are more difficult to perform.

7.5 Crystallisation

BtGH84 crystals grow over night in the known conditions 100 mM imidazole pH 8.0, 10% PEG 8K, 3% 2M TMAO and 15% ethylene glycol serving as cryoprotectant (Figure 7.11). Ligands were tested for soaking with final concentrations of about 20 mM and 10% DMSO. Data sets were collected at Diamond for ysbl000298, ysbl000299, ysbl000317 and ysbl000509 (Figures 7.10 and 7.13 a). After processing data with Mosflm and Xia2, structures were solved by molecular replacement with Balbes. Refinement and rebuilding was performed with Refmac5 and Coot (Methods in Section 10.9.2, page 187). The structure of BtGH84 is shown in Figure 7.12.



(a) ysbl000298 (b) ysbl000317 (c) ysbl000509

Figure 7.10: **Compounds soaked into BtGH84 crystals.**

Crystals grew in space group P21212 with two molecules in the asymmetric unit and diffracted between 2.0 and 2.7 Å. None of the structures contained the added ligand which leads to the conclusion that soaking conditions must be optimised. The need to optimise soaking was also suggested by other colleagues (Dr. Jens Landström). None of the structures were totally refined because they did not show density for the ligands.

Some of the structures had density of a metal ion in the A-chain between the carbonyl oxygen of Glu33 and the carboxyl groups of Glu62 and Asp65 side chains. Using data with the usual geometry obtained from http://tanna.bch.ed.ac.uk/qg3.htm, the metal ion is most likely to be a Nickel ion (an example in Figure 7.13).

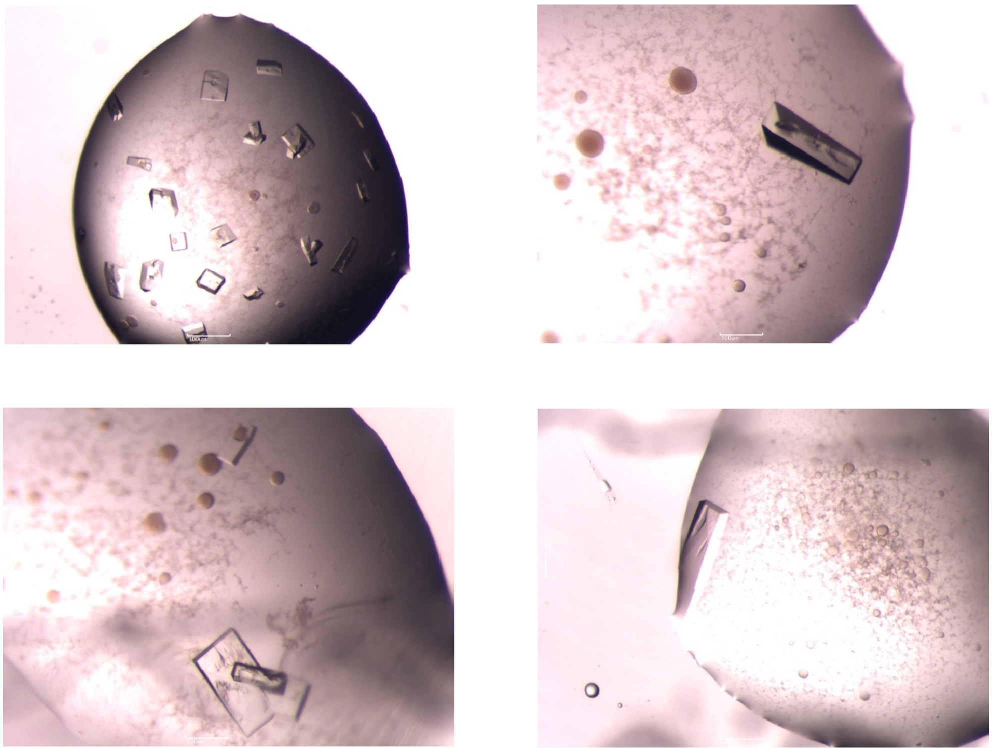


Figure 7.11: **BtGH84 Crystals.**

7.6 Final Remarks about BtGH84

BtGH84 is a very interesting target and easy to handle protein. Almost 500 compounds were screened against the target. Interesting hits were further tested for their K_D and crystal structures attempted to be obtained.

This chapter confirmed the binding constant of PUGNAc to be in the low micromolar range with a K_D around 6 µM, similar to data found by Dr. Yuan He and Dr. Jens Landström. Since Dennis et al. (2006) found PUGNAc binding to have nanomolar affinity, the inconsistency of PUGNAc affinity could be explained with reasons such as PUGNAc falling apart or PUGNAc being synthesised by different people.

There are still many open questions. First there are hits from the TSA screen which got missed with the initial erroneous T_m determination. These could be further tested with the Biacore. Soaking conditions need to be improved, such as using a slower soak, i.e. totally exchanging the mother drop with ligand solution in buffer.

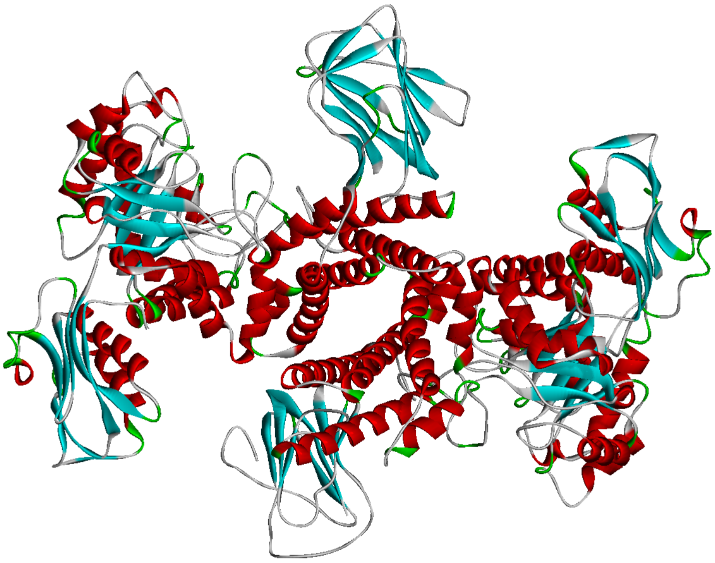


Figure 7.12: **Ribbon diagram of BtGH84.** Ribbon diagram of BtGH84 how it was solved when the ligand ysbl000509 was soaked. (Data was not totally refined after it was clear that no ligand was present in the crystal structure.) The image was created using Discovery Studio.

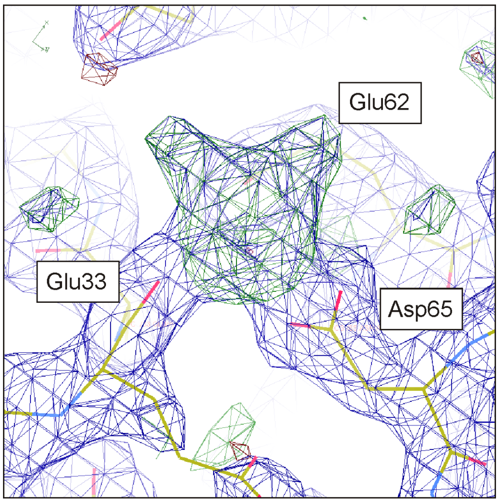


Figure 7.13: **Metal ion in BtGH84 structure.** Some crystal structures of BtGH84 contained a bound metal ion. Here the structure of BtGH84 with a test soak of ligand ysbl000314. Unexplained density was found in the structure (the $2F_o$-F_c map is contoured at 1.5 σ and the F_o-F_c map at 3 σ respectively in Coot). The density indicates most likely a bound Nickel ion.

Chapter 8

Melting Point Definition

The thermal shift data reported in chapters 6 and 7 were analysed with the program MTSA described in Chapter 4. This program analyses the change in fluorescence from a reporter dye as the protein (and ligand) sample is heated to identify a melting temperature $\mathrm{T_m}$. An increase in $\mathrm{T_m}$ suggests that a ligand has an effect on the ligand stability, and hence is considered as a hit in screening. During the analysis of the results presented in this thesis and the preparation of a manuscript describing the program, there was considerable debate about the way in which $\mathrm{T_m}$ is determined. Different conventions for fitting the experimental data and determining $\mathrm{T_m}$ gives small variations in the values obtained which can be significant for identifying whether a fragment is a hit. This chapter goes into more detail about the different procedures and whether the TSA method is suitable for fragment screening.

8.1 Boltzmann, Midpoint and Inflection

Different methods to analyse thermal shift data are available in the literature (Section 4.2, page 86). At the start of the work described in this thesis, decisions were made about which method would be most useful for the project. The more complicated thermodynamic model developed by the inventors of ThermoFluor® which required knowledge of heat capacity (not obtainable by TSA experiments), was discarded (Section 4.2.1, page 87). Instead the Boltzmann equation was initially used to fit the data (more in Section 4.2.2, page 91; Ericsson et al., 2006; Niesen et al., 2007; Sorrell et al. 2010) (Equation 8.1). The equation is a four-parameter logistic model:

$$\gamma(T)_{Boltzmann} = min + \frac{max - min}{1 + e^{\frac{T' - T}{a}}} \tag{8.1}$$

where the parameter T' is equal to the midpoint of the curve and to the point of inflection and thus defined as the melting temperature T_m:

$$T_{m,Boltzmann} = T' \tag{8.2}$$

The initial usage of the Boltzmann fit gave no satisfactory fits (Figure 4.3, page 96). That led to the introduction of an additional parameter c in order to better refine the fitting of the asymmetry of the curves. (Refer also to Chapter 4, page 85). The five-parameter logistic model is referred to as the Sigmoid-5 equation:

$$\gamma(T)_{Sigmoid-5} = min + \frac{max - min}{(1 + e^{\frac{T' - T}{a}})^c} \tag{8.3}$$

For $c = 1$, i.e. when the curve is fully symmetric, Sigmoid-5 becomes the Boltzmann equation. When $c \neq 1$ then the T_m can be defined as either the midpoint:

$$T_{m,Midpoint} = T' - a \times ln(2^{\frac{1}{c}} - 1) \tag{8.4}$$

or as the point of inflection:

$$T_{m,Inflection} = T' - a \times ln(\frac{1}{c}) \tag{8.5}$$

The equations described above lead to three different ways of possible $\mathrm{T_m}$ determination with the software MTSA. These three methods will be referred to as "Boltzmann", "Midpoint" and "Inflection" in the following discussion. All of the equations describe a sigmoidal shaped curve with *max* and *min* as the asymptotes and *a* being the slope. Depending on which fitting equation is used (i.e. Boltzmann where $c = 1$ or Sigmoid-5, with $c \neq 1$) the midpoint and point of inflection are the same points or are separated. The parameter are further illustrated in Figure 8.1.

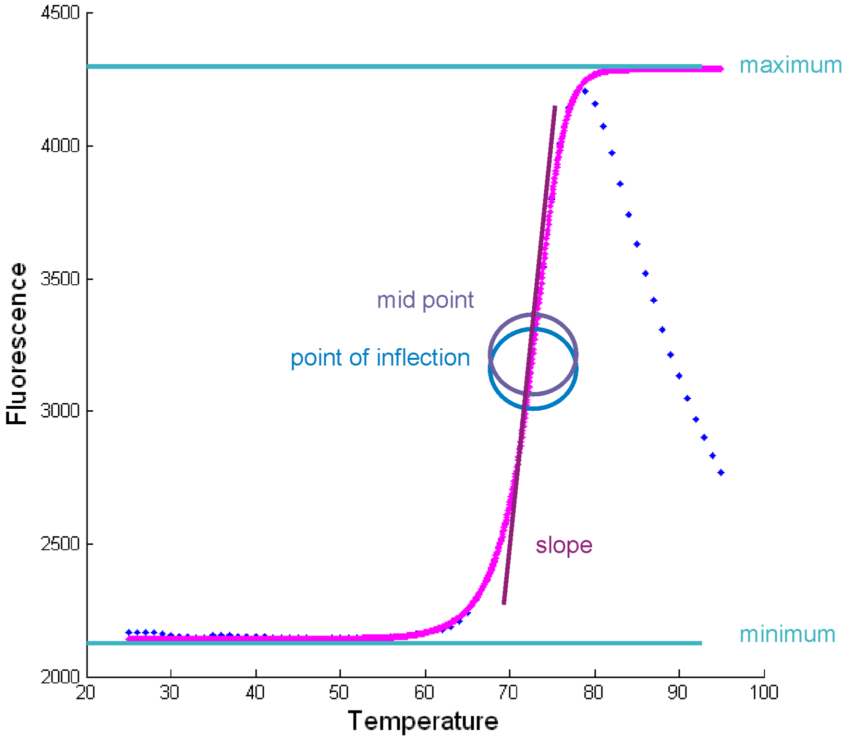


Figure 8.1: **Parameters of the fitting equations.** The parameter *min* and *max* are the same for both fitting equations and corresponds to the minimum and the maximum of the curve. a is the slope of the curve. For the Boltzmann equation, the midpoint is equal to the point of inflection, thus $T_{m,Boltzmann} = T'$. In the Sigmoid-5 equation these two points will be different if $c \neq 1$. $\mathrm{T_m}$ could be defined as the midpoint $T_{m,midpoint} = T' - a \times ln(2^{\frac{1}{c}} - 1)$ or as the point of inflection $T_{m,inflection} = T' - a \times ln(\frac{1}{c})$.

Initially, experiments were analysed with the Inflection method. However, during analysis questions arose as whether this was the most suitable descriptor for the melting point. By definition, the $\mathrm{T_m}$ is the transition midpoint of the curve (i.e., equal concentrations of folded and denatured protein), so the fitting equation should give the $\mathrm{T_m}$ equal to the midpoint.

To answer this question, the midpoint was manually determined from the raw data set: The raw data was transformed into a table of temperature and corresponding fluorescence. In a spreadsheet, the midpoint of the maximum and the minimum fluorescence was determined. As the temperature increases in 1 ℃/min steps, the corresponding temperature can only be estimated. Knowing that the transition around the T_m is almost linear, fluorescence values were estimated. Comparing the manually determined midpoints to “Boltzmann”, “Midpoint” and “Inflection” reveals that “Inflection” over-estimates the T_m value the most. In almost all cases, the manually found T_m and the mathematically determined T_m do not conform. An example is shown in Table 8.1 for the compound ysbl000701 (Figure 8.2; library “Jens” screened against BtGH84, Section 7.3.3, page 140). This analysis provided the motivation to explore the differences between the methods in more detail.

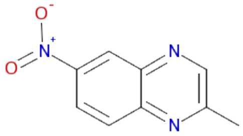


Figure 8.2: **Structure of ysbl000701.**

Table 8.1: **Manually estimated versus mathematically determined T_m.** Values for compound ysbl000701 are shown which are obtained from the “Jens” screen against BtGH84.

	Blank	**1 mM**	**10 mM**	**Blank**	**1 mM**	**10 mM**
Est. Midpoint	55.9	56.1	56.6	56.0	56.1	56.6
Boltzmann	55.9	56.1	56.8	56.0	56.2	56.8
Midpoint	56.0	56.2	56.8	56.1	56.3	56.8
Inflection	56.6	56.7	56.9	56.7	56.8	57.0
Δ to Boltzmann	-0.2	0.0	0.7	-0.1	0.0	0.7
Δ to Midpoint	-0.1	0.1	0.7	0.0	0.1	0.7
Δ to Inflection	0.5	0.6	0.8	0.6	0.7	0.9

The following sections discuss the three methods more into detail and gather information about reproducibility, comparison and hit identification.

8.2 Reproducibility

To begin the in-depth analysis, it was first tested if there were differences in reproducibility for the three T_m determination methods Boltzmann, Midpoint and Inflection. A sensitivity test with HEWL was performed with a protein concentration repeatedly diluted by half from 500 µg/ml down to 0.24 µg/ml (Section 6.4.2, page 128). Each concentration was measured in octuplicates. In terms of reproducibility defined as the standard deviation of aliquoted repeats of the same solution, the Boltzmann, Midpoint and Inflection methods are comparable. The standard deviations for each octuplicate (compared in Table 8.2) show that all three methods give similar reproducibility.

Table 8.2: **Reproducibility of Tm generated by all three methods** The standard deviations of the three different T_m techniques do not differ significantly. At the concentration 62.5 µg/ml one value was excluded as an outlier. At 31.25 µg/ml two curves could not be fitted with any of the fitting equations. (Data obtained with HEWL, n=8.)

	Lysozyme in µg/ml	**500**	**250**	**125**	**62.5**	**31.25**
Boltzmann	average T_m (°C)	72.41	72.66	72.98	72.99	73.70
	stdev (°C)	0.12	0.09	0.12	0.38	0.62
Midpoint	average T_m (°C)	72.54	72.67	72.81	72.68	72.49
	stdev (°C)	0.12	0.05	0.18	0.53	1.08
Inflection	average T_m (°C)	73.31	73.33	73.51	73.37	74.50
	stdev (°C)	0.10	0.14	0.35	0.37	1.43

8.3 Boltzmann Equation versus Sigmoid-5 Equation

In a second step, the comparability of the newer Midpoint and Inflection methods with the conventional Boltzmann method were assessed. The following figures derive from a representative 96-well plate with BtGH84 which contained 32 blanks and 8 compounds screened as quadruplicates. Experimental details can be found in Section 7.3 on page 137. In Figure 8.3, the T_m generated with the new

Midpoint and Inflection methods are plotted against the conventional Boltzmann T_m. One can see that the Midpoint method estimates the T_m slightly higher than Boltzmann, but in a very comparable manner. However the Inflection method produces more variation and obvious differences towards the Boltzmann method. The T_m determined with that method is significantly higher.

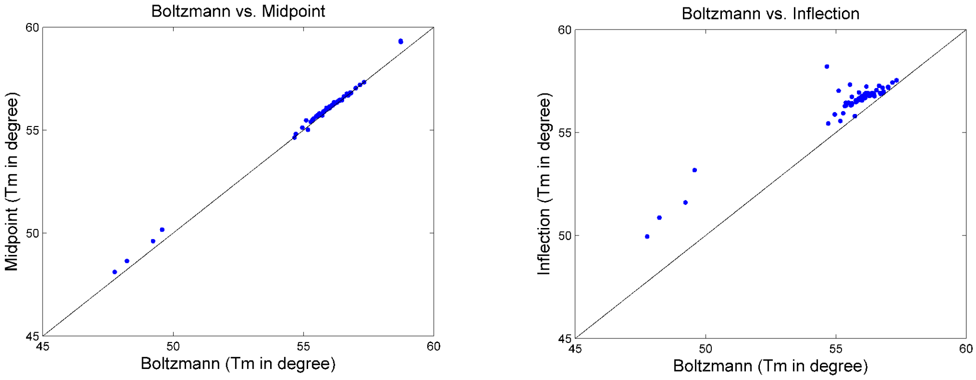


Figure 8.3: **Boltzmann compared to Midpoint and Inflection.** The figure shows a comparing plot of a representative screening plate with BtGH84. On this example plate, the Midpoint method differs less from the conventional Boltzmann method than the Inflection method. Nevertheless the Sigmoid-5 equation produces slightly higher T_m with the Midpoint method, and significantly higher and more spread T_m with the Inflection method.

8.4 Comparison of the Produced Shifts

After comparing the reproducibility and the value of the melting temperatures, the individual shifts produced with Boltzmann, Midpoint and Inflection were compared on an example of a representative screening plate (library "Jens" screened against BtGH84, Section 7.3.3, page 140). The bar chart on Figure 8.4 reveals that depending on how the T_m was determined, different thermal shifts are produced. The shift of all 96 wells was calculated with the Boltzmann (blue), the Midpoint (purple) and the Inflection (pink) method and compared as bars.

The first striking difference occurs for well number six. Boltzmann and Midpoint generate a negative shift whereas Inflection produces a positives shift. In such a case it is required to look in detail at the melting curves. Figure 8.5 displays the

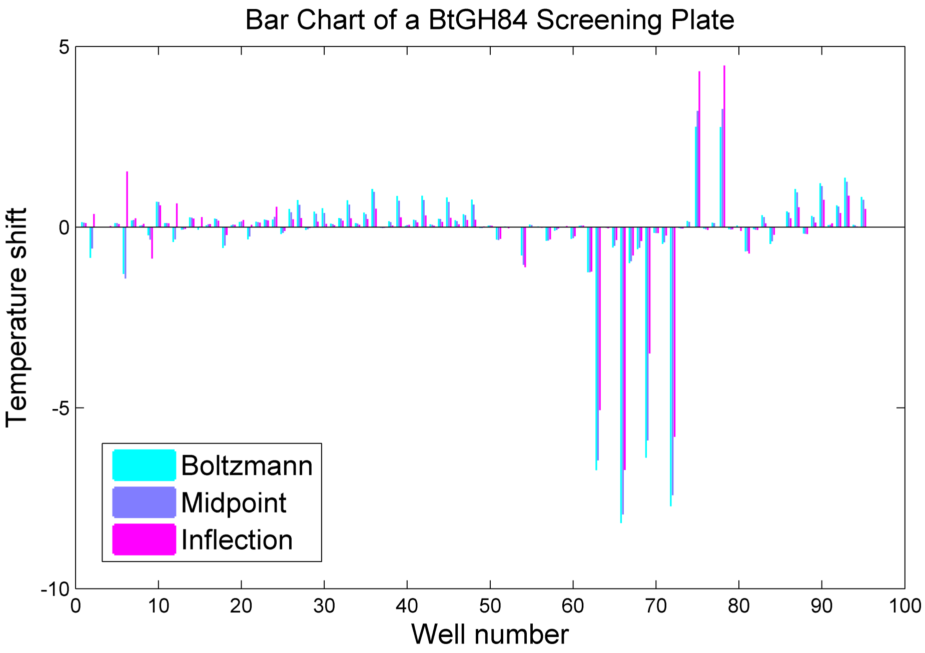
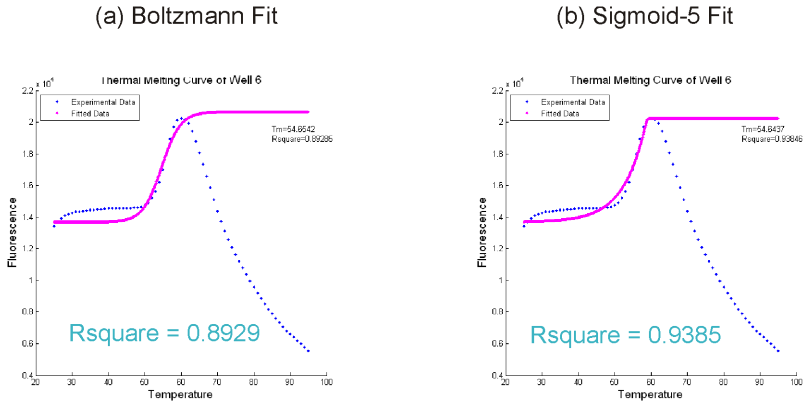


Figure 8.4: **Shifts on a 96-well plate with all three methods.** A representative BtGH84 screening plate with 32 blanks and 8 ligands in quadruplets. The shifts determined by the three different methods are shown in cyan (Boltzmann), purple (Midpoint) and pink (Inflection).

actual data curves and it becomes obvious that in this case the data were poor and thus resulted in low quality hits. This is an example where the curves would be discarded.

In addition, the same example plate illustrates that even though some compounds produce high quality data, the different methods result in significantly different shifts. Table 8.3 lists two wells (well number 36 and well number 48) with different compounds where the compound is classified differently (either as hit or non-hit following the rule that the shift should be above three times the standard deviation) depending on the method used for T_m determination. Additional, Figure 8.6 shows differences in hit classification also occur when data curves and fits are of good quality. Well number 36 is shown and illustrates the trustworthiness of the newly five-parameter equation.

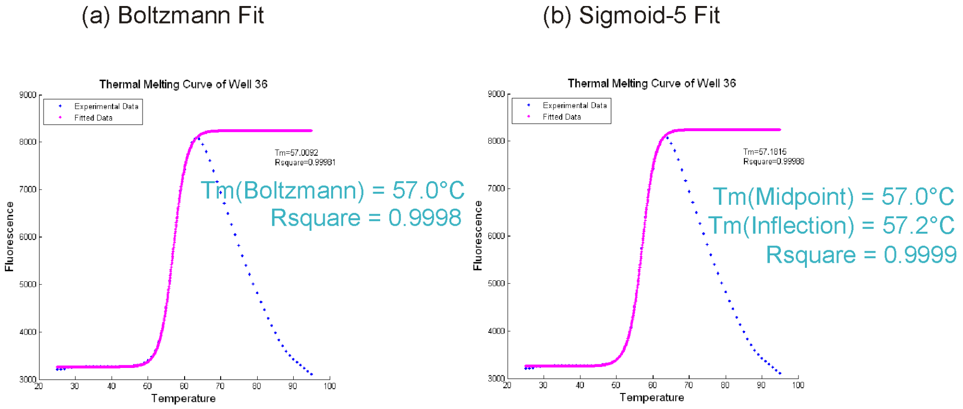
In summary, if a curve is misshaped (due to multiple unfolding transitions, ligands having intrinsic fluorescence etc.) none of the fitting equations will produce reliable data.



(a) Boltzmann Fit

(b) Sigmoid-5 Fit

Figure 8.5: **Bad fits resulting in odd shifts.** Well number 6 from the above presented screening plate showed strongly opposed shifts in the hit analysis. However the inspection of the actual data curve makes clear that the fit is poor and the data cannot be trusted.



(a) Boltzmann Fit

(b) Sigmoid-5 Fit

Figure 8.6: **Good fit with different Tm depending on method.** Well number 36 from the above presented screening plate is very well fitted with an rsquare of 0.9999. Despite, this could be considered either as a hit or not depending on the method used. The rsquare is calculated for data points included in the fit, i.e. for the points from the left hand minimum to the maximum of the curves (see Chapter 4).

8.5 Known Binders

All of the examples mentioned above derive from screening random compounds. Binding and affinity are unknown. This section now examines how some known binders to BtGH84 behave when analysed by these methods. The compounds PUGNAc and ysbl000293 bind to BtGH84 with an IC_{50} in low micromolar range

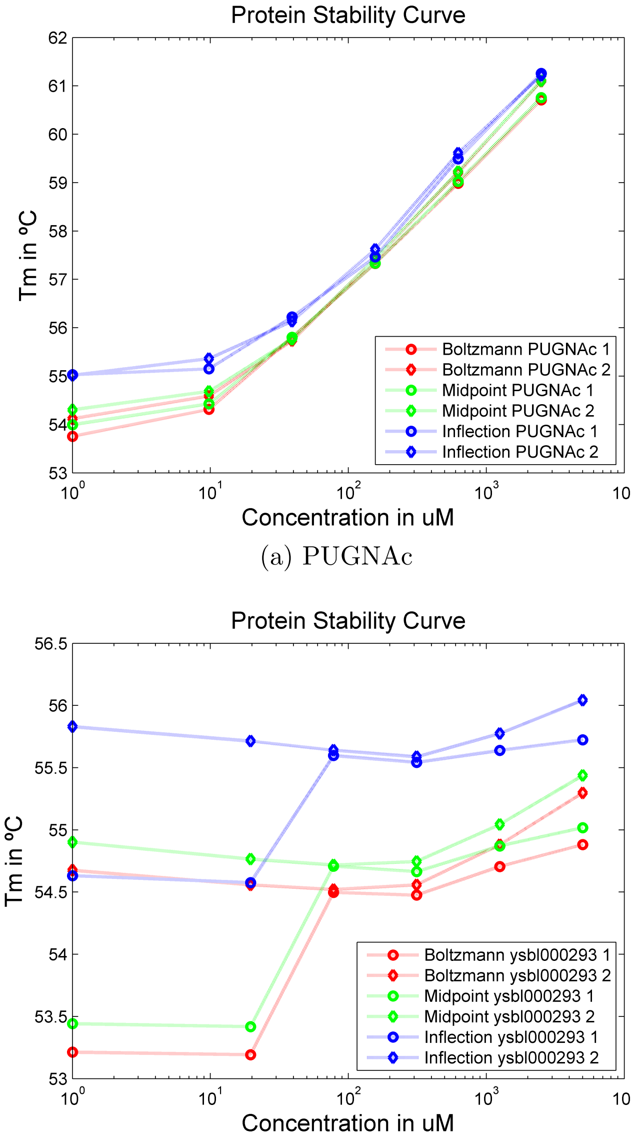
Table 8.3: **Hit dependency on Tm determination** One screening plate of neighbours with 32 blanks contains more than one example where the T_m determination is crucial for hit classification.

	32 Blanks		**Well 36**		
Method	**Average**	**Stdev**	**Tm**	**Shift**	**Rsquare**
Boltzmann	55.95	0.24	57.009	1.057	0.9998
Midpoint	56.07	0.23	57.043	0.977	0.9999
Inflection	56.67	0.20	57.182	0.512	0.9999
	32 Blanks		**Well 48**		
Method	**Average**	**Stdev**	**Tm**	**Shift**	**Rsquare**
Boltzmann	55.95	0.24	56.712	0.760	0.9999
Midpoint	56.07	0.23	56.690	0.623	1.0000
Inflection	56.67	0.20	56.875	0.206	1.0000

and 600 µM respectively. Figure 8.7 shows the protein stability curves of both compounds analysed with all three T_m methods recorded at duplicate concentration series. Blanks were plotted as 1 µM for logarithmic purposes. The curves for PUGNAc (Figure 8.7 a) show an increase in melting temperature with every concentration step.

Dose-response curves created with TSA are not typical dose-response curves. As the protein-ligand complex is heated up and denatures, the concentration of native protein decreases and (probably) the free concentration of the ligand will increase. In addition, the solubility of the ligand and its binding affinity for the native and denatured protein could be affected. This combination of effects will introduce variability from one ligand to another, and could also affect the detailed shape of the TSA curves .

The T_m values obtained with all three methods are in good agreement. However, for the blanks, the Inflection T_m is about 1℃ higher than the other two methods. The greater the ligand concentration, the smaller the difference in T_m. At 2.5 mM the calculated T_m are almost the same across the three methods. A similar trend can be observed for the fragment ysbl000293 (Figure 8.7 b). However, the results are far noisier for the weaker binding compound. There is a significant difference between the line with the circle markers (1) and with the diamond markers (2) for the blanks and the lowest concentrations of fragment. Nevertheless, Midpoint and Boltzmann produce the larger temperature shift for both control compounds.



(a) PUGNAc

(b) ysbl000293

Figure 8.7: **Comparison of the Tm methods using positive controls.** Protein stability curves of BtGH84 with positive controls with all three T_m methods. The point for 0 µM ligand was modified to 1 µM to allow a logarithmic plotting. (a) PUGNAc (b) Fragment ysbl000293

8.6 Discussion

Due to the time restraints for completing this thesis, the analysis presented above is on representative data. It would be very interesting to perform a statistical analysis on a wider collection of screening data. However, some general conclusions can be made.

The previous examples illustrate that the method of data processing used is crucial for determining if a fragment is a hit or a non-hit. In addition, there is no method for straight-forward readout of the transition point of a half unfolded protein ($\mathrm{T_m}$). The $\mathrm{T_m}$ must be estimated using a fitting model. However a different thermal shift is produced by the different methods used calculate the $\mathrm{T_m}$. This is probably because the physical model of the experiment is incomplete and there are additional physical events happening during unfolding which are not accounted for in the existing models. These could for example be changing heat capacity or cooperative unfolding. Nevertheless the fit statistics deriving from the Sigmoid-5 equation are mostly better (never worse) than those of the Boltzmann model. For this reason the Sigmoid-5 equation is the more robust model. The effect is self explanatory since an additional parameter is used in refinement. From the data produced in this thesis it cannot be clearly said which method of $\mathrm{T_m}$ determination is the most appropriate.

In theory, the observed signal in TSA is independent of ligand size and affinity. However, the thermal shift for a strong binder like PUGNAc will be significant, whereas it is less for small fragments with a weak affinity. As a result, the stability curves become more noisy and the compound is likely to be missed as a hit with the TSA method. Binding affinity detectable with TSA has also been limited by others to 100 µM–1 M $\mathrm{K_D}$ (Kranz and Schalk-Hihi, 2011).

All the examples mentioned in this chapter lead to the conclusion that TSA is a good screening method to test stronger binders. However, the method seems unsuitable for fragment screening due to the failure of proper $\mathrm{T_m}$ determination.

The Boltzmann equation and general inflection point methods are widely accepted for $\mathrm{T_m}$ determination (Sections 4.2.2 and 4.2.3). To remain in accordance with the definition of the melting temperature (where the protein is half unfolded and half folded), the $\mathrm{T_m}$ corresponds to the midpoint of a curve and *not* to the inflection point. However, the Sigmoid-5 equation represents a more robust way of fitting than the Boltzmann equation. In addition, the heat capacity of unfolding is not independent from the temperature (Gomez et al., 1995) as inadequately presupposed by the thermodynamic models (Section 4.2.1). The additional parameter of the Sigmoid-5 equation could account for the changing heat capacity and justifies that an asymmetric fitting equation is more appropriate for thermal denaturation curves.

Altogether, this leads to my recommendation that the midpoint of the Sigmoid-5 equation is the most suitable T_m definition for future applications. Nevertheless, users of that technique should be more aware that the T_m readout is not unambiguous.

Chapter 9

Summary and Future Perspectives

9.1 Summary

The work presented in this thesis comprises three major parts: Fragment library design, the coding of the program MTSA to analyse thermal shift data and the assessment of protein targets for screening.

9.1.1 Fragment Library Design

The aim of the library design presented in this thesis (Chapter 3) was the development of Pipeline Pilot protocols which automatically select a fragment set from an input library. Compound diversity and representation of the input library in the form of substructures was optimised. The presumptions will guarantee applicability to the SAR by catalogue approach where fragment hits are searched for superstructure neighbours in a chemical compound catalogue. Five protocols were developed which are named as follows:

- *Cluster All*
- *Cluster Fragments*
- *SIM within Cluster*

- *Substructure Count* and *Substructure Map*
- *Iterative Removal*

All library design protocols can be downloaded from the fragments website http://www.ysbl.york.ac.uk/fragments/protocols/ and are straight forward to use.

An evaluation strategy was also implemented in Pipeline Pilot which generates information about similarity to known drugs and inter-library similarity. Each of the protocols named above was used to generate three different fragment libraries with input compounds from different suppliers. These fragments libraries were assessed with the evaluation protocols and profiled for physicochemical representation of the input libraries and the protocols were compared in order to produce the best quality fragment library. It was found that the procedure *Iterative Removal* produces the best quality library.

In two rounds, the in-house fragment libraries "Michele_1" and "Michele_2" were generated with the *Iterative Removal* protocol, purchased, controlled for quality and set up in the database InstantJChem (Appendix C).

Different strategies to design fragment libraries have been approached by scientists: Intense Filtering and visual inspection, 3-dimensional shape-based fragments, focussed libraries and more (Section 1.3).

The procedure presented in this thesis gives a new perspective and is in particular useful for the SAR by catalogue approach. The fragments in the libraries generated with the *Iterative Removal* protocol represent substructures of an input library. Hits from screening campaigns can be quickly searched for neighbours in the input library which emphasises the advantage of this user-friendly protocol. Researchers can use this protocol to design a fragment library where evolution of the fragments can be rapidly trialled as multiple near neighbours are accessible for assay. These can be taken from internal collections, designed compound libraries or the commercially available compounds.

Groups new to fragment screening can use the protocols to quickly generate their own small fragment library based on laboratory compounds or selecting from their favourite supplier. This way they can easily perform initial trials with fragments without needing a chemistry team designing such a library.

Further, the protocol can be used to design a fragment library that provides a "window" into a corporate collection to complement high throughput screening strategies. For example, an organization could maintain a fragment library that maximally represents the screening collection. A preliminary fragment screen can be performed on which to judge the ligandability of the target and/or as a precursor to using the fragment hits to select larger compounds for assay. Increasingly, fragment screening is being performed alongside HTS and having such a representative fragment library could bring benefits.

9.1.2 MTSA

The program MTSA was written to facilitate the analysis of the thermal shift experiments for which no free software was available by the start of that project (Chapter 4). MTSA was implemented with Matlab and provides an easy way to process the raw fluorescence data of the thermal shift experiment. For each experiment, the program cuts the relevant data and fits the curve with a logistic (sigmoidal) four- or five-parameter equation and extracts the melting temperature T_m. If screening plates were set up with a recurring blank (e.g. in the first column of the plate), the program also calculates the average T_m of the blanks, $\overline{T_m}$, and the thermal shift of all experiments in respect to the average blank ΔT_m. The program warns the user if the standard deviation of the blanks surpasses a certain threshold because this could be a sign of errors in the assay. Further, for each experiment an image is generated.

The choice of the fitting equation is crucial with this technique (Chapter 8). The four-parameter logistic equation, also known as the Boltzmann equation, has the property that the T_m is situated at inflection- and midpoint. If using the five-parameter equations, the T_m can be defined either at the midpoint or at the inflection point. All three definitions deliver different results for T_m and for the thermal shift which leaves room for suggestions that the method is very fragile when applied to weak binding fragments.

In summary, MTSA significantly improves and accelerates the analysis of thermal shift experiments and is especially helpful for high throughput of plates. It can be downloaded from the fragments website http://www.ysbl.york.ac.uk/fragments/MTSA.

Concerning the experimental technique itself, TSA is increasingly used for fragment screening. Giving the reason that fragments only produce small temperature shifts places particular demands on the methods for analysis of the experimental data. The work on MTSA led to the question whether using TSA for fragment screening is valuable or not. One challenge here are false negatives. Not every ligand that binds to a protein will also stabilise it significantly which is especially true for smaller, weak binding compounds. Tight binding ligands usually translate into a high thermal shift and TSA might not be suitable for very small fragments. Also ligands can interfere with the fluorescent dye or have intrinsic fluorescence. And as common with fragment screening, the thermal shift assay is limited by compound solubility.

Offering a medium to high throughput, TSA would be a good primary screening method for fragments. The difficulty of obtaining affinity data also accounts for the use as a primary technique. (The preference for cross-validation screen should lay in a method which easily provides affinity data.) In contrast, a technique with a number of false negatives should rather be considered as an orthogonal assay.

The low material consumption is outweighed by the need to test every fragment multiple times because the repeatability of the results is limited (in this thesis, fragments were screened as quadruplicates).

As every screening method, TSA has pros and cons which should be kept in mind when setting up a screening assay. Notwithstanding, TSA is a good technique for what it was developed originally - to identify ligands and buffers that stabilise a protein and an aid to crystallisation.

9.1.3 Fragment Screening

The third part of this work deals with some of the practical aspects of fragment based ligand discovery in the form of the targets NMT, HEWL and BtGH84.

N-myristoyl transferase (NMT) from the parasites *Leishmania donovani* (ld), *Leishmania major* (lm) and *Trypanosoma brucei* (tb) were assessed for their suitability for fragment screening (Chapter 5). The parasites are responsible for the neglected diseases visceral and cutaneous Leishmania and sleeping sickness. The project was planned as an exciting collaboration between different institutions. My part of the project was originally described as performing a fragment screen with one of the NMT proteins, cross validating hits coming from a high throughput screen and solving co-crystal structures with promising hits.

ldNMT was successfully subcloned to obtain a protein construct with a double His-tag. Expression and purification trials were performed with both single His-tagged lmNMT and double His-tagged ldNMT. However, production of a suitable amount of stable protein failed. Nevertheless, initial SPR experiments were performed to assess the suitability using that method to screen the fragment libraries. Experiments turned out to be challenging. Issues derived from producing a stable baseline of the protein and observing co-factor binding.

While this project was in process, the widely used test protein HEWL came up during some experiments. The route of the over-all project was changed to establish some fragment screening assays with the model protein hen egg white lysozyme (HEWL). Thus the NMT project was postponed to a later stage.

Although the glycanhydrolase HEWL is often used as a model system, there are no small molecule inhibitors known to date. Together with being a protein which can be purchased at low costs which also easily crystallises, HEWL seemed a suitable target to screen the previously generated fragment library (Chapter 6).

A novel Biacore assay was developed where HEWL binds to the dextran surface of a CM5 chip and almost 50 fragments from "Michele_1" were screened. Interesting hits were confirmed with an established enzyme activity assay using whole cells as substrate. Another 19 fragments from "Michele_2" were screened with thermal shift analysis. Initial crystallisation trials were also performed. Nevertheless HEWL turned out to be more challenging than expected and does not behave straightforwardly in other assays. Results from the different assays could not be reproduced and optimisation of the experiments was tricky.

At the time of the initial crystallisation experiments with HEWL, the attention moved to the protein BtGH84 which turned out to be a much simpler target. Therefore HEWL was put to the side for possible later explorations.

Bacteroides thetaiotaomicron VPI-5482 glycoside hydrolase from GH family 84 (BtGH84) is a close homologue to the human enzyme O-GlcNAcase which is considered to play a key role in Alzheimer's disease, cancer and diabetes type II. Besides being a scientifically strongly relevant target, the protein is also easy to handle and reasonably stable. Almost 500 compounds were screened against the target with thermal shift analysis (Chapter 7). Interesting hits were further tested for their K_D with Biacore experiments and crystal structures with soaked ligands were solved. Unfortunately, the soaking conditions proved to be unsuitable and no ligands were present in the structures. Nevertheless the numerous results show the assays to be good starting points for further studies.

This part of the thesis proved that not every target is suitable for fragment screening. The bottle neck is the protein which needs to be producible in suitable amounts and stable in a wide range of conditions. Interesting drug targets can offer a great challenge for ligand discovery. Fragment screening can be a greatly useful tool to advance the project as with BtGH84, but might not work for other targets such as NMT. Proteins which behave well with the challenging X-ray crystallography can propose problems with other screening techniques - like HEWL forming aggregates at a wide pH range. No experimental screen in is suitable for every target.

Testing these different targets gave insight into the different fragment screening methods and the chance to test the suitability of the prior designed fragment library. The results of BtGH84 should be used for follow-up experiments to confirm the initial work.

9.2 Future Perspectives

9.2.1 Analysing the Fragment Library

The generated fragment libraries will now be used by other members of YSBL. It remains fascinating to see how the generated fragment library would apply to different targets. I would suggest screening the library against a couple of proteins of different classes (i.e. one kinase, one phosphatase, one sugar binding protein, one transferase), evaluate the generated hit rates and classify the types of hits. Results of such an evaluation would give information about the general applicability of the library and more details about the usefulness of the individual compounds as member of a screening library.

9.2.2 Improving MTSA

The program MTSA provides a huge platform for extensions and improvement. First of all, one could add functions for the user to choose if the analysis should be performed with the T_m methods Boltzmann, Midpoint or Inflection, or combinations of them. The output could be improved, e.g. with tables and histograms. One has to make sure that the program still remains applicable for use with different qPCR machines and operating systems. One could also envision the implementation of an automatic t-test which would be performed to assess the statistical difference between two curves. Currently, commercial interest was shown in the program. There may be opportunities for a collaboration so that I can make further improvements.

9.2.3 Melting Point Discussion

Another exciting aspect would be the gathering and analysis of more information about the actual melting point definition. First, it would be interesting using the improved version of MTSA to analyse other targets for their difference in thermal shift depending on the use of the different T_m methods. If enough data from different targets were obtained, a complex analysis could be generated. It would be especially intriguing to find out if the Inflection method always generates higher melting temperatures than Boltzmann and Midpoint. One could also keep the Hill factor a constant for one target, and assess the behaviour of the asymmetric factor c.

It would be useful to develop a new thermodynamic model which includes the temperature dependency of the heat capacity of unfolding. Such a model is likely to be in good accordance with the deployed asymmetric Sigmoid-5 equation.

9.2.4 Further Work on HEWL

For HEWL, work on the co-crystal structures could be continued. Although the screening results were not greatly reproducible, there is a strong suggestion that the compounds N-acetylglycine, o-toluic acid and benzothiazole alter the protein activity. Crystal structures with the compounds would give further conclusions and ideas how to proceed.

The fragment ysbl000497 altered the shape of the melting curve while producing a light double-hump at the top of the curve. It would be interesting to perform more experiments with this compound to assess how it is affecting the unfolding process. The NMR spectra to control the quality of the compound indicated the compound was only present at low concentration, so it would be reasonable to perform future experiments with much lower concentration. Thus disturbances by precipitate would be ruled out.

In addition, the novel Biacore assay could be improved and tested for suitability also with other sugar binding targets.

9.2.5 Following Up BtGH84

An evaluation of all TSA data with all three T_m methods suggests very intriguing results. The new profiling will give different conclusions about possible binders. The hits could be followed up via crystal structure determination. To do so, the soaking conditions for BtGH84 crystals need be improved to ensure the ligands are entering the crystals. These structures would give more starting points for further investigation.

The influence of DMSO on stability of BtGH84 is an interesting factor which should be further explored. Whether there is an the influence on stability of the dyes SYPRO orange and deep purple could also be explored further.

Chapter 10

Materials and Methods

10.1 Fragment Library

The fragment library was designed with the *Iterative Removal* procedure in two steps. The second iteration was built on the first one and passed some additional selection filters using JChem to remove duplicates and compounds >80% Tanimoto similarity to members of the first library. (The details are in Chapter 3 and Appendix C.)

10.1.1 Handling and Storage of the Fragments

The compounds were purchased from Specs, Maybridge, Asinex and Sigma Aldrich and prepared as 200 mM stock solutions in deuterated DMSO (DMSO-d_6). The solutions were stored on plates at room temperature in the dark.

10.1.2 Quality Control

The quality of the members of fragment library "Michele_2" was checked using 1D ^{1}H spectra with 32 scans on a JEOL 400 MHz NMR spectrometer. Compounds were tested at 2 mM in D_2O with 100 µM TSP as a reference. Spectra were assessed with the help of Dr. Jens Landström. The quality of the members of fragment library "Michele_1" had been demonstrated by Dr. Kerrin Bright.

10.2 Subcloning ldNMT

10.2.1 Template

The clone of ldNMT with a single N-terminal His$_6$-tag (H6ldNMT) was obtained from Dr. Jim Brannigan. The gene was in vector pSKB2 also known as pET28-PPX with a protease cleavage site. The vector was diluted to 6 ng/µl in water.

10.2.2 Cloning Strategy

The coding sequence of ldNMT was PCR amplified using primers (Eurofins MWG Operon) containing NheI (AGGCTATAAGCTAGCCATCATCATCATCATCAC AGCAGC) and SacI (ATTGCAGTGGTGGAGCTCGAGCTACAA CATCAC) restriction sites. pET28a vector and the PCR amplified insert were digested with NheI (Promega) and SacI (Fermentas) restriction enzymes for ligation to generate double (N-terminal) His$_6$-tagged ldNMT (2H6ldNMT) clone. Primers were used at 25 pmol/µl.

10.2.3 PCR

The PCR followed a touchdown protocol (Don et al., 1991) with the annealing temperatures 63℃ → 61℃ → 59℃ → 57℃ → followed by 20 cycles at 55℃ and was carried out in 50 µl volume with dNTPS (deoxynucleotides), $MgSO_4$, Hot Start KOD DNA polymerase (all Novagen), forward and reverse primers (Eurofins MWG Operon) and H6ldNMT template.

10.2.4 Agarose Gel Extraction

Amplified DNA and vector digests were separated by agarose gel electrophoresis with 1% agarose and 0.7% agarose respectively, in 70 ml TAE and 1 µl Sybr safe. A 2log DNA ladder (NEB) was used to determine the size of DNA fragments. Products were extracted using a gel extraction kit (Qiagen). The final concentrations were quantified by absorbance at 260 nm using a NanoDrop ND-1000 spectrometer (Thermo Scientific).

10.2.5 DNA Hydolysis with Restriction Endonucleases

To create sticky ends for ligation, the PCR product and the vector were digested with NheI (Promega) and SacI (Fermentas) restriction enzymes in multi core buffer (Promega) with BSA (Bovine Serum Albumin, Promega) overnight at 37°C.

10.2.6 Dephosphorylation

The purified vector was dephosphorylated with CIAP (calf intestinal alkaline phosphatase) (Promega). Therefore, 3 µl CIAP were added and incubated at 37°C for 15 minutes, at 56°C for 15 minutes, and repeated these steps once more (protocol from Promega CIAP manual "Dephosphorylation of 5' Recessed or Blunt Ends").

10.2.7 Ligation

A three fold molar excess of PCR product to digested vector was used in the ligation reaction (including ligation buffer and T4 DNA ligase, both BioLabs). One control ligation without PCR product was also set up. The mixtures were incubated at room temperature for one hour and then overnight at 16°C.

10.2.8 Transformation

The ligation mixtures were transformed into 50 µl competent *Escherichia coli* XL10 Gold cell aliquots containing β-mercaptoethanol. 5 µl of ligation or control ligation were added to aliquots of competent cells. For transformation the cells were incubated for 30 minutes on ice, heat-shocked for 1 minute at 42°C and subsequently cooled on ice for 2 minutes (Stratagene transformation protocol). 1 ml prewarmed LB medium was added to each transformation mixture and the tubes were incubated at 37°C for 2.5 hours at 220 rpm. Afterwards, 100 µl of each cell suspension were plated out on LB plates supplemented with 30 µg/ml kanamycin and incubated overnight at 37°C.

A storage plate supplemented with 30 µg/ml kanamycin was prepared, and colonies were used to inoculate 5 ml LB supplemented with 30 µg/ml kanamycin, and these colonies were also used to streak a LB agar plate containing 30 µg/ml kanamycin. The storage plate was incubated at 37℃ overnight. The 5 ml cultures were incubated at 37℃ overnight at 180 rpm. The storage plate was stored at 4℃ for later use. The plasmids were extracted using a GeneElute Plasmid Mini PrepKit (Sigma Aldrich) following the manufacturer's instructions, from which the OptiWash step was excluded.

10.2.9 Assessing Cloning Success

All five plasmids were digested in 9 µl volume in multi core buffer (Promega) containing NheI (Promega) and SacI (Fermentas) at 37℃ for 3 hours. The plasmid concentration was measured by absorbance at 260 nm using a NanoDrop ND-1000 spectrometer (Thermo Scientific). DNA samples were separated by agarose gel electrophoresis using 1 µl loading dye and 1% agarose gel. Two samples were selected for sequencing by the sequencing service of the TF (Technology Facility, Department of Biology, York University).

The sequences were analysed with programmes of the following websites:

http://bioinfo.hku.hk/services/menuserv.html
http://www.ebi.ac.uk/Tools/clustalw2/index.html

10.2.10 Recombinant 2H6ldNMT Expression Trial

The single His_6-tagged clone from Dr. Jim Brannigan has been expressed successfully in *Escherichia coli* Rosetta(DE3) pLysS. Rosetta strains provide seven codons rarely used in *Escherichia coli* on a chloramphenicol plasmid.

1 µl of the plasmid stock was added to 50 µl of competent Rosetta 2 (DE3). Cells were transformed as described above (Section 10.2.8). Subsequently, the transformed cells were added to 1 ml preheated LB containing 30 µg/ml chloramphenicol and 30 µg/ml kanamycin. After shaking for 2 hours at 37℃, 50 µl and 100 µl of the cell suspension were plated out on LB plates containing 30 µg/ml chloramphenicol and 30 µg/ml kanamycin, and incubated overnight at 37℃. One of the

overnight colonies was picked to inoculated 5 ml LB medium containing 30 µg/ml chloramphenicol and 30 µg/ml kanamycin, and shaken overnight at 37℃. Three 50 µl aliquots of the overnight culture were added to new Sterilin tubes containing 5 ml LB supplemented with 30 µg/ml chloramphenicol and 30 µg/ml kanamycin. They were incubated at 37℃ until they reached an OD600 (optical density measured at 600 nm) of 0.5 at which point expression was induced by addition of 1 mM IPTG and incubated at three temperatures:

16 ℃ → overnight

30 ℃ → 5 hrs

37 ℃ → 3 hrs

Cells were pelleted at 13,000 rpm for 3 minutes, resuspended and frozen. On the following day, the cells were collected into 1.5 ml Eppendorf tubes by centrifugation at 13,000 rpm for 3 minutes on table centrifuge (GenFuge Progen). The supernatant was discarded. The pellets were resuspended in 0.5 ml buffer as described below by vortexing and pipetting up and down. The cells were lysed by sonication (Soniprep 150) for 3 x 5 seconds and cells were stored on ice between sonication intervals. 5 µl aliquots of each sample were reserved as total cell lysate. Remaining lysates were pelleted at 13,000 rpm for 3 minutes. A 5 µl aliquot of the supernatant was taken to represent the soluble fraction. The pellet was resuspended to take a 5 µl aliquot of the insoluble fraction. The aliquots were assessed on a 12% SDS-PAGE (gel which was run at 200 V for 50 minutes using BioRad low molecular range marker.

A range of lysis buffers was tested to assess protein solubility, including a sugar buffer (40% sucrose, 30 mM NaCl, 50 mM tris-HCl pH 8.0, 0.1% Triton-X 100), a tris buffer (50 mM tris HCl pH 8.0, 300 mM NaCl) a tris glycerol buffer (50 mM tris HCl pH 8.0, 300 mM NaCl, 10% glycerol) and a phosphate buffer (100 mM potassium phosphate pH 8.0, 150 mM NaCl).

10.3 Large Scale Protein Production

10.3.1 Autoinduction Media

The following autoinduction media as described in Studier (2005) were used in the large-scale protein production described below:

- **50x5052:** 25 g glycerol, 73 ml Milli-Q water, 2.5 g glucose, 10 g α-lactose
- **25xM:** 3.6 g Na_2SO_4 anhydrous, 13.4 g NA_4Cl anhydrous, 17.0 g KH_2PO_4 anhydrous, 17.7 g Na_2HPO_4 anhydrous, dissolved sequentially in 200 ml Milli-Q water
- **ZY:** 5 g tryptone, 2.5 g yeast, 468.5 ml Milli-Q water

10.3.2 ldNMT

Protein Expression

Escherichia coli Rosetta2 DE3 cells containing a pET28a vector with N-terminal double His_6-tagged ldNMT (2H6ldNMT) were used to inoculate an overnight culture of 20 ml LB with 30 µg/ml kanamycin and 30 µg/ml chloramphenicol. 1 ml of the overnight culture was added to tubes of 20 ml LB with 30 µg/ml kanamycin and 30 µg/ml chloramphenicol and grown at 37℃ for 1 hour. Then each tube was added to a separate prewarmed autoinduction media (500 ml ZY media with 1 mM $MgSO_4$, 10 ml 50x5052, 20 ml 25xM, with 30 µg/ml kanamycin and 30 µg/ml chloramphenicol) and grown for 6 hours at 37℃. The temperature was reduced to 20℃ for overnight growth. The following day, cells were harvested for 15 minutes at 6,000 rpm and 4℃ using a Sorvall RC5B centrifuge. Cells were resuspended with 20 ml phosphate buffer and the soluble fraction isolated by centrifugation for 15 minutes at 6,000 rpm and 4℃ using a Sorvall RC5B centrifuge.

Protein Purification

The cells with double His-tagged ldNMT were resuspended in extraction buffer (nickel chelating buffer A with 0.5% Triton-X 100, a small amount of DNAse

(Sigma Aldrich), 10 mM $MgCl_2$, 1 tablet Complete EDTA-free protease inhibitor cocktail tablet, Roche) and lysed by sonication (Soniprep 150) 6 times for 30 seconds on ice. The crude lysate was loaded on a HisTrap crude 1 ml column (GE Healthcare) using a peristaltic pump at 4℃. The column was washed with nickel chelating buffer A (20 mM sodium phosphate pH 8.0, 300 mM NaCl, 20 mM imidazole). Elution was performed on an AKTA system at 4℃ with a gradient of nickel chelating buffer B (20 mM sodium phosphate pH 8.0, 300 mM NaCl, 1 M imidazole) over 20–30 column volumes.

The eluted protein was slowly diluted 1:10 in anion exchange buffer A (20 mM tris pH 8.0, 20 mM NaCl, 1 mM DTT). If there were remaining particles, the sample was filtered with a 0.45 µM filter unit and loaded on a HiTrap Q FF 5 ml column (GE Healthcare). An ion exchange chromatography was performed using an AKTA system at 4℃.

The elution was performed using a step gradient with increasing amounts of anion exchange buffer B (20 mM tris pH 8.0, 500 mM NaCl, 1 mM DTT). The elution started at 40% buffer B for 10 column volumes, increased to 100% buffer B over 5 column volumes, followed by a wash step the duration of 5 column volumes at 100% buffer B.

10.3.3 lmNMT

Protein expression followed an autoinduction protocol (Studier, 2005) adapted by Dr. Jim Brannigan, similar to above.

Protein Expression

Escherichia coli BL21pLysS transformed pET15b and N-terminal His_6-tagged lmNMT (H6lmNMT) were used to inoculate an overnight culture of 20 ml LB with 100 µg/ml ampicillin and 30 µg/ml chloramphenicol. 1 ml of the overnight culture was added to tubes of 20 ml LB with 100 µ/ml ampicillin and 30 µg/ml chloramphenicol and grown at 37℃ for 1 hour. Each culture was added to separate prewarmed autoinduction medium (500 mL ZY medium with 1 mM $MgSO_4$, 10 ml 50x5052, 20 ml 25xM, with 100 µg/ml ampicillin and 30 µg/ml chloramphenicol) and grown for 6 hours at 37℃. The temperature was reduced to 20℃ for overnight growth. The following day, soluble fractions were isolated by centrifugation at 6,000 rpm for 15 minutes and 4℃ using a Sorvall RC5B centrifuge. Cells were resuspended in 20 ml phosphate buffer and centrifuged again for 15 minutes at 6,000 rpm and 4℃ on the Sorvall RC5B.

Protein Purification

The purification was performed at 4℃. Cells were resuspended in extraction buffer (40% sucrose, 300 mM NaCl, 10 mM imidazole, 50 mM tris-HCl, 0.1% Triton-X, pH 8.0), containing a small amount of DNAse (Sigma Aldrich) and one tablet Complete EDTA free protease inhibitor cocktail tablet, Roche, per tube, and were lysed using a French Press (Constant Disruption System). The crude lysate was loaded onto a NiChelating HisTrap crude 1 ml column (GE Healthcare) at 0.3 ml/min and washed with Nickel chelating buffer A (20 mM sodium phosphate pH 7.4, 300 mM NaCl, 20 mM imidazole). Elution was performed with a linear gradient up to 100% buffer B (20 mM sodium phosphate pH 7.4, 300 mM NaCl, 500 mM imidazole) over 15 column volumes. The eluted protein was further slowly diluted 10fold in anion exchange buffer A (20 mM tris pH 8.0, 20 mM NaCl) and filtered through a 0.45 µm filter. The protein was loaded onto a HiTrap 5 ml column (GE Healthcare), and eluted with anion exchange buffer B (20 mM tris pH 8.0, 500 mM NaCl) using a step gradient composed of: 10 column volumes at 40% buffer B, 5 column volumes at 100% buffer B, and then washed at 100% buffer B for another 5 column volumes.

The protein was flash frozen in liquid nitrogen and then stored at -20℃.

10.3.4 tbNMT

The protein with a N-terminal His$_6$-tag for some initial experiments was kindly provided by Dr. Jim Brannigan.

10.3.5 BtGH84

The protein was expressed and purified as described by Dennis et al. (2006).

Protein Expression

Escherichia coli BL21(DE3) with the N-terminal His$_6$-tagged BtGH84 gene on ysblLIC pET28 vector were used to inoculate 8 ml LB medium containing 40 µg/ml kanamycin and which was grown overnight at 37℃. The next day, the small flasks were transferred into flasks containing 0.8 l sterile LB media supplemented with 40 µg/ml kanamycin. Cells were grown at 37℃ until an OD600 of 0.8 and were then induced with 1 mM IPTG. In the following, the temperature was reduced to 16℃ and the cells grew overnight. The following day, the cells were harvested at 5,000 rpm for 25 minutes at 4℃. The pellets were resuspended in 10 ml buffer A (20 mM HEPES pH 7.5, 200 mM NaCl, 20 mM imidazole) and frozen at -20℃.

Protein Purification

Cell pellets were defrosted and resuspended in 30 ml buffer A was with one tablet Complete EDTA free protease inhibitor cocktail tablet (Roche). Cells were sonicated (Soniprep 150) on ice 15 times for 10 seconds and 20 seconds rest between sonication steps. The lysate was pelleted by centrifugation for 40 minutes at 15,000 rpm and 4℃. The supernatant was filtered with a 0.22 µM filter and loaded onto a 5 ml precharged nickel column (GE Healthcare) The elution followed a gradient of 0 – 100% buffer B (20 mM HEPES pH 7.5, 200 mM NaCl, 500 mM imidazole) over 40 column volumes.

10.3.6 HEWL

HEWL was purchased from Sigma and dissolved in Milli-Q water to the required concentration. For some experiments the protein was further purified by buffer exchange to reduce the ion concentration.

10.4 Mass Spectrometry

The experiments on the electrospray ionisation (ESI) mass spectrometer were kindly performed by Simon Grist using a QSTAR orthogonal acceleration time of flight mass spectrometer (Applied Biosystems). The protein sample was concentrated to 2 mg/ml and further buffer exchanged into 2 mM tris pH 8.0. The concentrated protein was diluted to 4 µM in a solution of 50% acetonitrile and 0.05% formic acid to give the positive charge.

10.5 Surface Plasmon Resonance

When light comes from a medium with a higher refractive index it is partially refracted and partially reflected. Above a critical angle in incidence, light is only reflected. However, the electromagnetic field component penetrates a couple of hundreds of nm into the surface with the lower refractive index creating an evanescent wave. Under certain conditions such as the light being monochromatic and p-polarised and the interface between the two media is surfaced with gold, the intensity of reflected light reaches a minimum. This dip is called surface plasmon resonance (SPR). The resonance conditions are influenced by material which is adsorbed onto the metal film. SPR varies almost linearly with the weight of biologically relevant molecules and can be used to measure their concentration on the surface.

SPR can detect binding of small molecules via changing of the resonance conditions on a gold surface. The technique was applied to NMT, HEWL and BtGH84. Concentration series in six wells in 1/4 dilution steps were tested for affinity determination where the lowest concentration was usually a blank and the highest concentration ten times the expected K_D.

10.5.1 Testing NMT

Different test experiments with lmNMT, ldNMT and tbNMT were performed. Initially, the NMT proteins were tested for binding to an NTA chip (nitrilotriacetic acid chip, GE healthcare) via their His-tags. After washing the chip with HBS-P buffer (0.01 M HEPES pH 7.4, 0.15 M NaCl, 50 µM EDTA, 0.05% Surfactant P20), nickel was attached to the chip and 100–200 nM protein flushed over it. 100 nM myristoyl coenzyme A (MCoA, Sigma Aldrich) in running buffer (HBS-P or tris, 200 mM NaCl, 50 mM tris-HCl, pH 7.5) was used as a positive control. Covalent binding of the protein to the chip was used when too much bleeding-off occurred and no binding of the positive control was observed.

The proteins were covalently bound to the chip following the amine coupling kit by Biacore. 500 µM $NiCl_2$ in HBS-P+ buffer was injected to bind to the NTA groups. The dextran surface of the chip was activated with 0.2 M/0.05 M EDC/NHS at 5 µl/min for 8 minutes. His-tagged protein at 10 µg/ml was injected at 5 µl/min to bind covalently to the chip until the response reached 6,000 to 8,000 units. Thereafter, the remaining active groups were deactivated with ethanolamine at 5 µl /min for five minutes.

Buffers such as HBS-P and tris were tested. The standard regeneration solution was HBS-P buffer containing 350 mM EDTA. If a stronger regeneration was needed 5 mM NaOH was used. The nickel solution was prepared by dissolving 0.5 mM $NiCl_2$ in HBS-P buffer. Protein was handled at a concentration of 10 µg/ml. The running buffer was filtered using a 0.2 µm filter.

10.5.2 Binding Analysis with BtGH84

The chip surface was washed with regeneration solution (10 mM HEPES, 150 mM NaCl, 0.05% surfactant P20, 350 mM EDTA, pH 7.4) followed by a wash with HBS-P+ buffer (10 mM HEPES, 150 mM NaCl, 0.05% surfactant P20, 50 µM EDTA, pH 7.4). 500 µM $NiCl_2$ in HBS-P+ buffer was injected to bind to the NTA groups. BtGH84 was covalently bound to an NTA chip following the standard protocol with the Biacore amine coupling kit as mentioned above (10.5.1).

The buffer for the final experiment was changed to HBS-P+ containing 5% DMSO. The ligands were measured in concentration series in four times concentration steps up to 10 mM or up to 1 mM in a second experiment. A solvent correction was recorded for eight different DMSO concentrations. The protein remained stable on the chip while it was stored in buffer until it was used for a second experiment. The activity was tested with 50 µM the known binder PUGNAc (Tocris bioscience, batch number 3A/103806). K_D values were determined with the Biacore software. In some cases outliers were excluded for the analysis which is clearly indicated in the results chapters.

10.5.3 The Alternative Biacore Screen with HEWL

HEWL was buffer exchanged to get rid of additional salts. The running buffer (RB) was PBS (10 mM phosphate, 2.7 mM KCl, 137 mM NaCl, pH 7.4; made from 10x stock adding 5% DMSO and 0.05% Surfactant P20). Fragments were screened at 2 mM in 200 µl running buffer with a final concentration of 5% DMSO (2 µl fragment plus 198 µl RB in 4.2% DMSO). The final protein concentration was 70 µM (1 mg/ml). The fragment screen was set-up as an automated procedure with the Biacore software. Screening of one plate takes 16 hours. It is possible to screen 48 fragments (two wells per fragment) on one 96-well plate. The plate was set up to have in the first well fragment alone, and in the second well a mixture of protein and fragment. The third well contained the second fragment and so forth. Chitotriose (570 µM, from an 80 mM stock in Milli-Q) served as a positive control and was injected after each fragment cycle.

In order to prohibit non-specific binding, a solution of 1 mg/ml carboxymethyl dextran sodium salt in 0.15 M NaCl containing 0.02% NaN_2 was tested.

10.5.4 Binding Data with Alternative Biacore Screen

The aim of the screen was to confirm the binding of the first screen and determine affinity data in the form of a dose-response curve. The positive control was chitobiose (stock 200 mM in d_6-DMSO) and the negative control was one of the fragments (ysbl000261)which was shown in the first screen not to bind to the protein. The procedure was implemented using the Biacore software. Every sixth injection was a wash step with 5 mM NaOH in running buffer with DMSO (as mentioned above). Every cycle (i.e. six wells) was repeated once. The 96-well plate was designed using one row per fragment.

Each fragment was tested at six different concentrations from 0 to 10 mM in running buffer with 5% DMSO on its own and then in the same concentration together with HEWL. After every concentration series, the chip was washed with 5 mM NaOH. The screening took ten hours.

10.5.5 Plotting and Analysis

The Biacore T100 system has the ability to output tables with three points per injection: baseline, binding and stability. There are further absolute and relative responses measured in response units (RU). For the analysis, only relative responses were used.

For all the experiments where cycles contained more than two injections (the initial tests and the affinity screen), protein buildup on the chip needed to be corrected. Therefore the values were further corrected using the relationship: $RU_{corrected} = RU_{binding} - RU_{stability}$. For the fragment screen $RU = RU_{binding}$ was used.

For logarithmic plotting purposes, the blank concentration 0 mM was assigned with a value of 0.01 mM. The plots were created with Matlab.

10.6 IC$_{50}$ Determination

Plots were created with a script in Matlab. Where possible the data were fitted with a four-parameter logistic function to extract the IC$_{50}$:

$$y = min + \frac{max - min}{(1 + \frac{x}{IC_{50}})^a} \tag{10.1}$$

min and *max* are the asymptotes which are forced by the fitting script to be within ±2% of the experimentally determined values. a is the Hill coefficient which describes the slope of the curve at its midpoint. The IC_{50} is the corresponding x-coordinate of the inflection point. The function assumes symmetry around IC_{50}.

10.7 Enzyme Activity Assay for HEWL

HEWL lyses cell walls of bacteria. *Micrococcus luteus* (Sigma Aldrich) is especially susceptible to lysozyme and is used as a substrate. When the protein lyses the cells, the solution becomes more transparent. The IC$_{50}$ of a ligand can be determined by following changes in the turbidity of the solution. The change of turbidity was measured on a plate reader (Hidex Plate Chameleon) in 96-well format at a wavelength of 450 nm. Glucose, chitobiose and chitotriose served as controls and 14 fragments in the "Michele_2" library were tested. The final reaction volume was 300 µl. The working concentration of lysozyme was in the range of 2.3 µM (33 µg/ml)–4.6 µM (66 µg/ml) and 0.33–0.5 mg/ml of *M. luteus*. The stock of cells was prepared freshly before every experiment. The experiment was performed in phosphate buffer (50 mM phosphate pH 4.7,150 mM NaCl). Concentration series were tested in six wells in 1/4 dilution steps of the ligand, where the lowest concentration was zero and the highest concentration ten times the expected K_D. The reaction started by dispensing lysozyme into the reaction mixture. The plate was shaken for six seconds at 3 mm amplitude. Ideally, the initial concentration gave an absorbance of 0.8 and the rate should be between 0.015 and 0.040 ΔA_{450}/minute to give non-noisy results. For most of the experiments, the absorbance was recorded for 200–300 counts (approximately 1 count/minute). The slope from the first 30 seconds was excluded in the analysis to guarantee a linear decrease in the absorbance to determine the slope. The slope of the linear part was determined in Excel. The negative slope was multiplied by -100,000 and plotted against the concentration to obtain the dose-response curves.

10.8 Thermal Shift Assay

With the thermal shift method (TSA), the unfolding curve of a protein can be recorded following the signal of a fluorescent dye. When the protein unfolds and hydrophobic groups are exposed, the fluorescence of the dye which was before quenched in the aqueous environment increases. 96-well plates were heated in 1℃ steps from 25 to 95℃. The analysis was performed using the software MTSA, written in Chapter 4 on page 85. The programme was devised to enable fitting of an unlimited number of curves automatically and to output error statistics as well as the thermal melting point.

10.8.1 HEWL

For the sensitivity test, a master mixture for a final concentration of 50 mM citric acid pH 3.8, 150 mM NaCl and 1x SYPRO orange (Sigma Aldrich) was prepared. The mixture was aliquoted onto a master plate. 500 µg/ml HEWL were added to the first column to a final volume of 30 µl per well. The first column was diluted by half for each of the following columns to obtain octuplicates for 12 concentrations ranging from 500 µg/ml to 0.24 µg/ml. Three aliquots were taken from the master plate to test identical plates on three different qPCR machines.

HEWL was further tested against the hits discovered in the SPR screen mentioned in Section 10.5.3. 200 µg/ml HEWL was screened in 50 mM citric acid buffer pH 3.8, 150 mM NaCl, 1x SYPRO orange in 40 µl solution per well. The final DMSO concentration in the wells was 2.5%. All screening experiments with HEWL were performed as quadruplicates if not otherwise mentioned. The fragments were tested at 0 mM, 500 µM and 5 mM with 32 blank samples (protein in buffer without ligand) per plate.

10.8.2 BtGH84

BtGH84 was screened for ligand binding against the fragment libraries "Michele_1" and "Michele_2" as well as against some nearest neighbour compounds of the library "Jens". The compounds were screened as quadruplicates at 1 mM and 10 mM in 96-well plates. Each plate contained 8–32 blanks. The screen was carried out with 50 mM sodiumphosphate buffer pH 6.0, 150 mM NaCl, 1x SYPRO orange and 1–2 µM protein. The screens were performed on the Agilent Stratagene qPCR machine. SYPRO orange was used at 1x concentration, Deep Purple (GE Healthcare) at 6x. The final volume per well was 40 µl with containing 5% DMSO. ΔT_m is the difference of the T_m of the sample to that of the averaged blanks.

10.9 Crystallisation

Protein crystallisation is one of the major techniques to determine the structure of proteins. With the reinforced signal of repeating protein units, a three dimensional diffraction pattern can be recorded with which help the three dimensional structure can be calculated. To evolve fragments, it is extremely important to know where and how the compound is binding. If the protein is suitable for crystallisation, the binding site is exposed to solvent channels in the crystal lattice, and if a native structure is already known, the method molecular replacement may generate a solution to the phase problem, after which point positive difference density may be evident for the bound fragment.

10.9.1 HEWL

Known crystallisation conditions for HEWL were reproduced. Sitting drop trays with different conditions were set up: One 96-well plate (tray 1) was prepared with 100 mM sodium acetate trihydrate, pH 4.5–6.5 with drops of 150 nl mother liquor and 150 nl protein solution, and 300 nl of each respectively. Conditions were adopted from the "Lysozyme Kit" by Hampton Research. One 96-well plate each with the CSS I& II screen (Clear Strategy Screen, Molecular Dimensions, CSS II was developed at York; Brzozowksi and Walton, 2001; tray 2) and one with the PGA screen (Poly-γ-Glutamic Acid polymer; Molecular Dimensions; Hu et al., 2008; tray 3) with the buffers tris pH 8.0 and HEPES pH 7.5 with 55 µl per well were prepared. Drops of 250 nl mother liquor mixed with 250 nl protein were added. As follow-up plates, one 48-well (tray 4) and one 96-well plate (tray 5) with 2 M sodium formate, 100 mM HEPES pH 7.5 and 0.8 M sodium formate, 10% PEG 8K, 10% PEG 1K were set up. The 48-well plate had a reservoir volume of 500 µl and drop volume of 1 µl mother liquor and 1 µl protein. The 96-well plate had a reservoir volume of 50 µl and drops of 250 nl mother liquor mixed with 250 nl protein. For each condition, drops were set with 10 mg/ml and with 30 mg/ml HEWL concentration. The nano litre drops were set with the Mosquito robot.

Fragments were soaked into wells of tray 1. The fragment solution in 500 mM DMSO was prediluted in distilled H_2O or mother liquor in ratios 1:1, 1:2, 1:3 and 1:4, and 0.2–0.5 µl of the diluted fragments were added to the protein crystals. In some cases, the crystal drop was enlarged by 1 µl of mother liquor to dilute the ligand solution further. Assuming that the original drop size was 300 or 600 nl, these soaking conditions resulted in final concentrations of 10–200 mM ligand

and 3–30% DMSO.

10.9.2 BtGH84

48-well plates with 100 µl mother liquor (100 mM imidazole pH 8.0, 10% PEG 8K, 3% 2 M TMAO with protein > 10 mg/ml and with 10, 12.5 and 15% ethyleneglycol as cryoprotectant) as sitting drop were prepared (personnel communication, Dr. Yuan He, 2011). The protein to mother liquor ratio was 1 µml : 1 µl and 0.75 µl : 1 µl. The protein concentration used was more than 10 mg/ml and crystals grew overnight. Ligands were soaked into the crystals with an estimated final concentration of 20 mM and 10% DMSO.

10.10 Structure Solution

BtGH84 crystals soaked with the fragments ysbl000298, ysbl000299, ysbl000314, ysbl000317, ysbl000370, ysbl000509, ysbl000545 and ysbl000548 were tested in-house and then sent to Diamond for data collection. BtGH84 data sets were processed with Mosflm (Leslie and Powell, 2007) and Xia2 (Winter, 2009). BtGH84 datasets were solved by molecular replacement with Balbes (Long et al., 2008). The solved structures were refined with Refmac5 (Vagin et al., 2004) and rebuilt in Coot (Emsley and Cowtan, 2004).

Appendix A

SMARTS Strings

The SMARTS strings listed below were implemented by Dr. Kerrin Bright in course of the collaboration of the fragment library design.

Table A.1: **SMARTS strings.** List of SMARTS strings used to define compounds from the Input Libraries with unwanted functionality, a brief description of the chemical feature and the number of compounds from each supplier that contained each feature.

SMARTS string	Functionality	Asinex	Maybridge	Specs
[OX2H]c1ccccc1[OX2H]	Catechol	417	31	209
[CX4](−[OX2])(−[OX2])(−[OX2])	Ortho ester	0	2	33
[$([NX3](= O) = o), $([NX3+](= O)[O−])][!#8]	Nitro	34477	6529	23314
O − O	Ether	0	0	3
[NX2] = [OX1]	Nitroso	155	73	63
[CX3] = ([NX2] − OH)	Oxime	1066	688	551
[CX3] − [CX3](= O) − [CX3]	Aliphatic ketone	5660	1125	2681
[CX4](−[OX2])(−[OX2])	Acetal	19437	1468	7065
c1nsnc1	Thiadiazole	876	320	180
C = C[H2]	Methylene	82066	4168	37903
[#6][C!H0] = O	Aldehyde	1523	7	1210
c1(NH2)cccs1	Aminothiophene	8013	677	13033
C1CN1	Aziridine	42	10	89
NC(= S)N	Thiourea	18077	4433	11669
c1ncns1	Thiadiazole	271	118	91
[NX3, NX4+][CX3](= [OX1])				

Continued on next page

Table A.1 – continued from previous page

SMARTS string	Functionality	Asinex	Maybridge	Specs
[SX2, SX1−]	Thiolactone	1576	96	2610
[CX4](−[SX2])(−[SX2])	Dithioacetal	58	308	503
[CX3; !R](= [SX1])[NH2]	Thioamide	465	384	323
c1(NH2)ccsc1	Aminothiophene	2133	916	2051
S − H	Thiol	2328	595	1590
[NX3+]	Tertiary amine	39219	8190	24954
N = C = O	N=C=O	1	0	1
S(= O)(= O)[OX2]	Sulphoxide	570	543	1153
c2ccc1nonc1c2	Benzoxadiazole	248	264	116
N = C = S	N=C=S	11	0	9
c1scnn1	Thiadiazole	10349	535	5249
[CX4](−[NX3])(−[OX2])	Aminal	6032	648	5037
[$([CX3]([#6])[#6]), $([CX3H][#6])] = [$([NX2][#6]), $([NX2H])]	Imine	19678	2133	13930
S − S	Thioether	565	49	410
[CX3] = [CX3][Cl, Br, F, I]	Alpiphatic C=C-Halogen	2032	389	1533
C(= O)[F, Cl, Br, I]	Acyl halide	37	0	21
O = C1CSCO1	Oxathiolane	2	0	6
[C; !R](= O) − [S; !R]	acyclic C(=O)-S	105	320	138
C1CO1	Epoxide	234	0	175
N − [NH2]	Hydrazide	1074	690	652
[CH2, H3; !r][CH2!r][CH2!r][CH2, H3; !r]	Aliphatic chain	11413	847	11047
[NX4+]	Quarternary amine	86993	4647	19080
[#6]C(= O)OC(= O)[#6]	Anhydride	0	0	32
O = C − C = [A; !O; !N; !S; !a]	R=C-C=O	139968	10466	69559
[*; !#1; !#6; !#7; !#8; !#9; !#16; !#17]	Not C,N,F,Cl,S	27118	3042	40808
[N; !R] = [C; !R] = [N; !R]	Acyclic N=C=N	1	4	1
[NX3, NX4+][CX3](= [OX1])[OX2, OX1−]	N-C-O acetal	4065	2229	2472
C − [CX3](= O)[CX4, CX3, CX2][Cl, Br, F, I]	Halo-acetophenone	184	109	376
[C; !R](= S) − [O; !R]	Acyclic C(=S)-O	41	62	51

Appendix B

Physicochemical Properties of All Fragment Sets

The tables in this appendix give a detailed overview about the physicochemical properties of the Fragment Libraries of the suppliers Asinex, Maybridge and Specs. Also all physicochemical properties of all 18 created Fragment Sets (six different procedures times three different suppliers) are listed.

Table B.1: **Physicochemical Profile of the Fragment Sets - Part 1.** Properties of the fragment set, generated by the different protocols for the three Fragment Libraries. MW: Molecular weight, AC: Number of heavy atoms (non hydrogen), FC: Formal charge, ALogP, HA: Number of hydrogen bond acceptors.

Library	**MW**	**AC**	**FC**	**ALogP**	**HA**
Asinex	216.5±33.35	15.3±2.30	-0.23±0.47	1.11±1.03	3.1 ±0.99
Cluster All	209.6±30.02	14.8±2.24	-0.07±0.29	1.38±0.93	2.6±1.00
Cluster Fragments	201.8±30.87	14.3±2.36	-0.09±0.38	1.24±0.92	2.7±1.10
SIM within Cluster	218.6±30.61	15.4±2.28	-0.09±0.37	1.21±0.99	3.0±1.06
Substructure Count	228.0±25.08	16.8±1.50	-0.05±0.38	1.79±0.63	2.7±0.99
Substructure Map	241.8±23.60	17.6±1.27	-0.03±0.25	1.91±0.73	2.6±0.90
Iterative Removal	196.7±30.87	13.8±2.24	-0.09±0.28	1.46±0.94	2.3±0.93
Maybridge	203.4±30.71	14.3±2.17	-0.18±0.42	1.36±0.89	2.9±1.03
Cluster All	194.1±28.72	13.5±2.16	-0.08±0.39	1.50±0.85	2.6±1.13
Cluster Fragments	191.5±28.58	13.2±2.17	-0.10±0.40	1.42±0.94	2.6±1.07
SIM within Cluster	204.5±33.14	14.2±2.28	-0.10±0.44	1.60±0.81	2.8±1.07
Substructure Count	212.2±29.58	15.7±1.95	-0.08±0.44	1.80±0.71	2.5±1.03
Substructure Map	232.3±31.34	16.8±1.88	-0.04±0.26	1.78±0.67	2.5±0.96
Iterative Removal	189.2±26.90	13.1±1.97	-0.09±0.32	1.52±0.72	2.6±0.91
Specs	205.4±37.69	14.5±2.63	-0.11±0.42	1.35±0.94	2.8±1.03
Cluster All	180.1±37.56	12.7±2.87	0.01±0.40	1.32±0.91	2.5±1.18
Cluster Fragments	179.5±36.87	12.7±2.80	0.01±0.38	1.33±0.86	2.5±1.12
SIM within Cluster	201.9±38.42	14.3±2.79	-0.04±0.43	1.48±0.92	2.8±1.18
Substructure Count	220.7±28.24	16.3±1.74	0.00±0.40	1.95±0.62	2.7±1.07
Substructure Map	237.7±26.34	17.3±1.46	-0.01±0.22	2.10±0.65	2.6±0.93
Iterative Removal	174.9±31.58	12.2±2.23	-0.11±0.32	1.38±0.96	2.1±0.85

Table B.2: **Physicochemical Profile of the Fragment Sets of Specs - Part 2.** HD: Number of hydrogen bond donors, RB: Number of rotatable bonds, PSA: Polar surface area, LogS: Solubility, ArB: Number of aromatic bonds.

Library	**HD**	**RB**	**PSA**	**LogS**	**ArB**
Asinex	0.8±0.76	2.42±1.30	57.9±14.35	-2.30±0.76	7.09±3.57
Cluster All	0.7±0.71	2.41±1.29	53.2±16.73	-2.37±0.81	6.98±3.22
Cluster Fragments	0.7±0.72	2.14±1.18	53.3±16.31	-2.26±0.88	6.98±3.58
SIM within Cluster	0.7±0.68	2.49±1.31	57.3±15.76	-2.35±0.85	7.46±4.06
Substructure Count	0.9±0.75	2.43±1.11	51.7±16.13	-3.03±0.25	12.54±1.99
Substructure Map	0.6±0.85	1.68±0.86	48.3±15.94	-2.61±0.57	8.47±4.08
Iterative Removal	0.6±0.70	2.08±1.35	47.3±16.61	-2.39±0.73	6.71±3.15
Maybridge	0.7±0.75	1.94±1.27	56.6±14.78	-2.34±0.69	7.10±3.50
Cluster All	0.7±0.74	1.74±1.16	52.8±16.16	-2.37±0.76	7.17±3.60
Cluster Fragments	0.6±0.74	1.76±1.24	54.2±16.75	-2.32±0.68	6.90±3.89
SIM within Cluster	0.7±0.79	1.82±1.27	55.8±14.77	-2.40±0.78	7.41±3.98
Substructure Count	0.8±0.77	1.87±1.00	50.3±16.51	-2.84±0.35	12.30±2.02
Substructure Map	0.7±0.79	1.58±0.76	49.6±16.86	-2.75±0.52	8.94±4.28
Iterative Removal	0.5±0.68	1.76±1.17	50.0±14.61	-2.31±0.60	7.27±2.54
Specs	0.8±0.71	2.18±1.45	54.1±15.35	-2.33±0.74	6.90±3.31
Cluster All	0.7±0.69	1.44±1.30	49.8±17.84	-2.12±0.86	6.77±4.13
Cluster Fragments	0.7±0.71	1.59±1.35	48.9±17.60	-2.11±0.84	6.54±3.92
SIM within Cluster	0.7±0.67	1.90±1.43	52.3±16.13	-2.30±0.88	7.51±3.98
Substructure Count	0.8±0.78	2.21±1.11	49.3±16.67	-3.01±0.26	12.73±2.03
Substructure Map	0.7±0.81	1.77±0.93	47.2±17.05	-2.79±0.44	8.76±3.92
Iterative Removal	0.4±0.62	1.54±1.26	42.7±16.13	-2.17±0.77	6.02±2.98

Table B.3: **Physicochemical Profile of the Fragment Sets of Specs - Part 3.** R: Number of rings, ArR: Number of aromatic rings, RA: Number of ring assemblies.

Library	**R**	**ArR**	**RA**
Asinex	1.9±0.63	1.3±0.69	1.5±0.52
Cluster All	1.8±0.69	1.2±0.60	1.5±0.54
Cluster Fragments	1.8±0.76	1.2±0.69	1.5±0.55
SIM within Cluster	2.0±0.70	1.4±0.76	1.6±0.58
Substructure Count	2.4±0.53	2.2±0.43	2.1±0.35
Substructure Map	3.0±0.52	1.5±0.74	2.2±0.43
Iterative Removal	1.6±0.64	1.2±0.60	1.4±0.54
Maybridge	1.8±0.73	1.3±0.67	1.4±0.55
Cluster All	1.7±0.74	1.3±0.67	1.4±0.55
Cluster Fragments	1.6±0.76	1.2±0.73	1.3±0.57
SIM within Cluster	1.9±0.73	1.3±0.75	1.4±0.56
Substructure Count	2.4±0.50	2.2±0.42	2.0±0.44
Substructure Map	2.8±0.78	1.6±0.77	2.1±0.48
Iterative Removal	1.6±0.61	1.3±0.48	1.4±0.49
Specs	1.8±0.71	1.2±0.63	1.4±0.52
Cluster All	1.7±0.78	1.2±0.77	1.3±0.51
Cluster Fragments	1.6±0.74	1.2±0.72	1.3±0.51
SIM within Cluster	1.9±0.73	1.4±0.75	1.4±0.55
Substructure Count	2.4±0.52	2.2±0.43	2.0±0.33
Substructure Map	2.9±0.68	1.5±0.71	2.1±0.44
Iterative Removal	1.5±0.70	1.0±0.54	1.2±0.45

Appendix C

Detailed Fragment Library Generation

C.1 Library Generation – First Iteration

The first iteration to set up a new fragment library was performed with a version of the *Iterative Removal* protocol that does not contain the *Diverse Molecules* component, but runs on the Pipeline Pilot Student Edition (Figure C.1).

After the separation in Fragments and Non-Fragments (Figure 3.1, page 63), both Libraries are compared and the number of nearest neighbours is calculated for each fragment (Figure C.1 2). The fragment with the highest number of nearest neighbours is selected and written to a separate file. The reference compounds to that fragment are determined. Subsequently the selected fragment is removed from the Fragments list and the nearest neighbours are removed from the Non-Fragment Library (Figure C.1 3). The loop executes as many times as members required for the Fragment Set. In the final step, all selected fragments are merged to a Final Fragment Set (Figure C.1 4).

500 fragments from Specs were selected. The compounds were visually inspected with the help of Dr. Gideon Grogan and Dr. Kerrin Bright to discard reactive compounds which unintentionally passed the *unwanted groups* filter (Appendix A). 201 fragments with the highest score of nearest neighbours were purchased, and dissolved at a concentration of 200 mM in DMSO-d_6.

95% of the compounds were soluble which confirms the adequate estimation of the solubility threshold. The library was named "Michele_1".

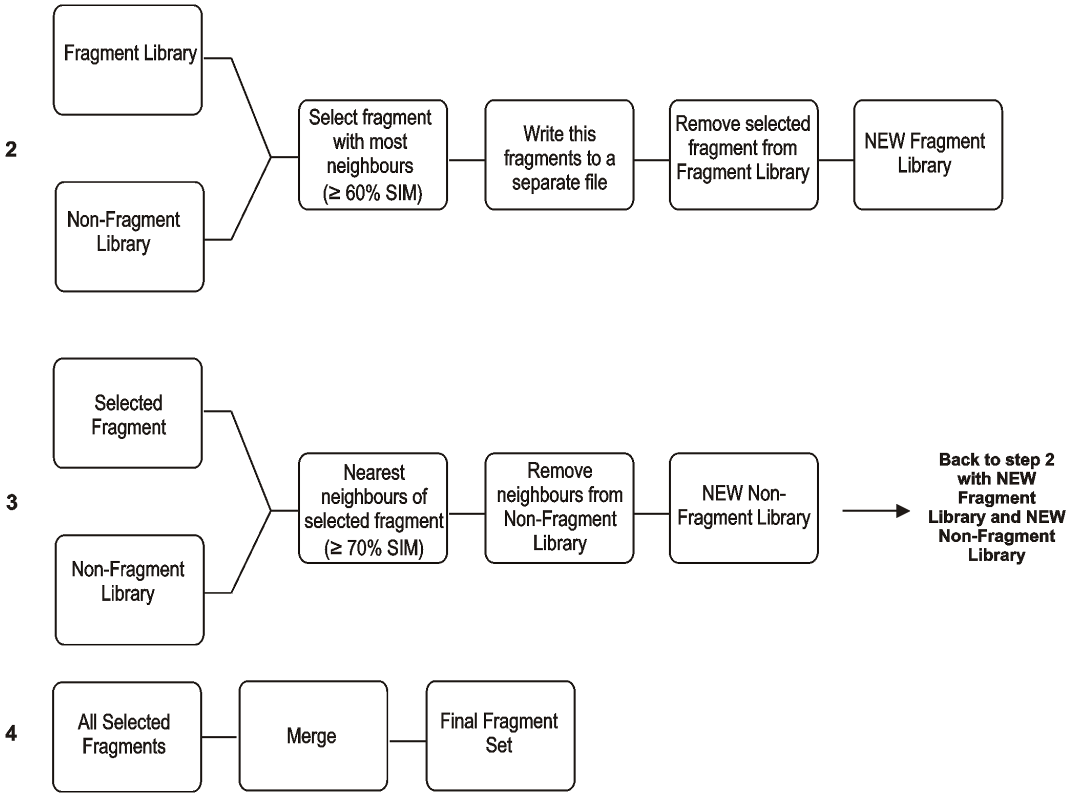

2
Fragment Library
Non-Fragment Library
Select fragment with most neighbours (≥ 60% SIM)
Write this fragments to a separate file
Remove selected fragment from Fragment Library
NEW Fragment Library
3
Selected Fragment
Non-Fragment Library
Nearest neighbours of selected fragment (≥ 70% SIM)
Remove neighbours from Non-Fragment Library
NEW Non-Fragment Library
Back to step 2 with NEW Fragment Library and NEW Non-Fragment Library
4
All Selected Fragments
Merge
Final Fragment Set

Figure C.1: **Iterative Removal – Student Version.**

C.2 Library Generation – Second Iteration

After the generation of Michele_1, a library extending Fragment Set ("Michele_2") was selected. Four different suppliers were chosen: Asinex, Maybridge, Sigma Aldrich and Specs, which offered more than one million compounds. The protocol version of the *Iterative Removal* on page 70 was used.

200 fragments were selected. The aim was breaking this number down to 100 compounds physically available at 200 mM stock solutions. Two programmes of the JChem programme suite (ChemAxon) were used for the following steps: LibraryMCS and InstantJChem. LibraryMCS (Maximum Common Substructure) clusters structures based on common substructures and outputs cluster trees. InstantJChem is a database for chemical compounds.

All duplicate compounds (about 50) already contained in Michele_1 were removed. For the remaining compounds, cluster trees were generated with LibraryMCS. Two compounds were deleted because they had a close neighbour in the generated trees. Within InstantJChem, the overlap component was used to run a superstructure search which lead to the removal of further 27 compounds (the superstructures). A Tanimoto similarity search was performed with $\geq$ 80% SIM. Four compounds were removed. The remaining library contained 108 compounds. Two further fragments were deleted because one was a superstructure of compounds in Michele_1 and the other one had no protons, which is unsuitable for NMR experiments. 97 compounds from the remaining list were available from the above named suppliers. Three compounds were too volatile or not soluble at 200 mM in DMSO, which reduced Michele_2 to 94 compounds.

C.3 Purchasing Compounds and Quality Control

NMR spectra for Michele_2 were recorded. The analysis of the spectra was performed with the help of Dr. Jens Landström. Eight compounds showed an incorrect spectrum (ysbl000468, ysbl000474, ysbl000492, ysbl000506, ysbl000507, ysbl000525, ysbl000527, ysbl000539). Further eight spectra showed that the compounds were not fully soluble (ysbl000467, ysbl000490, ysbl000496, ysbl000497, ysbl000501, ysbl000506, ysbl000529, ysbl000540).

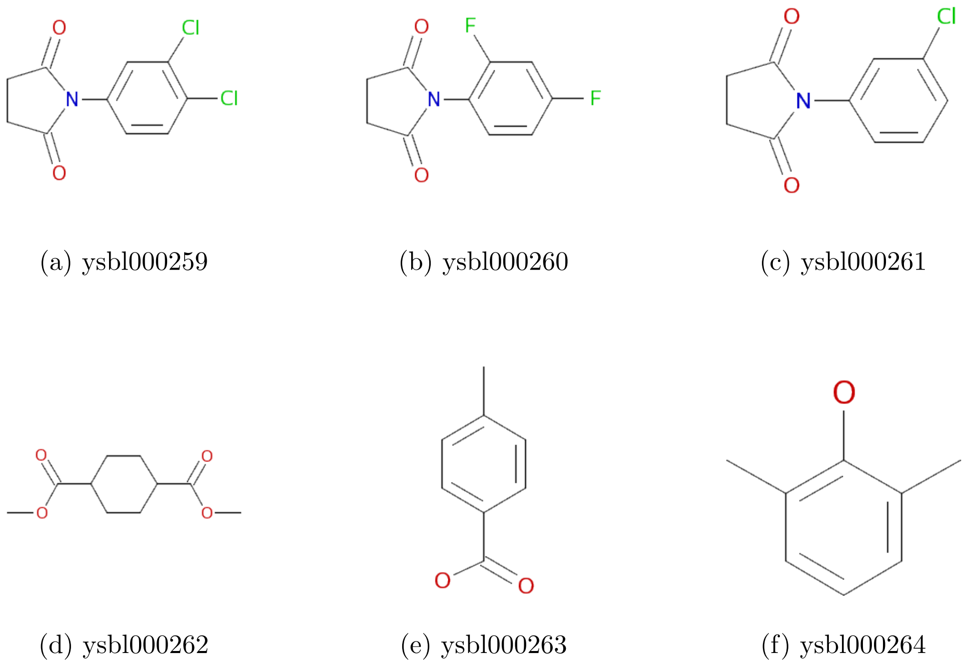
C.4 Creating Database

A database for all available compounds in YSBL was created in InstantJChem (ChemAxon). The new “YSBL Library” contains the fragment sets Michele_1 and Michele_2, “Yasu” (Dr. Yasuhiko Kanda), “Kerrin” (Dr. Kerrin Bright) and “Jens” (Dr. Jens Landström). All 805 compounds were sorted and numbered with a consecutive ysbl number to facilitate the fragment screening for future colleagues in YSBL.

Appendix D

YSBL Compounds

All in this thesis mentioned YSBL compounds are shown with their structure below:



(a) ysbl000259

(b) ysbl000260

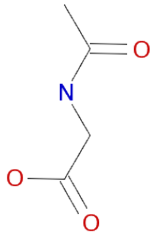
(c) ysbl000261

(d) ysbl000262

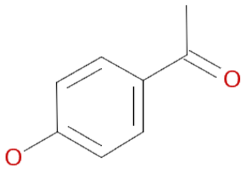
(e) ysbl000263

(f) ysbl000264

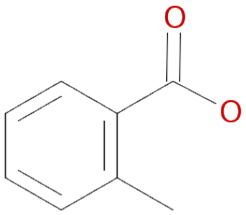
Figure D.1: **Ysbl Compounds - Part 1.**



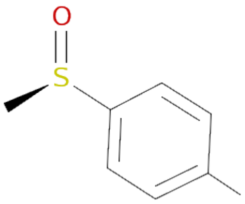
(a) ysbl000265



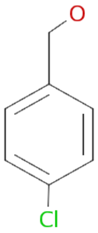
(b) ysbl000266



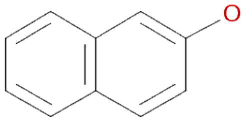
(c) ysbl000267



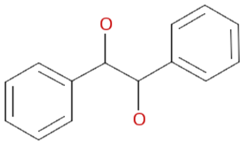
(d) ysbl000268



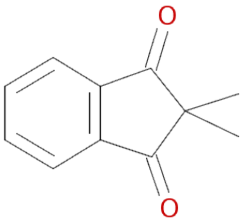
(e) ysbl000269



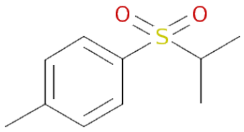
(f) ysbl000270



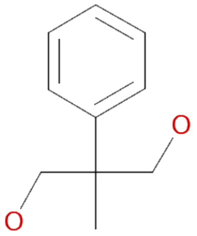
(g) ysbl000271



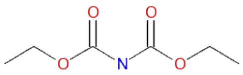
(h) ysbl000272



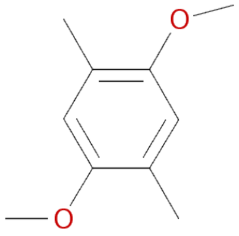
(i) ysbl000273



(j) ysbl000274

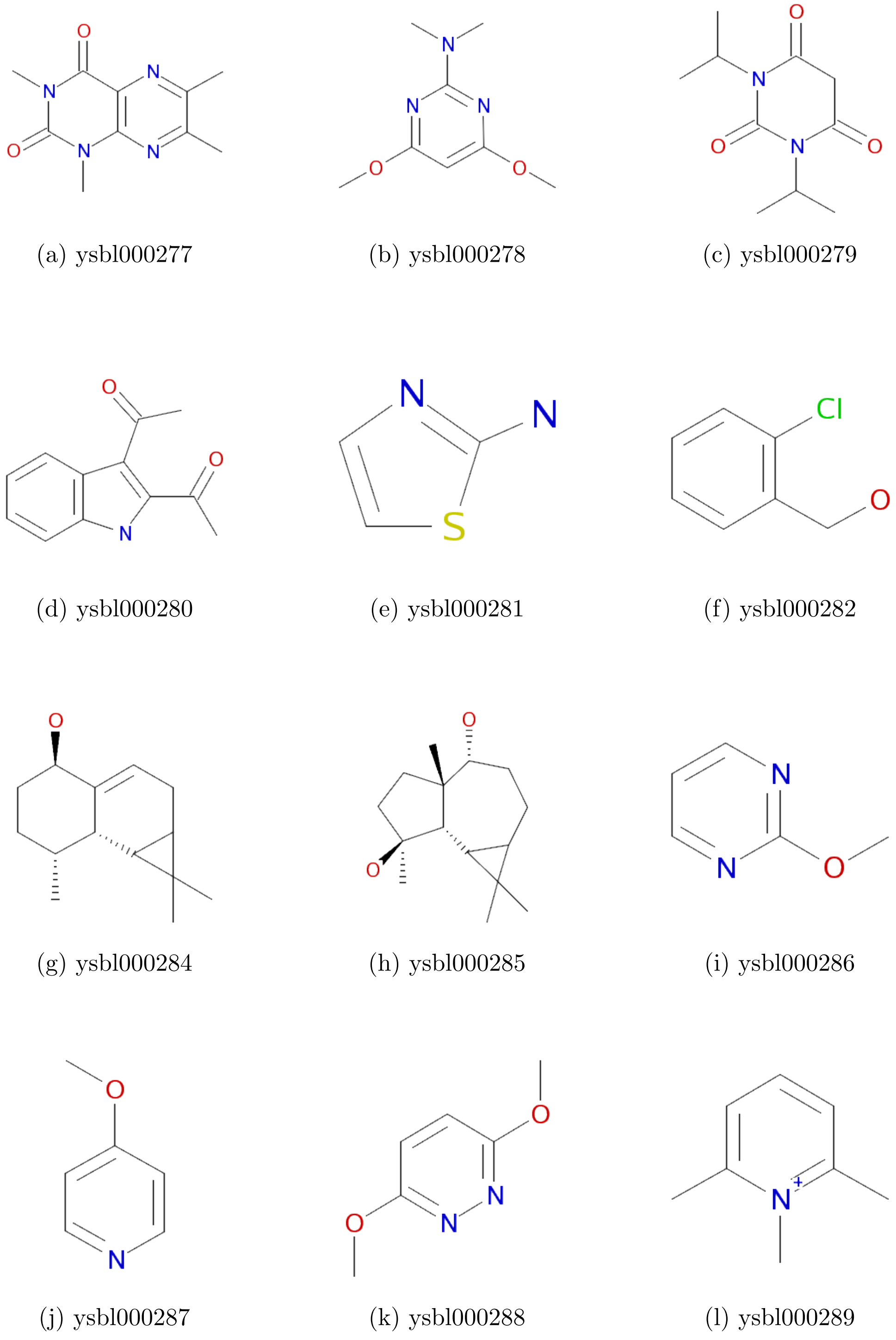


(k) ysbl000275



(l) ysbl000276

Figure D.2: **Ysbl Compounds - Part 2.**



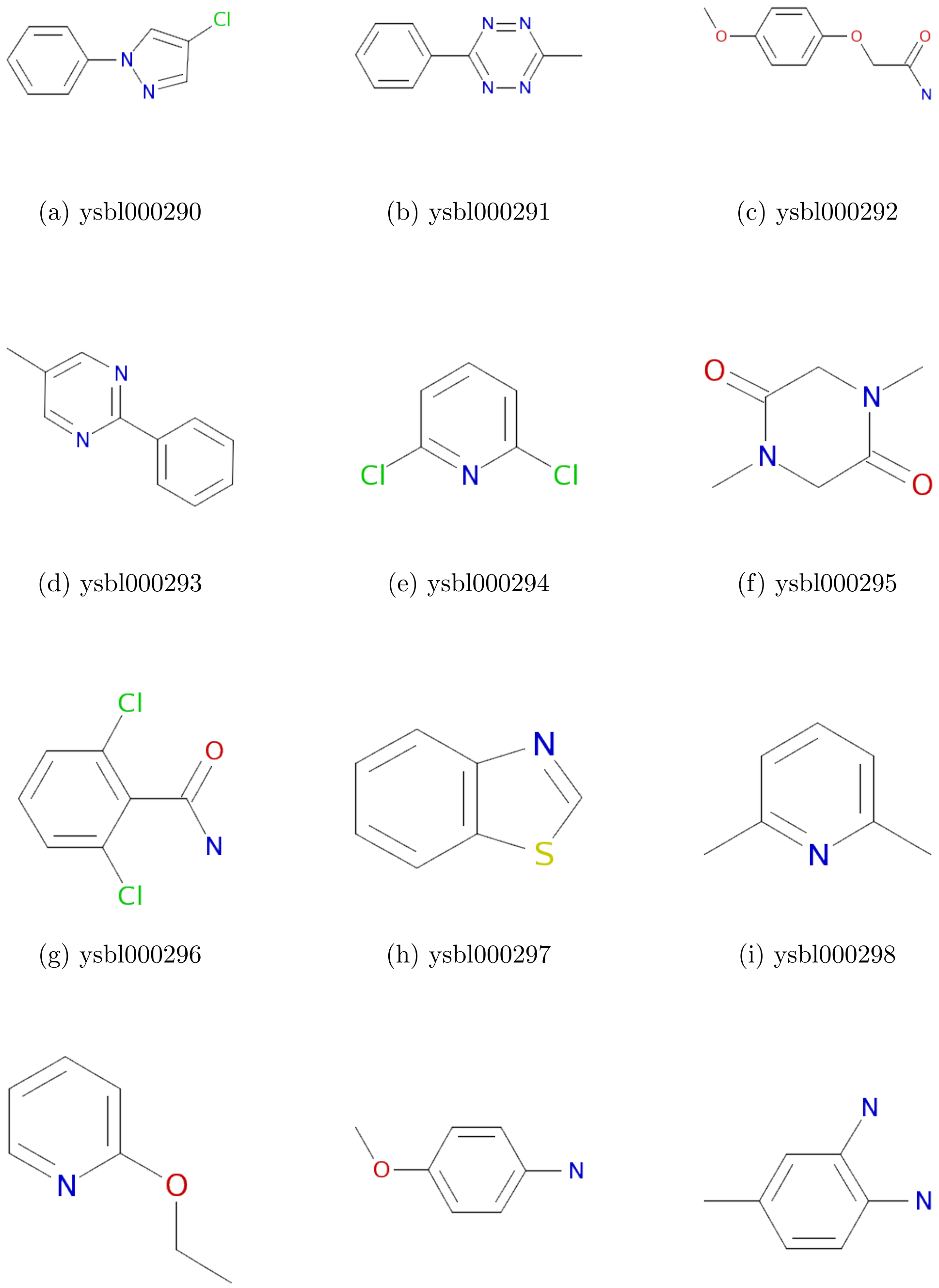
(a) ysbl000277 (b) ysbl000278 (c) ysbl000279

(d) ysbl000280 (e) ysbl000281 (f) ysbl000282

(g) ysbl000284 (h) ysbl000285 (i) ysbl000286

(j) ysbl000287 (k) ysbl000288 (l) ysbl000289

Figure D.3: **Ysbl Compounds - Part 3.**



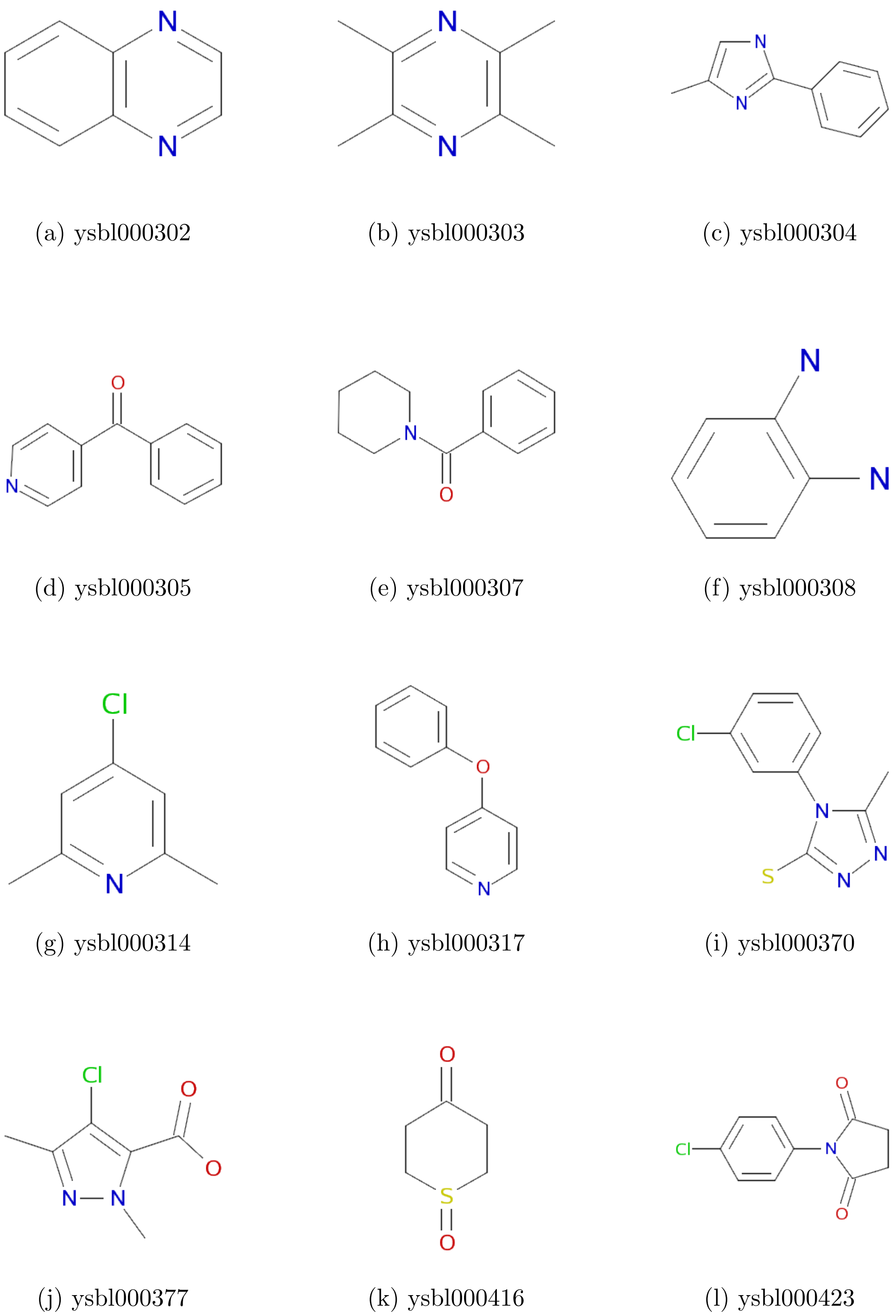
(a) ysbl000290 (b) ysbl000291 (c) ysbl000292

(d) ysbl000293 (e) ysbl000294 (f) ysbl000295

(g) ysbl000296 (h) ysbl000297 (i) ysbl000298

(j) ysbl000299 (k) ysbl000300 (l) ysbl000301

Figure D.4: **Ysbl Compounds - Part 4.**



(a) ysbl000302

(b) ysbl000303

(c) ysbl000304

(d) ysbl000305

(e) ysbl000307

(f) ysbl000308

(g) ysbl000314

(h) ysbl000317

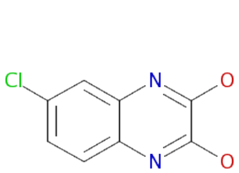
(i) ysbl000370

(j) ysbl000377

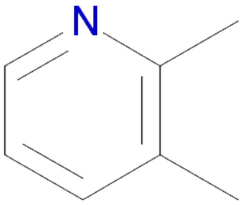
(k) ysbl000416

(l) ysbl000423

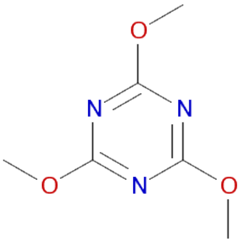
Figure D.5: **Ysbl Compounds - Part 5.**



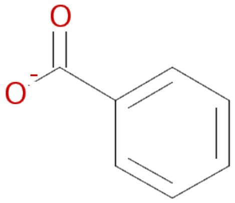
(a) ysbl000438



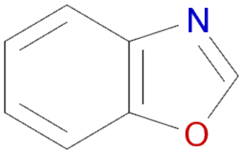
(b) ysbl000461



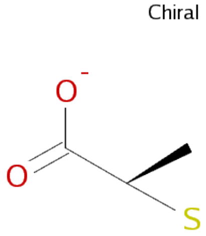
(c) ysbl000464



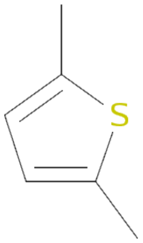
(d) ysbl000469



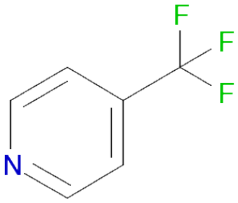
(e) ysbl000470



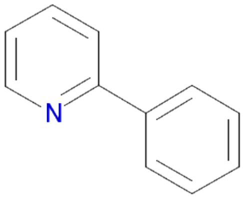
(f) ysbl000471



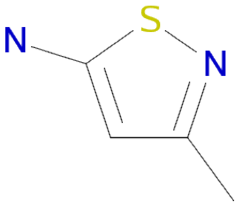
(g) ysbl000473



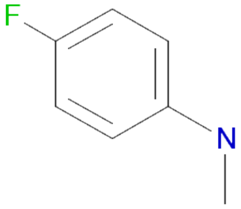
(h) ysbl000477



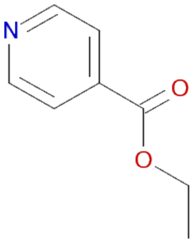
(i) ysbl000481



(j) ysbl000483

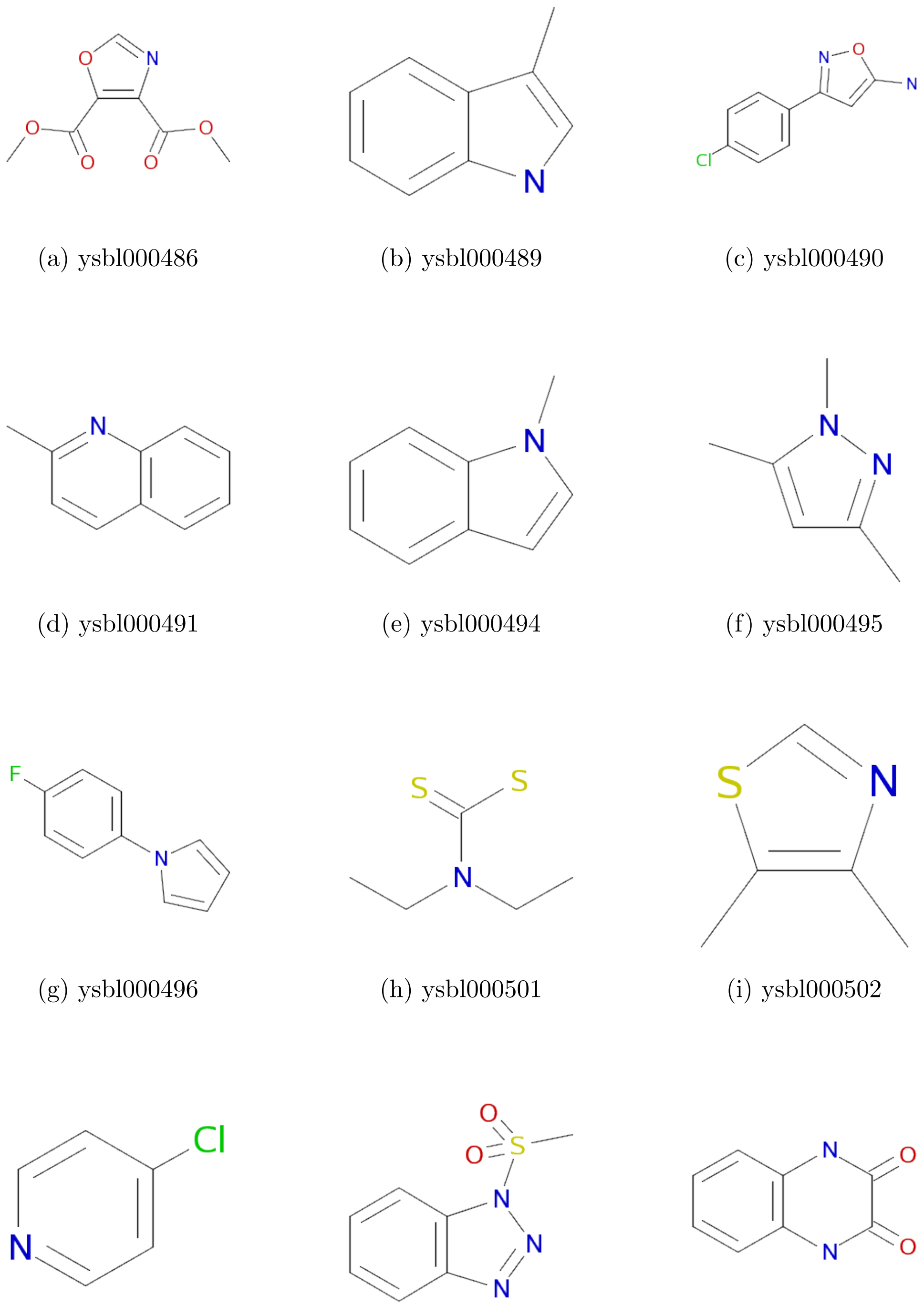


(k) ysbl000484



(l) ysbl000485

Figure D.6: **Ysbl Compounds - Part 6.**



(a) ysbl000486

(b) ysbl000489

(c) ysbl000490

(d) ysbl000491

(e) ysbl000494

(f) ysbl000495

(g) ysbl000496

(h) ysbl000501

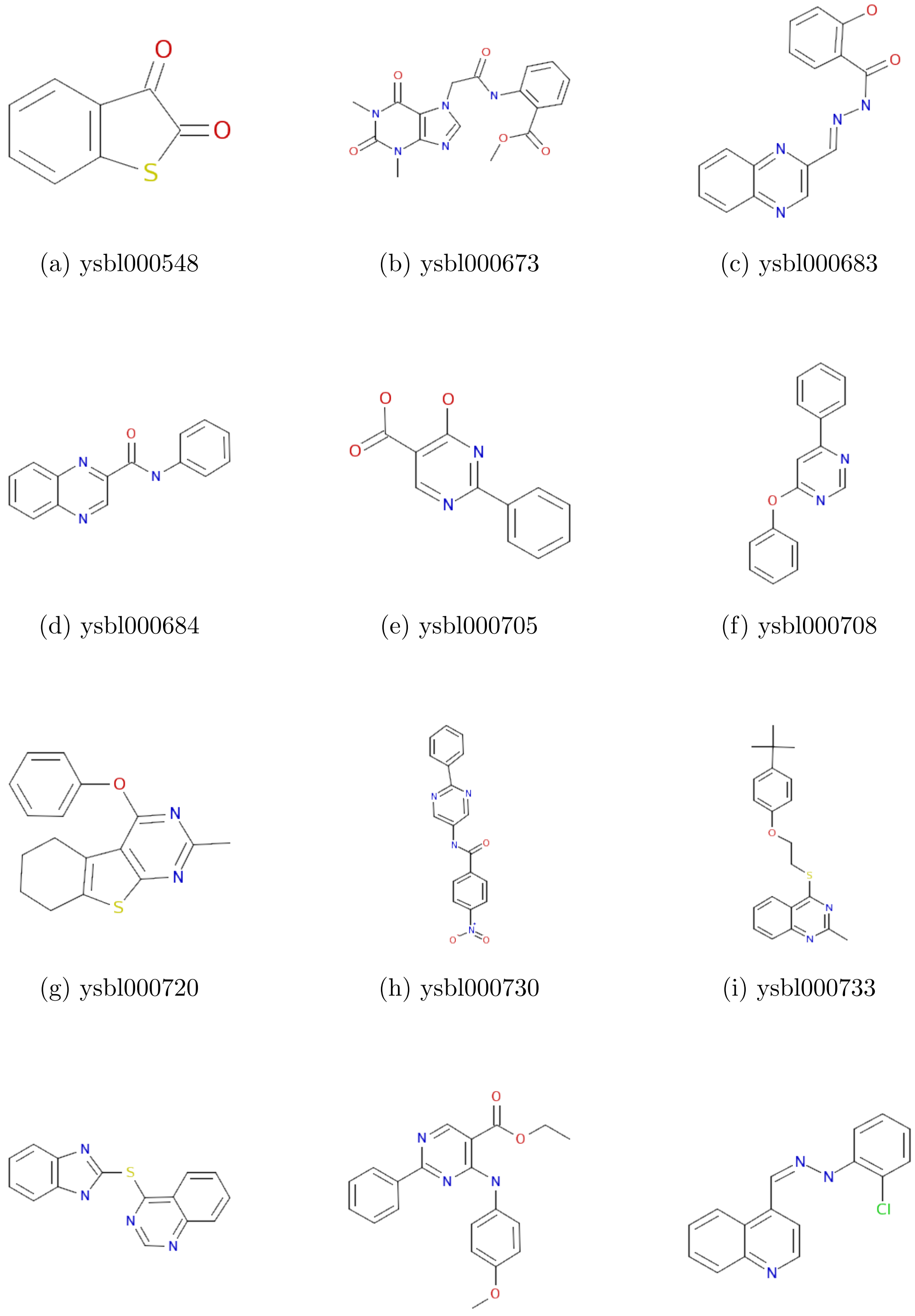
(i) ysbl000502

(j) ysbl000507

(k) ysbl000509

(l) ysbl000545

Figure D.7: **Ysbl Compounds - Part 7.**



(a) ysbl000548

(b) ysbl000673

(c) ysbl000683

(d) ysbl000684

(e) ysbl000705

(f) ysbl000708

(g) ysbl000720

(h) ysbl000730

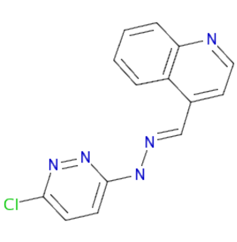
(i) ysbl000733

(j) ysbl000734

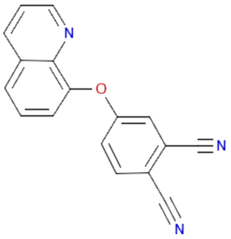
(k) ysbl000737

(l) ysbl000749

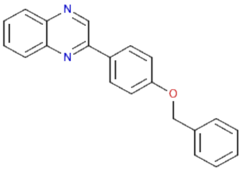
Figure D.8: **Ysbl Compounds - Part 8.**



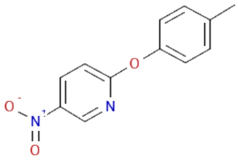
(a) ysbl000752



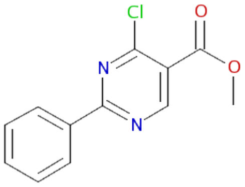
(b) ysbl000767



(c) ysbl000770



(d) ysbl000775



(e) ysbl000779

Figure D.9: **Ysbl Compounds - Part 9.**

Definitions

Boltzmann Equation: Four-parameter logistic model

Boltzmann Method: Midpoint and inflection point of Boltzmann equation

$\mathbf{IC_{50}}$ Half maximal inhibitory concentration

Jens Compound library containing nearest neighbours

Inflection Inflection point of Sigmoid-5 equation

$\mathbf{k_a}$ Association rate constant, also $\mathrm{k_1}$

$\mathbf{k_{cat}}$ Product forming rate constant

$\mathbf{k_d}$ Dissociation rate constant, also $\mathrm{k_{-1}}$

$\mathbf{K_D}$ Dissociation constant, $\frac{k_{-1}}{k_1}$ or $\frac{k_d}{k_a}$

$\mathbf{K_i}$ Inhibition constant; competing ligand concentration binding to half of the receptors in absence of a ligand in equilibrium

$\mathbf{K_M}$ Michaelis-Menten constant; substrate concentration where the enzyme activity is half maximal

Kerrin Fragment library

Michele_1 Fragment library

Michele_2 Fragment library

Midpoint Midpoint of Sigmoid-5 equation

Sigmoid-5 Five-parameter logistic model

T_m Melting temperature in a thermal unfolding curve; equal concentration of folded and denatured protein

Yasu Fragment library

Glossary

Asp Aspartic acid

CAPS N-cyclohexyl-3-aminopropanesulfonic acid

CHES N-Cyclohexyl-2-aminoethanesulfonic acid

D_2O Deuterium oxide

DMSO Dimethyl sulfoxide

DMSO-d_6 Deuterated DMSO

DTT Dithiothreitol

EDC 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide

EDTA Ethylenediaminetetraacetic acid

ESI Electrospray ionisation

GlcNAc N-Acetylglucosamine

Glu Glutamic acid

HBS HEPES buffered saline

HBS-P HBS buffer with surfactant P20

HEPES 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

IPTG Isopropyl β-D-1-thiogalactopyranoside

LB Lysogeny broth (medium)

MCoA Myristoyl coenzyme A

MES 2-(N-morpholino)ethanesulfonic acid

MOPS 3-(N-morpholino)propanesulfonic acid

MW Molecular weight

NHS N-hydroxysuccinimide

NMR Nuclear magnetic resonance

NTA Nitrilotriacetic acid

OD Optical density

PCR Polymerase chain reaction

PEG Polyethylene glycol

PIPES Piperazine-N,N-bis(2-ethanesulfonic acid

PUGNAc O-(2-acetoamido-2-deoxy-D-glucopyra-nosylidene)amino-N-phenylcarbamate

QPCR Quantitative Real Time Polymerase Chain Reaction

RU Response units

SDS-PAGE Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SIM Similarity

SPR Surface plasmon resonance

TAE Tris-acetate-EDTA

TSA Thermal shift analysis

TSP Trimethylsilyl propanoic acid

References

ABAD-ZAPATERO, C. & METZ, J. T. 2005. Ligand efficiency indices as guideposts for drug discovery. Drug Discov Today, 10, 464-9.

ACCELRYSSOFTWAREINC. 2007. Pipeline Pilot. In: SCITEGIC (ed.) 6.1.5.0 Student Edition ed. San Diego: Accelrys Software Inc.

AGRAWAL, A., JOHNSON, S. L., JACOBSEN, J. A., MILLER, M. T., CHEN, L. H., PELLECCHIA, M. & COHEN, S. M. 2010. Chelator fragment libraries for targeting metalloproteinases. ChemMedChem, 5, 195-9.

AKRITOPOULOU-ZANZE, I. & HAJDUK, P. J. 2009. Kinase-targeted libraries: the design and synthesis of novel, potent, and selective kinase inhibitors. Drug Discov Today, 14, 291-7.

ALBERT, J. S., BLOMBERG, N., BREEZE, A. L., BROWN, A. J., BURROWS, J. N., EDWARDS, P. D., FOLMER, R. H., GESCHWINDNER, S., GRIFFEN, E. J., KENNY, P. W., NOWAK, T., OLSSON, L. L., SANGANEE, H. & SHAPIRO, A. B. 2007. An integrated approach to fragment-based lead generation: philosophy, strategy and case studies from AstraZeneca's drug discovery programmes. Curr Top Med Chem, 7, 1600-29.

ALLEN, K. N., BELLAMACINA, C. R., DING, X., JEFFEREY, C. J., MATTOS, C., PETSKO, G. A. & RINGE, D. 1996. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. J. Med. Chem, 100, 2605-2611.

ANDREWS, P. R., CRAIK, D. J. & MARTIN, J. L. 1984. Functional group contributions to drug-receptor interactions. J Med Chem, 27, 1648-57.

ANTONYSAMY, S., HIRST, G., PARK, F., SPRENGELER, P., STAPPENBECK, F., STEENSMA, R., WILSON, M. & WONG, M. 2009. Fragment-based discovery of JAK-2 inhibitors. Bioorg Med Chem Lett, 19, 279-82.

ANTONYSAMY, S. S., AUBOL, B., BLANEY, J., BROWNER, M. F., GIANNETTI, A. M., HARRIS, S. F., HEBERT, N., HENDLE, J., HOPKINS, S., JEFFERSON, E., KISSINGER, C., LEVEQUE, V., MARCIANO, D., MCGEE, E., NAJERA, I., NOLAN, B., TOMIMOTO, M., TORRES, E. & WRIGHT, T. 2008. Fragment-based discovery of hepatitis C virus NS5b RNA polymerase inhibitors. Bioorg Med Chem Lett, 18, 2990-5.

ARNAUDOV, L. N. & DE VRIES, R. 2005. Thermally induced fibrillar aggregation of hen egg white lysozyme. Biophys J, 88, 515-26.

BABAOGLU, K. & SHOICHET, B. K. 2006. Deconstructing fragment-based inhibitor discovery. Nat Chem Biol, 2, 720-3.

BARNETT, J. M., CADMAN, A., GOR, D., DEMPSEY, M., WALTERS, M., CANDLIN, A., TISDALE, M., MORLEY, P. J., OWENS, I. J., FENTON, R. J., LEWIS, A. P., CLAAS, E. C., RIMMELZWAAN, G. F., DE GROOT, R. & OSTERHAUS, A. D. 2000. Zanamivir susceptibility monitoring and characterization of influenza virus clinical isolates obtained during phase II clinical efficacy studies. Antimicrob Agents Chemother, 44, 78-87.

BAURIN, N., ABOUL-ELA, F., BARRIL, X., DAVIS, B., DRYSDALE, M., DYMOCK, B., FINCH, H., FROMONT, C., RICHARDSON, C., SIMMONITE, H. & HUBBARD, R. E. 2004. Design and characterization of libraries of molecular fragments for use in NMR screening against protein targets. J Chem Inf Comput Sci, 44, 2157-66.

BLANEY, J., NIENABER, V. & BURLEY, S. K. 2006. Fragment-based Lead Discovery and Optimization Using X-Ray Crystallography, Computational Chemistry, and High-throughput Organic Synthesis. In: ERLANSON, W. J. A. D. A. (ed.) Fragment-based Approaches in Drug Discovery. Weinheim, Germany: WILEY-VCH.

BLOMBERG, N., COSGROVE, D. A., KENNY, P. W. & KOLMODIN, K. 2009. Design of compound libraries for fragment screening. J Comput Aided Mol Des.

BOHACEK, R. S., MCMARTIN, C. & GUIDA, W. C. 1996. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev, 16, 3-50.

BOHM, H. J. 1992. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. J Comput Aided Mol Des, 6, 593-606.

BOTTING, R. M. 2006. Inhibitors of cyclooxygenases: mechanisms, selectivity and uses. J Physiol Pharmacol, 57 Suppl 5, 113-24.

BOWYER, P. W., TATE, E. W., LEATHERBARROW, R. J., HOLDER, A. A., SMITH, D. F. & BROWN, K. A. 2008. N-myristoyltransferase: a prospective drug target for protozoan parasites. ChemMedChem, 3, 402-8.

BRADSHAW, J. 1997. Introduction to Tversky similarity measure. http://www.daylight.com/meetings/mug97/Bradshaw/MUG97/tv_tversky.html.

BRANDTS, J. F. & LIN, L. N. 1990. Study of strong to ultratight protein interactions using differential scanning calorimetry. Biochemistry, 29, 6927-40.

BREWER, M., ICHIHARA, O., KIRCHHOFF, C., SCHADE, M. & WHITTAKER, M. 2008. Assembling a Fragment Library. In: ZARTLER, E. R. & SHAPIRO, M. J. (eds.) Fragment-Based Drug Discovery. Wiltshire, UK: Wiley.

BRIGHT, K. J. 2009. Fragment-Based Hit Discovery. PhD, University of York.

BROUGH, P. A., AHERNE, W., BARRIL, X., BORGOGNONI, J., BOXALL, K., CANSFIELD, J. E., CHEUNG, K. M., COLLINS, I., DAVIES, N. G., DRYSDALE, M. J., DYMOCK, B., ECCLES, S. A., FINCH, H., FINK, A., HAYES, A., HOWES, R., HUBBARD, R. E., JAMES, K., JORDAN, A. M., LOCKIE, A., MARTINS, V., MASSEY, A., MATTHEWS, T. P., MCDONALD, E., NORTHFIELD, C. J., PEARL, L. H., PRODROMOU, C., RAY, S., RAYNAUD, F. I., ROUGHLEY, S. D., SHARP, S. Y., SURGENOR, A., WALMSLEY, D. L., WEBB, P., WOOD, M., WORKMAN, P. & WRIGHT, L. 2008. 4,5-diarylisoxazole Hsp90 chaperone inhibitors: potential therapeutic agents for the treatment of cancer. J Med Chem, 51, 196-218.

BROWN, F. K. 1998. Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery. In: JAMES, A. B. (ed.) Annual Reports in Medicinal Chemistry. Academic Press.

BRZOZOWSKI, A. M. & WALTON, J. 2001. Clear strategy screens for macromolecular crystallization. Journal of Applied Crystallography, 34, 97-101.

BUTINA, D. 1999. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster

Small and Large Data Sets. J. Chem. Inf. Comput. Sci., 39, 747-750.

CAMPBELL, S. F. 2000. Science, art and drug discovery: a personal perspective. Clin Sci (Lond), 99, 255-60.

CARD, G. L., BLASDEL, L., ENGLAND, B. P., ZHANG, C., SUZUKI, Y., GILLETTE, S., FONG, D., IBRAHIM, P. N., ARTIS, D. R., BOLLAG, G., MILBURN, M. V., KIM, S. H., SCHLESSINGER, J. & ZHANG, K. Y. 2005. A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. Nat Biotechnol, 23, 201-7.

CETINBAS, N., MACAULEY, M. S., STUBBS, K. A., DRAPALA, R. & VOCADLO, D. J. 2006. Identification of Asp174 and Asp175 as the key catalytic residues of human O-GlcNAcase by functional analysis of site-directed mutants. Biochemistry, 45, 3835-44.

CHEN, W. L. 2006. Chemoinformatics: past, present, and future. J Chem Inf Model, 46, 2230-55.

CHENG, X., COLE, R. N., ZAIA, J. & HART, G. W. 2000. Alternative O-glycosylation/O-phosphorylation of the murine estrogen receptor beta. Biochemistry, 39, 11609-20.

CHENG, Y. & PRUSOFF, W. H. 1973. RELATIONSHIP BETWEEN INHIBITION CONSTANT (K1) AND CONCENTRATION OF INHIBITOR WHICH CAUSES 50 PER CENT INHIBITION (I50) OF AN ENZYMATIC-REACTION. Biochemical Pharmacology, 22, 3099-3108.

CHOU, C. F., SMITH, A. J. & OMARY, M. B. 1992. Characterization and dynamics of O-linked glycosylation of human cytokeratin 8 and 18. J Biol Chem, 267, 3901-6.

CHOU, T. Y. & HART, G. W. 2001. O-linked N-acetylglucosamine and cancer: messages from the glycosylation of c-Myc. Adv Exp Med Biol, 491, 413-8.

CHUNG, F., TISNE, C., LECOURT, T., DARDEL, F. & MICOUIN, L. 2007. NMR-guided fragment-based approach for the design of tRNA(Lys3) ligands. Angew Chem Int Ed Engl, 46, 4489-91.

CIMMPERMAN, P., BARANAUSKIENE, L., JACHIMOVICIUTE, S., JACHNO, J., TORRESAN, J., MICHAILOVIENE, V., MATULIENE, J., SEREIKAITE, J., BUMELIS, V. & MATULIS, D. 2008. A quantitative model of thermal stabilization and destabilization of proteins by ligands. Biophys J, 95, 3222-31.

COLMAN, P. 2006. Anti-Influenza Drugs from Neuramidase Inhibitors. In: HUBBARD, R. E., CLORE, M. & LILLEY, D. M. (eds.) Structure-Based Drug Discovery. Cambridge: RSC Publishing.

CONGREVE, M., AHARONY, D., ALBERT, J., CALLAGHAN, O., CAMPBELL, J., CARR, R. A., CHESSARI, G., COWAN, S., EDWARDS, P. D., FREDERICKSON, M., MCMENAMIN, R., MURRAY, C. W., PATEL, S. & WALLIS, N. 2007. Application of fragment screening by X-ray crystallography to the discovery of aminopyridines as inhibitors of beta-secretase. J Med Chem, 50, 1124-32.

CONGREVE, M., CARR, R., MURRAY, C. & JHOTI, H. 2003. A 'rule of three' for fragment-based lead discovery? Drug Discov Today, 8, 876-7.

CONGREVE, M., CHESSARI, G., TISI, D. & WOODHEAD, A. J. 2008. Recent developments in fragment-based drug discovery. J Med Chem, 51, 3661-80.

CRISMAN, T. J., BENDER, A., MILIK, M., JENKINS, J. L., SCHEIBER, J., SUKURU, S. C., FEJZO, J., HOMMEL, U., DAVIES, J. W. & GLICK, M. 2008. "Virtual fragment linking": an approach to identify potent binders from low affinity fragment hits. J Med Chem, 51, 2481-91.

CROWTHER, G. J., NAPULI, A. J., THOMAS, A. P., CHUNG, D. J., KOVZUN, K. V., LEIBLY, D. J., CASTANEDA, L. J., BHANDARI, J., DAMMAN, C. J., HUI, R., HOL, W. G., BUCKNER, F. S., VERLINDE, C. L., ZHANG, Z., FAN, E. & VAN VOORHIS, W. C. 2009. Buffer optimization of thermal melt assays of Plasmodium proteins for detection of small-molecule ligands. J Biomol Screen, 14, 700-7.

CUMMINGS, M. D., FARNUM, M. A. & NELEN, M. I. 2006. Universal screening methods and applications of ThermoFluor. J Biomol Screen, 11, 854-63.

DAHLQUIST, F. W., JAO, L. & RAFTERY, M. 1966. On the binding of chitin oligosaccharides to lysozyme. Proc Natl Acad Sci U S A, 56, 26-30.

DALBY, A., NOURSE, J. G., HOUNSHELL, W. D., GUSHURST, A. K. I., GRIER, D. L., LELAND, B. A. & LAUFER, J. 1992. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. Journal of Chemical Information and Computer Sciences, 32, 244-255.

DANIEL, E. & WEBER, G. 1966. Cooperative effects in binding by bovine serum albumin. I. The binding of 1-anilino-8-naphthalenesulfonate. Fluorimetric titrations. Biochemistry, 5, 1893-900.

DENNIS, R. J., TAYLOR, E. J., MACAULEY, M. S., STUBBS, K. A., TURKENBURG, J. P., HART, S. J., BLACK, G. N., VOCADLO, D. J. & DAVIES, G. J. 2006. Structure and mechanism of a bacterial beta-glucosaminidase having O-GlcNAcase activity. Nat Struct Mol Biol, 13, 365-71.

DON, R. H., COX, P. T., WAINWRIGHT, B. J., BAKER, K. & MATTICK, J. S. 1991. 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res, 19, 4008.

EGNER, U., KRATZSCHMAR, J., KREFT, B., POHLENZ, H. D. & SCHNEIDER, M. 2005. The target discovery process. Chembiochem, 6, 468-79.

EISEN, M. B., WILEY, D. C., KARPLUS, M. & HUBBARD, R. E. 1994. HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. Proteins, 19, 199-221.

EMSLEY, P. & COWTAN, K. 2004. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr, 60, 2126-32.

ENGEL, T. 2006. Basic overview of chemoinformatics. J Chem Inf Model, 46, 2267-77.

ENGLISH, A. C., DONE, S. H., CAVES, L. S., GROOM, C. R. & HUBBARD, R. E. 1999. Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2

ERICSSON, U. B., HALLBERG, B. M., DETITTA, G. T., DEKKER, N. & NORDLUND, P. 2006. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal Biochem, 357, 289-98.

ERLANSON, D. A., BRAISTED, A. C., RAPHAEL, D. R., RANDAL, M., STROUD, R. M., GORDON, E. M. & WELLS, J. A. 2000. Site-directed ligand discovery. Proc Natl Acad Sci U S A, 97, 9367-72.

FEJZO, J., LEPRE, C. A., PENG, J. W., BEMIS, G. W., AJAY, MURCKO, M. A. & MOORE, J. M. 1999. The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. Chem Biol, 6, 755-69.

FINK, T., BRUGGESSER, H. & REYMOND, J. L. 2005. Virtual exploration of the small-molecule chemical universe below 160 Daltons. Angew Chem Int Ed Engl, 44, 1504-8.

FISCHER, M. & HUBBARD, R. E. 2009. Fragment-based ligand discovery. Mol Interv, 9, 22-30.

FISCHER, M., LEECH, A. P. & HUBBARD, R. E. 2011. Comparative Assessment of Different Histidine-Tags for Immobilization of Protein onto Surface Plasmon Resonance Sensorchips. Anal Chem.

FREIRE, E. 2008. Do enthalpy and entropy distinguish first in class from best in class? Drug Discov Today, 13, 869-74.

GAO, Y., WELLS, L., COMER, F. I., PARKER, G. J. & HART, G. W. 2001. Dynamic O-glycosylation of nuclear and cytosolic proteins: cloning and characterization of a neutral, cytosolic beta-N-acetylglucosaminidase from human brain. J Biol Chem, 276, 9838-45.

GHOSE, A. K. & CRIPPEN, G. M. 1986. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. Journal of Computational Chemistry, 7, 565-577.

GHOSE, A. K. & CRIPPEN, G. M. 1987. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. Journal of Chemical Information and Computer Sciences, 27, 21-35.

GHOSE, A. K., PRITCHETT, A. & CRIPPEN, G. M. 1988. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. Journal of Computational Chemistry, 9, 80-90.

GHOSE, A. K., VISWANADHAN, V. N. & WENDOLOSKI, J. J. 1998. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using

Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. The Journal of Physical Chemistry A, 102, 3762-3772.

GIANTI, E. & SARTORI, L. 2008. Identification and selection of "privileged fragments" suitable for primary screening. J Chem Inf Model, 48, 2129-39.

GILLET, V., JOHNSON, A. P., MATA, P., SIKE, S. & WILLIAMS, P. 1993. SPROUT: a program for structure generation. J Comput Aided Mol Des, 7, 127-53.

GOMEZ, J., HILSER, V. J., XIE, D. & FREIRE, E. 1995. The heat capacity of proteins. Proteins, 22, 404-12.

GOODFORD, P. J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem, 28, 849-57.

GRIFFITH, L. S. & SCHMITZ, B. 1995. O-linked N-acetylglucosamine is upregulated in Alzheimer brains. Biochem Biophys Res Commun, 213, 424-31.

HAJDUK, P. J. 2006. Fragment-based drug design: How big is too big? Journal of Medicinal Chemistry, 49, 6972-6976.

HAJDUK, P. J. 2006. SAR by NMR: putting the pieces together. Mol Interv, 6, 266-72.

HAJDUK, P. J. & GREER, J. 2007. A decade of fragment-based drug design: strategic advances and lessons learned. Nat Rev Drug Discov, 6, 211-9.

HAJDUK, P. J. & SAUER, D. R. 2008. Statistical analysis of the effects of common chemical substituents on ligand potency. J Med Chem, 51, 553-64.

HALL, L. H. & KIER, L. B. 1995. ELECTROTOPOLOGICAL STATE INDEXES FOR ATOM TYPES - A NOVEL COMBINATION OF ELECTRONIC, TOPOLOGICAL, AND VALENCE STATE INFORMATION. Journal of Chemical Information and Computer Sciences, 35, 1039-1045.

HALL, L. H., MOHNEY, B. & KIER, L. B. 1991. THE ELECTROTOPOLOGICAL STATE - AN ATOM INDEX FOR QSAR. Quantitative Structure-Activity Relationships, 10, 43-51.

HAMALAINEN, M. D., ZHUKOV, A., IVARSSON, M., FEX, T., GOTTFRIES, J., KARLSSON, R. & BJORSNE, M. 2008. Label-free primary screening and affinity ranking of fragment libraries using parallel analysis of protein panels. J Biomol Screen, 13, 202-9.

HANN, M., HUDSON, B., LEWELL, X., LIFELY, R., MILLER, L. & RAMSDEN, N. 1999. Strategic pooling of compounds for high-throughput screening. J Chem Inf Comput Sci, 39, 897-902.

HANN, M. M., LEACH, A. R. & HARPER, G. 2001. Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci, 41, 856-64.

HANSCH, C. 1993. QUANTITATIVE STRUCTURE-ACTIVITY-RELATIONSHIPS AND THE UNNAMED SCIENCE. Accounts of Chemical Research, 26, 147-153.

HARTSHORN, M. J., MURRAY, C. W., CLEASBY, A., FREDERICKSON, M., TICKLE, I. J. & JHOTI, H. 2005. Fragment-based lead discovery using X-ray crystallography. J Med Chem, 48, 403-13.

HE, Y. 2011. Mechanism and Inhibition of a Bacterial O-GlcNAcase. Doctor of Philosophy, University of York.

HE, Y., BUBB, A. K., STUBBS, K. A., GLOSTER, T. M. & DAVIES, G. J. 2011. Inhibition of a bacterial O-GlcNAcase homologue by lactone and lactam derivatives: structural, kinetic and thermodynamic analyses. Amino Acids, 40, 829-39.

HOHWY, M., SPADOLA, L., LUNDQUIST, B., HAWTIN, P., DAHMEN, J., GROTH-CLAUSEN, I., NILSSON, E., PERSDOTTER, S., VON WACHENFELDT, K., FOLMER, R. H. & EDMAN, K. 2008. Novel prostaglandin D synthase inhibitors generated by fragment-based drug design. J Med Chem, 51, 2178-86.

HOPKINS, A. L., GROOM, C. R. & ALEX, A. 2004. Ligand efficiency: a useful metric for lead selection. Drug Discov Today, 9, 430-1.

HOWARD, S., BERDINI, V., BOULSTRIDGE, J. A., CARR, M. G., CROSS, D. M., CURRY, J., DEVINE, L. A., EARLY, T. R., FAZAL, L., GILL, A. L., HEATHCOTE, M., MAMAN, S., MATTHEWS, J. E., MCMENAMIN, R. L., NAVARRO, E. F., O'BRIEN, M. A., O'REILLY, M., REES, D. C., REULE, M., TISI, D., WILLIAMS, G., VINKOVIC, M. & WYATT, P. G. 2009. Fragment-Based Discovery of the Pyrazol-4-yl Urea (AT9283), a Multitargeted Kinase Inhibitor with Potent Aurora Kinase Activity. Journal of Medicinal Chemistry, 52, 379-388.

HU, T. C., KORCZYNSKA, J., SMITH, D. K. & BRZOZOWSKI, A. M. 2008. High-molecular-weight polymers for protein crystallization: poly-gamma-glutamic acid-based precipitants. Acta Crystallogr D Biol Crystallogr, 64, 957-63.

HUBBARD, R. E. 2006. 3D Structure and the Drug Discovery Process. In: HUBBARD, R. E., CLORE, M. & LILLEY, D. M. (eds.) Structure-Based Drug Discovery. Cambridge: RSC Publishing.

HUBBARD, R. E. 2008. Fragment approaches in structure-based drug discovery. J Synchrotron Radiat, 15, 227-30.

HUBBARD, R. E., DAVIS, B., CHEN, I. & DRYSDALE, M. J. 2007. The SeeDs approach: integrating fragments into drug discovery. Curr Top Med Chem, 7, 1568-81.

HUNG, A. W., RAMEK, A., WANG, Y., KAYA, T., WILSON, J. A., CLEMONS, P. A. & YOUNG, D. W. 2011. Route to three-dimensional fragments using diversity-oriented synthesis. Proc Natl Acad Sci U S A, 108, 6799-804.

HUTH, J. R., PARK, C., PETROS, A. M., KUNZER, A. R., WENDT, M. D., WANG, X., LYNCH, C. L., MACK, J. C., SWIFT, K. M., JUDGE, R. A., CHEN, J., RICHARDSON, P. L., JIN, S., TAHIR, S. K., MATAYOSHI, E. D., DORWIN, S. A., LADROR, U. S., SEVERIN, J. M., WALTER, K. A., BARTLEY, D. M., FESIK, S. W., ELMORE, S. W. & HAJDUK, P. J. 2007. Discovery and design of novel HSP90 inhibitors using multiple fragment-based design strategies. Chem Biol Drug Des, 70, 1-12.

HUUSKONEN, J. 2000. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. Journal of Chemical Information and Computer Sciences, 40, 773-777.

IRWIN, J. J. & SHOICHET, B. K. 2005. ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model, 45, 177-82.

JACCARD, P. 1901. Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Socit Vaudoise des Sciences Naturelles, 37, 547-579.

JENCKS, W. P. 1981. On the attribution and additivity of binding energies. Proc Natl Acad Sci U S A, 78, 4046-50.

JHOTI, H. 2007. Fragment-based drug discovery using rational design. Ernst Schering Found Symp Proc, 169-85.

JHOTI, H., CLEASBY, A., VERDONK, M. & WILLIAMS, G. 2007. Fragment-based screening using X-ray crystallography and NMR spectroscopy. Current Opinion in Chemical Biology, 11, 485-493.

JORDAN, V. C. 2006. Tamoxifen (ICI46,474) as a targeted therapy to treat and prevent breast cancer. Br J Pharmacol, 147 Suppl 1, S269-76.

KAMEMURA, K., HAYES, B. K., COMER, F. I. & HART, G. W. 2002. Dynamic interplay between O-glycosylation and O-phosphorylation of nucleocytoplasmic proteins: alternative glycosylation/phosphorylation of THR-58, a known mutational hot spot of c-Myc in lymphomas, is regulated by mitogens. J Biol Chem, 277, 19229-35.

KOSTER, H., CRAAN, T., BRASS, S., HERHAUS, C., ZENTGRAF, M., NEUMANN, L., HEINE, A. & KLEBE, G. 2011. A small nonrule of 3 compatible fragment library provides high hit rate of endothiapepsin crystal structures with various fragment chemotypes. J Med Chem, 54, 7784-96.

KRANZ, J. K. 19/03/2012 2012. RE: Discussion - Literature of Thermal Shift Analysis. Type to SCHULZ, M. N.

KRANZ, J. K. & SCHALK-HIHI, C. 2011. Protein thermal shifts to identify low molecular weight fragments. Methods Enzymol, 493, 277-98.

KREBS, M. R., WILKINS, D. K., CHUNG, E. W., PITKEATHLY, M. C., CHAMBERLAIN, A. K., ZURDO, J., ROBINSON, C. V. & DOBSON, C. M. 2000. Formation and seeding of amyloid fibrils from wild-type hen lysozyme and a peptide fragment from the beta-domain. J Mol Biol, 300, 541-9.

KREPPEL, L. K., BLOMBERG, M. A. & HART, G. W. 1997. Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. J Biol Chem, 272, 9308-15.

KUMAR, S., RAVI, V. K. & SWAMINATHAN, R. 2009. Suppression of lysozyme aggregation at alkaline pH by tri-N-acetylchitotriose. Biochim Biophys Acta, 1794, 913-20.

KUNTZ, I. D., CHEN, K., SHARP, K. A. & KOLLMAN, P. A. 1999. The maximal affinity of ligands. Proc Natl Acad Sci U S A, 96, 9997-10002.

KUO, L. C. 2011. Fragment-based drug design - tools, practical approaches, and examples. Preface. Methods Enzymol, 493, xxi-xxii.

LAURI, G. & BARTLETT, P. A. 1994. CAVEAT: a program to facilitate the design of organic molecules. J Comput Aided Mol Des, 8, 51-66.

LAW, R., BARKER, O., BARKER, J. J., HESTERKAMP, T., GODEMANN, R., ANDERSEN, O., FRYATT, T., COURTNEY, S., HALLETT, D. & WHITTAKER, M. 2009. The multiple roles of computational chemistry in fragment-based drug design. J Comput Aided Mol Des.

LAYTON, C. J. & HELLINGA, H. W. 2010. Thermodynamic analysis of ligand-induced changes in protein thermal unfolding applied to high-throughput determination of ligand affinities with extrinsic fluorescent dyes. Biochemistry, 49, 10831-41.

LEACH, A. R. & GILLET, V. J. 2003. An Introduction to Chemoinformatics, Dordrecht, The Netherlands, Springer.

LEACH, A. R., HANN, M. M., BURROWS, J. N. & GRIFFEN, E. J. 2006. Fragment Screening: An Introduction: An Overview. In: HUBBARD, R. E. (ed.) Structure-Based Drug Discovery. Cambridge, United Kingdom: RSCPublishing.

LEESON, P. D. & SPRINGTHORPE, B. 2007. The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov, 6, 881-90.

LEO, A., HANSCH, C. & ELKINS, D. 1971. Partition coefficients and their uses. Chemical Reviews, 71, 525-616.

LEO, A. J. 1993. CALCULATING LOG P(OCT) FROM STRUCTURES. Chemical Reviews, 93, 1281-1306.

LEPRE, C. 2007. Fragment-based drug discovery using the SHAPES method. Expert Opinion on Drug Discovery, 2, 1555-1566.

LESLIE, A. G. W. & POWELL, H. R. 2007. Processing Diffraction Data with Mosflm Evolving Methods for Macromolecular Crystallography.

LEW, W., CHEN, X. & KIM, C. U. 2000. Discovery and development of GS 4104 (oseltamivir): an orally active influenza neuraminidase inhibitor. Curr Med Chem, 7, 663-72.

LINDSAY, M. A. 2003. Target discovery. Nat Rev Drug Discov, 2, 831-8.

LIPINSKI, C. A., LOMBARDO, F., DOMINY, B. W. & FEENEY, P. J. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev, 46, 3-26.

LONG, F., VAGIN, A. A., YOUNG, P. & MURSHUDOV, G. N. 2008. BALBES: a molecular-replacement pipeline. Acta Crystallogr D Biol Crystallogr, 64, 125-32.

MANNING, G., WHYTE, D. B., MARTINEZ, R., HUNTER, T. & SUDARSANAM, S. 2002. The protein kinase complement of the human genome. Science, 298, 1912-34.

MASON, J. S., MORIZE, I., MENARD, P. R., CHENEY, D. L., HULME, C. & LABAUDINIERE, R. F. 1999. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. J Med Chem, 42, 3251-64.

MATTOS, C. & RINGE, D. 1996. Locating and characterizing binding sites on proteins. Nat Biotechnol, 14, 595-9.

MATULIS, D., KRANZ, J. K., SALEMME, F. R. & TODD, M. J. 2005. Thermodynamic stability of carbonic anhydrase: measurements of binding affinity and stoichiometry using ThermoFluor. Biochemistry, 44, 5258-66.

MCCANN, J., SPINGARN, N. E., KOBORI, J. & AMES, B. N. 1975. Detection of carcinogens as mutagens: bacterial tester strains with R factor plasmids. Proc Natl Acad Sci U S A, 72, 979-83.

MEINDL, P. & TUPPY, H. 1969. [2-Deoxy-2,3-dehydrosialic acids. II. Competitive inhibition of Vibrio cholerae neuraminidase by 2-deoxy-2,3-dehydro-N-acylneuraminic acids]. Hoppe Seylers Z Physiol Chem, 350, 1088-92.

MIRANKER, A. & KARPLUS, M. 1991. Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins, 11, 29-34.

MOHAN, H. & VASELLA, A. 2000. An improved synthesis of 2-acetamido-2-deoxy-D-gluconohydroximolactone (PUGNAc), a strong inhibitor of beta-N-acetylglucosaminidases. Helvetica Chimica Acta, 83, 114-118.

MORSKY, P. 1983. Turbidimetric determination of lysozyme with Micrococcus lysodeikticus cells: reexamination of reaction conditions. Anal Biochem, 128, 77-85.

MUKHERJEE, D. 2002. Selective cyclooxygenase-2 (COX-2) inhibitors and potential risk of cardiovascular events. Biochem Pharmacol, 63, 817-21.

NEUMANN, T., JUNKER, H. D., SCHMIDT, K. & SEKUL, R. 2007. SPR-based fragment screening: advantages and applications. Curr Top Med Chem, 7, 1630-42.

NIESEN, F. H., BERGLUND, H. & VEDADI, M. 2007. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat Protoc, 2, 2212-21.

OLTERSDORF, T., ELMORE, S. W., SHOEMAKER, A. R., ARMSTRONG, R. C., AUGERI, D. J., BELLI, B. A., BRUNCKO, M., DECKWERTH, T. L., DINGES, J., HAJDUK, P. J., JOSEPH, M. K., KITADA, S., KORSMEYER, S. J., KUNZER, A. R., LETAI, A., LI, C., MITTEN, M. J., NETTESHEIM, D. G., NG, S., NIMMER, P. M., O'CONNOR, J. M., OLEKSIJEW, A., PETROS, A. M., REED, J. C., SHEN, W., TAHIR, S. K., THOMPSON, C. B., TOMASELLI, K. J., WANG, B., WENDT, M. D., ZHANG, H., FESIK, S. W. & ROSENBERG, S. H. 2005. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature, 435, 677-81.

ONDETTI, M. A., RUBIN, B. & CUSHMAN, D. W. 1977. Design of specific inhibitors of angiotensin-converting enzyme: new class of orally active antihypertensive agents. Science, 196, 441-4.

OPREA, T. I., DAVIS, A. M., TEAGUE, S. J. & LEESON, P. D. 2001. Is there a difference between leads and drugs? A historical perspective. J Chem Inf Comput Sci, 41, 1308-15.

PACE, C. N. & MCGRATH, T. 1980. Substrate stabilization of lysozyme to thermal and guanidine hydrochloride denaturation. J Biol Chem, 255, 3862-5.

PANTOLIANO, M. W., PETRELLA, E. C., KWASNOSKI, J. D., LOBANOV,

V. S., MYSLIK, J., GRAF, E., CARVER, T., ASEL, E., SPRINGER, B. A., LANE, P. & SALEMME, F. R. 2001. High-density miniaturized thermal shift assays as a general strategy for drug discovery. J Biomol Screen, 6, 429-40.

PATTERSON, A. W., WOOD, W. J. L. & ELLMAN, J. A. 2007. Substrate activity screening (SAS): a general procedure for the preparation and screening of a fragment-based non-peptidic protease substrate library for inhibitor discovery. Nature Protocols, 2, 424-433.

PAUL, S. M., MYTELKA, D. S., DUNWIDDIE, C. T., PERSINGER, C. C., MUNOS, B. H., LINDBORG, S. R. & SCHACHT, A. L. 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov, 9, 203-14.

PELLECCHIA, M., BERTINI, I., COWBURN, D., DALVIT, C., GIRALT, E., JAHNKE, W., JAMES, T. L., HOMANS, S. W., KESSLER, H., LUCHINAT, C., MEYER, B., OSCHKINAT, H., PENG, J., SCHWALBE, H. & SIEGAL, G. 2008. Perspectives on NMR in drug discovery: a technique comes of age. Nat Rev Drug Discov, 7, 738-45.

PEROLA, E. 2010. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. J Med Chem, 53, 2986-97.

PRAKESCH, M., DENISOV, A. Y., NAIM, M., GEHRING, K. & ARYA, P. 2008. The discovery of small molecule chemical probes of Bcl-X(L) and Mcl-1. Bioorg Med Chem, 16, 7443-9.

PROLL, F., FECHNER, P. & PROLL, G. 2009. Direct optical detection in fragment-based screening. Anal Bioanal Chem, 393, 1557-62.

RECHT, M. I., DE BRUYKER, D., BELL, A. G., WOLKIN, M. V., PEETERS, E., ANDERSON, G. B., KOLATKAR, A. R., BERN, M. W., KUHN, P., BRUCE, R. H. & TORRES, F. E. 2008. Enthalpy array analysis of enzymatic and binding reactions. Anal Biochem, 377, 33-9.

REYMOND, J. L., BLUM, L. C. & VAN DEURSEN, R. 2011. Exploring the chemical space of known and unknown organic small molecules at www.gdb.unibe.ch. Chimia (Aarau), 65, 863-7.

ROGERS, D. & HAHN, M. 2010. Extended-connectivity fingerprints. J Chem Inf Model, 50, 742-54.

RUBIN, B., ANTONACCIO, M. J. & HOROVITZ, Z. P. 1978. Captopril (SQ 14,225) (D-3-mercapto-2-methylpropranoyl-L-proline): a novel orally active inhibitor of angiotensin-converting enzyme and antihypertensive agent. Prog Cardiovasc Dis, 21, 183-94.

SAXTY, G., WOODHEAD, S. J., BERDINI, V., DAVIES, T. G., VERDONK, M. L., WYATT, P. G., BOYLE, R. G., BARFORD, D., DOWNHAM, R., GARRETT, M. D. & CARR, R. A. 2007. Identification of inhibitors of protein kinase B using fragment-based lead discovery. J Med Chem, 50, 2293-6.

SCHNEIDER, G. & FECHNER, U. 2005. Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov, 4, 649-63.

SCHULZ, M. N. & HUBBARD, R. E. 2009. Recent progress in fragment-based lead discovery. Curr Opin Pharmacol, 9, 615-21.

SCHULZ, M. N., LANDSTROM, J., BRIGHT, K. & HUBBARD, R. E. 2011. Design of a fragment library that maximally represents available chemical space. J Comput Aided Mol Des, 25, 611-20.

SCHULZ, M. N., LANDSTROM, J. & HUBBARD, R. E. 2012. MTSA-A Matlab program to fit thermal shift data. Anal Biochem, [Epub ahead of print]

SHUGAR, D. 1952. The measurement of lysozyme activity and the ultra-violet inactivation of lysozyme. Biochim Biophys Acta, 8, 302-9.

SHUKER, S. B., HAJDUK, P. J., MEADOWS, R. P. & FESIK, S. W. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. Science, 274, 1531-4.

SIEGAL, G., AB, E. & SCHULTZ, J. 2007. Integration of fragment screening and library design. Drug Discov Today, 12, 1032-9.

SOPHIANOPOULOS, A. J. & VAN HOLDE, K. E. 1961. Evidence for dimerization of lysozyme in alkaline solution. J Biol Chem, 236, PC82-PC83.

SOPHIANOPOULOS, A. J. & VANHOLDE, K. E. 1964. Physical Studies of Muramidase (Lysozyme). Ii. Ph-Dependent Dimerization. J Biol Chem, 239, 2516-24.

SORRELL, F. J., GREENWOOD, G. K., BIRCHALL, K. & CHEN, B. 2010. Development of a differential scanning fluorimetry based high throughput screening assay for the discovery of affinity binders against an anthrax protein. J Pharm Biomed Anal, 52, 802-8.

STUDIER, F. W. 2005. Protein production by auto-induction in high density shaking cultures. Protein Expr Purif, 41, 207-34.

TANIMOTO, T. T. 1957. IBM Internal Report 17th Nov. 1957.

TAYLOR, R. 1995. Simulation Analysis of Experimental Design Startegies for Screening Random Compounds as Potential New Drugs and Agrochemicals. J. Chem. Inf. Comput. Sci., 35, 59-67.

TETKO, I. V., TANCHUK, V. Y., KASHEVA, T. N. & VILLA, A. E. P. 2001. Estimation of aqueous solubility of chemical compounds using E-state indices. Journal of Chemical Information and Computer Sciences, 41, 1488-1493.

THOMPSON, D. C., DENNY, R. A., NILAKANTAN, R., HUMBLET, C., JOSEPH-MCCARTHY, D. & FEYFANT, E. 2008. CONFIRM: connecting fragments found in receptor molecules. J Comput Aided Mol Des, 22, 761-72.

TORRES, C. R. & HART, G. W. 1984. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. J Biol Chem, 259, 3308-17.

TUMMINO, P. J. & COPELAND, R. A. 2008. Residence time of receptor-ligand complexes and its effect on biological function. Biochemistry, 47, 5481-92.

TVERSKY, A. 1977. Features of Similarity. Psychological Review, 84, 327-352.

ULLMANN, J. R. 1976. An Algorithm for Subgraph Isomorphism. J. ACM, 23, 31-42.

VAGIN, A. A., STEINER, R. A., LEBEDEV, A. A., POTTERTON, L., MCNICHOLAS, S., LONG, F. & MURSHUDOV, G. N. 2004. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr D Biol Crystallogr, 60, 2184-95.

VEDADI, M., NIESEN, F. H., ALLALI-HASSANI, A., FEDOROV, O. Y., FINERTY, P. J., JR., WASNEY, G. A., YEUNG, R., ARROWSMITH, C., BALL, L. J., BERGLUND, H., HUI, R., MARSDEN, B. D., NORDLUND, P., SUNDSTROM, M., WEIGELT, J. & EDWARDS, A. M. 2006. Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. Proc Natl Acad Sci U S A, 103, 15835-40.

VERDONK, M. L. & REES, D. C. 2008. Group efficiency: a guideline for hits-to-leads chemistry. ChemMedChem, 3, 1179-80.

VERHEIJ, H. J. 2006. Leadlikeness and structural diversity of synthetic screening libraries. Mol Divers, 10, 377-88.

VISWANADHAN, V. N., GHOSE, A. K., REVANKAR, G. R. & ROBINS, R. K. 1989. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. Journal of Chemical Information and Computer Sciences, 29, 163-172.

VON ITZSTEIN, M., WU, W. Y., KOK, G. B., PEGG, M. S., DYASON, J. C., JIN, B., VAN PHAN, T., SMYTHE, M. L., WHITE, H. F., OLIVER, S. W. & ET AL. 1993. Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature, 363, 418-23.

VULPETTI, A., HOMMEL, U., LANDRUM, G., LEWIS, R. & DALVIT, C. 2009. Design and NMR-based screening of LEF, a library of chemical fragments with different local environment of fluorine. J Am Chem Soc, 131, 12949-59.

WALLIS, R. M., CORBIN, J. D., FRANCIS, S. H. & ELLIS, P. 1999. Tissue distribution of phosphodiesterase families and the effects of sildenafil on tissue cyclic nucleotides, platelet function, and the contractile responses of trabeculae carneae and aortic rings in vitro. Am J Cardiol, 83, 3C-12C.

WANG, C. K., WEERATUNGA, S. K., PACHECO, C. M. & HOFMANN, A. 2012. DMAN: a Java tool for analysis of multi-well differential scanning fluorimetry experiments. Bioinformatics, 28, 439-40.

WEBER, G. & LAURENCE, D. J. 1954. Fluorescent indicators of adsorption in aqueous solution and on the solid phase. Biochem J, 56, xxxi.

WEININGER, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, 28, 31-36.

WINTER, G. 2010. xia2: an expert system for macromolecular crystallography data reduction. Journal of Applied Crystallography, 43, 186-190.

WISWESSER, W. J. 1954. A line-formula chemical notation, New York, Crowell.

WYATT, P. G., WOODHEAD, A. J., BERDINI, V., BOULSTRIDGE, J. A., CARR, M. G., CROSS, D. M., DAVIS, D. J., DEVINE, L. A., EARLY, T. R., FELTELL, R. E., LEWIS, E. J., MCMENAMIN, R. L., NAVARRO, E. F., O'BRIEN, M. A., O'REILLY, M., REULE, M., SAXTY, G., SEAVERS, L. C. A., SMITH, D. M., SQUIRES, M. S., TREWARTHA, G., WALKER, M. T. & WOOLFORD, A. J. A. 2008. Identification of N-(4-piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a novel cyclin dependent kinase inhibitor using fragment-based X-ray crystallography and structure based drug design. Journal of Medicinal Chemistry, 51, 4986-4999.

YEH, A. P., MCMILLAN, A. & STOWELL, M. H. 2006. Rapid and simple protein-stability screens: application to membrane proteins. Acta Crystallogr D Biol Crystallogr, 62, 451-7.

ZHANG, R. & MONSMA, F. 2010. Fluorescence-based thermal shift assays. Curr Opin Drug Discov Devel, 13, 389-402.

ZUBRIENE, A., MATULIENE, J., BARANAUSKIENE, L., JACHNO, J., TORRESAN, J., MICHAILOVIENE, V., CIMMPERMAN, P. & MATULIS, D. 2009. Measurement of nanomolar dissociation constants by titration calorimetry and thermal shift assay - radicicol binding to Hsp90 and ethoxzolamide binding to CAII. Int J Mol Sci, 10, 2662-80.