The Development and Assessment of Computational Approaches to the Thermodynamics and Kinetics of Binding

Iva Lukac

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with AstraZeneca

December 2016

Acknowledgements

Abstract

Molecular recognition refers to the interaction between two or more molecules through complementary noncovalent bonding, for example, via hydrogen bonding, electrostatic interactions, van der Waals forces or hydrophobic forces. Molecular recognition plays an important role in biology and mediates interactions between receptors and ligands, antigens and antibodies, nucleic acids and proteins, proteins and proteins, enzymes and substrates, and nucleic acids with each other.

Many cellular processes are governed by a group of proteins acting in a coordinated manner; such complicated mechanisms are closely regulated: changes in the populations of particular complexes or changes in concentrations of the products of protein mediated reactions can switch cells from one state to another (from replication to apoptosis, for example). These small variations in molecular populations are caused by very delicate differences in the thermodynamics or kinetics of reactions. This implies that in order to understand not only biological systems in terms of their molecular components, but also to be able to predict and model system response to stimuli (whether it is a natural substrate or a drug), characterisation of the thermodynamic and kinetic components of the binding process is of paramount importance.

This combined computational-experimental project was focused on the development of new computational approaches able to predict the enthalpic component of ligand binding, using quantum mechanics. A concept of 'theoceptors' was developed, which are theoretical receptors constructed by computing the optimal geometry of ligands binding in the receptor. This project was supported by AstraZeneca, and it included an industrial placement in the Structural and Biophysical Sciences area, where the experimental data was generated to characterise the thermodynamics and kinetics of binding of a range of ligands to two biological targets, using two experimental techniques, isothermal titration calorimetry and surface plasmon resonance.

The findings contribute greatly to the process currently underway of expanding our understanding of the relevance of both of these aspects of biochemistry to drug discovery.

Table of contents

List of abbreviations

ΔH_i Ionisation enthalpy

ΔH_{obs} Observed enthalpy

ΔHvH van't Hoff enthalpy

ATP   Adenosine triphosphate

BK   Binding kinetics

DEC   Decrease

DFT   Density functional theory

DMSO   Dimethyl sulfoxide

DSC   Differential scanning calorimetry

FEP   Free-energy perturbation

GTO   Gaussian type orbitals

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

HTS   High-throughput screen

INC   Increase

InhA   Mtu Enoyl-ACP reductase

ITC   Isothermal titration calorimetry

LDHA   Lactate dehydrogenase A

LE   Ligand efficiency

LEM   Ligand efficiency metric

LIE Linear interaction energy

LLE Ligand lipophilic efficiency

MM Molecular mechanics

MMPA Matched molecular pair analysis

MOE Molecular operating environment

Mtb *Mycobacterium tuberculosis*

NAD Nicotinamide adenine dinucleotide

NED No effect determined

NMR Nuclear magnetic resonance

PD Pharmacodynamics

PDB Protein data bank

PHGDH 3-Phosphoglycerate dehydrogenase

PK Pharmacokinetics

QM Quantum mechanics

RSCC Real-space correlation coefficient

RU Response units

SAR Structure-activity relationship

SCF Self-consistent field

SPR Surface plasmon resonance

STO Slater type orbitals

T_m Melting temperature

TCEP Tris(2-carboxyethyl)phosphine hydrochloride

TI Thermodynamic integration

TRIS Tris(hydroxymethyl)aminomethane

Part I

Theoretical background and literature review

Chapter 1

Protein-ligand interactions

1.1 Introduction

In protein ligand complexes, complementarity of shape and chemical groups between the binding partners is usually observed: apolar groups come together, hydrogen bond donors match hydrogen bond acceptors and charged groups of the ligand orient so that they are neutralised by the oppositely charged groups of the protein. However, this image is only complete when changes in the interactions of both the binding partners with the surrounding environment are taken into consideration. Protein-ligand binding affinities are driven by the balance between a large number of factors, many of which cannot be accounted for by a simple analysis of a crystal structure. In solution, both water and ions can dramatically affect the electrostatics and change the environment found within certain binding sites.

Several types of non-covalent bonds are critical in both maintaining the shape of biological macromolecules and enabling one molecule to bind specifically but transiently to another (Figure 1.1). Although the energy released in formation of non-covalent interactions is low, when acting together, they lead to stable association between either different parts of a large molecule, or two different binding partners. Formulating the rules for molecular interactions is only possible within certain boundaries. Molecular interactions are non-additive and the same apparent interaction might result different amounts of observed energy in different contexts [1].
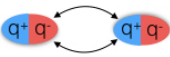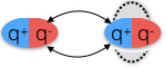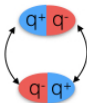
Interaction Type

**Dependence of
Energy on Distance**

charge-charge		longest-range force, nondirectional	$1/r$
charge-dipole		depends on dipole orientation	$1/r^2$
dipole-dipole		depends on mutual orientation of dipoles	$1/r^3$
charged-induced dipole		depends on polarisability of the molecule	$1/r^4$
dipole-induced dipole		depends on polarisability of the molecule	$1/r^5$
dispersion		involves mutual synchronisation of fluctuating charges	$1/r^6$
van der Waals repulsion		occurs when outer electron orbitals overlap	$1/r^{12}$
hydrogen bond		charge attraction+partial covalent bond	fixed length

Fig. 1.1 Noncovalent interactions.

1.2 Non-covalent interactions

1.2.1 Hydrogen bonds

Hydrogen bonds are attractive interactions between a hydrogen atom from a molecule or a fragment X-H, where X is more electronegative than H, and another electronegative atom or a group of atoms, such as nitrogen or oxygen.

An important feature of hydrogen bonds is directionality: these interactions are the strongest when the angle X-H$\cdots$Y is closest to 180°. Contribution of hydrogen bonds to an overall affinity is context dependent and may vary substantially in different environments. The spectrum of hydrogen bond strengths extends from 0.2-4 kcal/mol for weak bonds to 4-15 kcal/mol for moderate bonds and 15-40 kcal/mol for strong bonds [2]. In the protein apo state, almost all of its medium strength donor or acceptor groups will make hydrogen bonds with either other parts of the protein or the surrounding solvent. For a hydrogen bond to form, the desolvation of both the donor and acceptor has to occur. Because the effects of hydration and bond formation cancel out, these interactions do not necessarily add much binding free energy [3]. There is some evidence that energetic consequences

for not satisfying the hydrogen bond donor have more drastic effects than not satisfying the acceptor: the penalty for burying a donor in a desolvated state has been reported to be 4.3-5.3 kcal/mol for a hydroxyl group binding in a hydrophobic pocket [4]. Therefore, directionality and the energetic penalty for not forming hydrogen bonds makes hydrogen bonds a large source of specificity in biomolecular recognition.

An example of where an addition of hydrogen bonding moiety caused a dramatic effect on potency (so called 'activity cliff') are the two cyclin-dependent kinase 2 (CDK2) inhibitors in Figure 1.2. These two molecules are only distinguished by charged hydrogen bond interaction between a sulfonamide group in one of the purine- based inhibitors and residue Asp86 of the enzyme, causing a 166-fold potency increase [5].
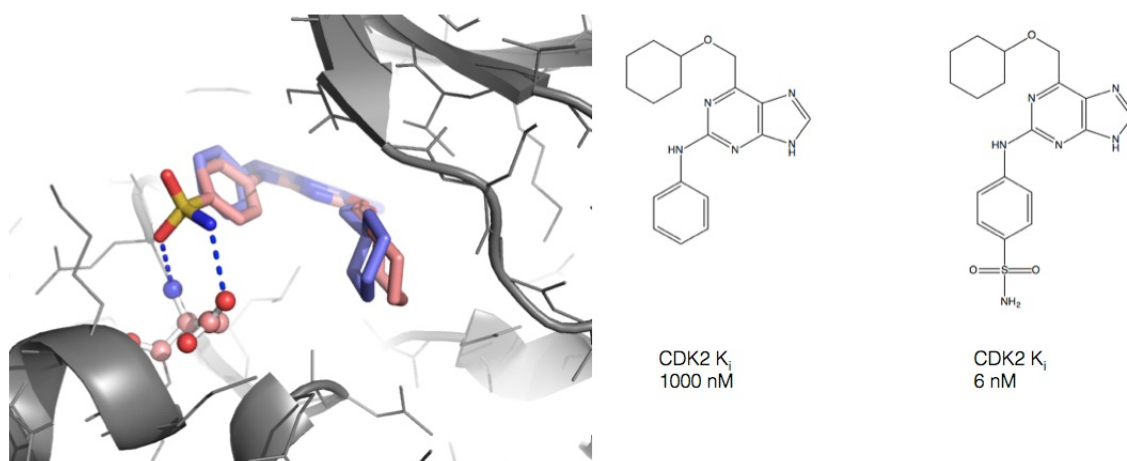
Fig. 1.2 Affinity data for two CDK2 purine-based inhibitors (PDB codes are 1H1Q and 1H1S).

1.2.2   Electrostatic interactions

Electrostatic interactions are attractive forces between opposite charges and can be approximately classified into three types: charge-charge, charge-dipole and dipole-dipole. Typical charge-charge (Coulombic) interactions are those between oppositely charged atoms, such as ligand functional groups and positively or negatively charged amino acid side chains. A dipole-induced dipole interaction is caused when a molecule with a permanent dipole approaches another molecule, causing it to become temporarily polarised. The two molecules are then attracted to interact with each other (Debye force); this is known as induction. Dipole-dipole interactions (Keesom forces) are electrostatic interactions between permanent dipoles in molecules.

1.2.3 Van der Waals interactions

A collective term to describe the non-Coulombic interactions is van der Waals interactions. These interactions have two components. Attractive force (London dispersion) is an interaction between two induced dipoles that dominates at longer distances, typically at 0.4-0.7 nm. Repulsive interaction, on the other hand, occurs at shorter distances as a result of the Pauli exclusion principle that prevents the collapse of the molecules (Figure 1.3). Van der Waals interactions are very weak, but because a large number of such interactions occur during the binding event, they contribute to the free energy significantly.

$$-\frac{A_{ij}}{r_{ij}^6}+\frac{B_{ij}}{r_{ij}^{12}}$$
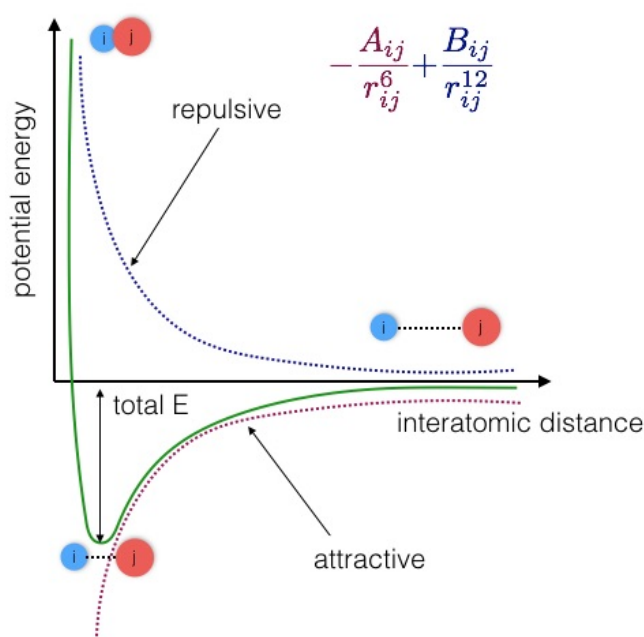
Fig. 1.3 Interatomic distances. Van der Waals attraction occurs at short range, and rapidly reduces as the interacting atoms move apart by a few Angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii. These effects are often modeled using a 6-12 equation, as shown in the figure. r is the distance of separation between both particles, A is $4\varepsilon\sigma^6$, B is $\varepsilon\sigma^{12}$, ε is the a measure of how strongly the two particles attract each other and σ is van der Waals radius.

1.2.4 Hydrophobic interactions and structural water

Hydrophobic interactions refer to the association or folding of non-polar molecules in aqueous solution. Interactions between ligands and hydrophobic amino acid side chains contribute significantly to the binding free energy. Hydrophobic interactions quantified by the amount of hydrophobic surface buried upon ligand binding have been shown to

correlate well with the binding free energy [6, 7]. However, a lot of this affinity is gained because of the sub-optimal solvation of the binding site in the apo state. The strength of hydrophobic interactions is not caused by an attractive direct force between non-polar molecules. Rather, it results from the system achieving greatest thermodynamic stability by minimising the number of ordered water molecules required to surround hydrophobic portions of the solute molecules. This is the concept of the classical hydrophobic effect (Figure 1.4). The classical hydrophobic effect may be overestimated as hydrogen bonds at hydrophobic surfaces are weaker [8] and water molecules retain a significant amount of residual mobility [9].
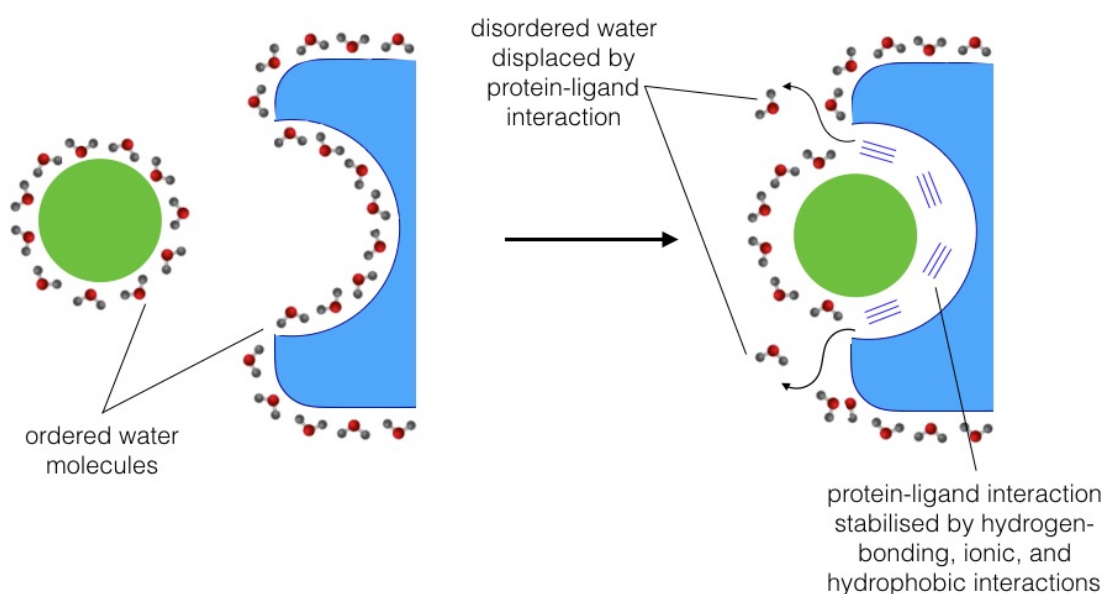
Fig. 1.4 Classic hydrophobic effect. While separate, both protein and ligand force neighbouring water molecules into an ordered shell. Binding of ligand to a protein releases some of the ordered water, and water then forms one larger "cage" structure around the complex. This maximises the amount of free water and causes an increase in entropy. If this mechanism was the only force driving the formation of the complex, all binding events involving hydrophobic ligands would be entropy-driven.

Therefore, it appears that there is no *single* hydrophobic effect. There is a fine balance between enthalpic and entropic contributions arising from the interactions with the solvent. These *effects* have different structural and thermodynamic origins in different molecular contexts, and can be classified as non classical hydrophobic effects [10].

An illustrative example where large gains in binding free energy were achieved when a lipophilic pocket was optimally occupied by nonpolar ligand atoms is optimisiation of the interactions in the S1 pocket of serine protease DPP-IV [11]. In this example, a 400-fold increase in binding affinity was achieved by substituting a meta-phenyl hydrogen atom with a -CH_2F group. The -CH_2F group forms five short hydrophobic contacts with

the residues in the S1 pocket. This subsituent had a 9-fold higher interaction strength compared to the methyl substituent, which can be explained by the slightly larger volume of the -CH_2F group that enables a tighter fit and better shape complementarity (Figure 1.5).
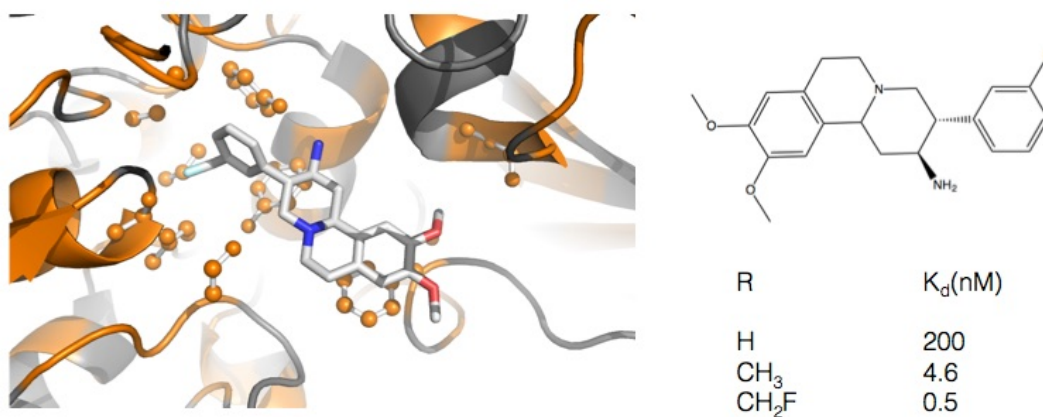
Fig. 1.5 Human DDP-IV with an aminobenzoquinolizone inhibitor (PDB code 3KWJ). Affinity data for the three related inhibitors.

Every ligand binding event displaces water molecules from the binding site. Most of these water molecules are highly disordered and are rarely crystallographically observed. Those that are observed need to be assessed as to whether they can be replaced or whether they are structural and therefore should be considered a part of the complex. Displacement of water molecules from the binding site may affect the resulting thermodynamic signature in a dramatic way. However, the direction of the influence cannot be predicted by simple rules because it is heavily dependent on the context: some proteins become more rigid [12], while others become more dynamic upon water binding [13] and displaced water molecules do not necessarily result in favourable entropy, particularly if they were less ordered in the bound state than they are in the bulk solvent.

In the design of inhibitors of Scytalone Dehydratase, a key target against the pathogenic fungus, *Magnaporthe grisea*, the causative agent of blast disease in rice crops, the inhibitory potency was increased by synthesising compounds where a nitrile moiety was directed into the space occupied by one of the crystallographic water molecules. Replacement of the nitrile with a hydrogen atom lowered binding affinity 18000-fold (Figure 1.6) [14]. Nitrile functionality displaces the water molecule while maintaining the hydrogen bond with the nearby Tyr50 residue.
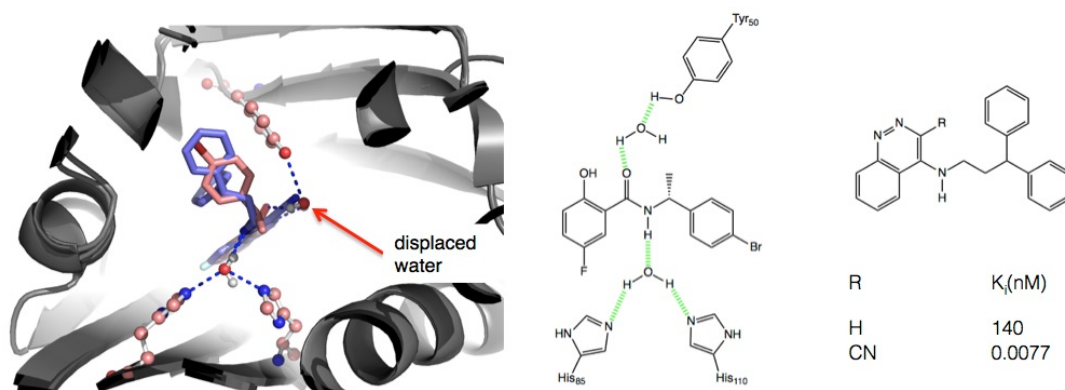
Fig. 1.6 An overlay of the two crystal structures of Scytalone Dehydratase inhibitors, 1STD and 3STD. Replacement of the nitrile with a hydrogen atom decreases binding affinity.

1.2.5 Interactions involving π systems

Aromatic rings can interact by a combination of electrostatic and dispersive interactions [15]. Interactions between the ligands and aromatic protein sidechains (Phe, Trp and Tyr) are very common in protein-ligand complexes. These interactions are strongly orientation dependent, and two arrangements are favoured: one where two rings are parallel to each other, and a perpendicular (T-shaped) arrangement. Quantum mechanical calculations of the dimerisation energy of benzene predict these two arrangements to be isoenergetic (D_e = -2.5 kcal/mol) [16], which is in good agreement with the experimental results (D_e=-1.6 to -2.4 kcal/mol) [17, 18]. An introduction of heteroatoms into the aromatic rings, as well as adding functional substituents affects the alignment of positive and negative charges, modulating π interactions [19]. Aromatic interactions are not limited only to $\pi - \pi$ systems: cation-π interactions have been extensively studied in protein structures [20] and it has been shown that these are rarely buried, contributing -0.8 to -0.5 kcal/mol to the overall binding free energy. This energy drops significantly with increased solvent exposure [21]. With the exception of cation-π interactions, energetic contribution interactions involving π systems make to the overall binding energy is low, but they are important in orientating and positioning the ligand within the binding site.

An interesting example is the S4 binding pocket of serine protease factor Xa, comprised of Tyr, Phe and Trp residues. Introducing a carbonyl group causes a polarisation of the ring CH_2 on top of Trp, increasing the CH_2-π interaction and causing a planarization of the morpholinone ring, bringing it into a perpendicular arrangement to the aryl ring. This effect is reflected by the 60-fold increase in potency (Figure 1.7) [22].
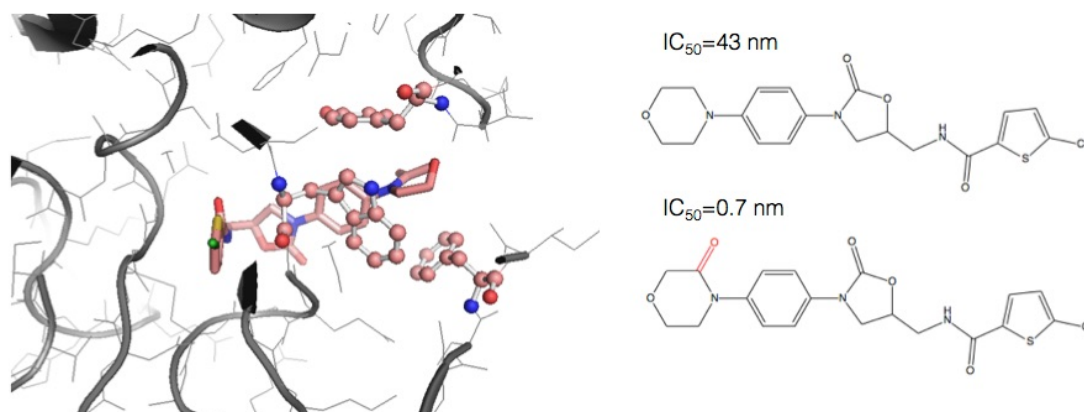
Fig. 1.7 Structures of Xa inhibitors. Introducing a carbonyl group causes a polarisation of the ring CH_2 on top of Trp (PDB code 2W26).

1.2.6 Halogen bonds

The concept of halogen bonds is similar to hydrogen bonds. Both interactions involve a relationship between an electron donor and electron acceptor. In a halogen bond, the halogen atom acts as a donor. Although widely present in protein-ligand complexes, the importance of these interactions in molecular recognition has not been recognised for a long time. This is because halogens were predominantly seen as hydrophobic moieties used to fill apolar protein cavities [23, 24]. The nature of halogens allows for a patch of negative charge to be formed around the central region of the halogen-carbon bond, leaving the outer region positively charged. This partial positive charge is known as the σ-hole [25, 26]. Halogens involved in bond formation are Cl, Br, and I, with the strength of the interaction increasing with the size of halogen electron donor. This type of interaction is mostly electrostatic in nature, and the strength is affected by the electronegativity of the binding partner [27]. Fluorine's small size and extremely high electronegativity prevents it from forming halogen bonds. Instead, it forms the so called 'fluorine bonds' [28]. In general, halogen bonds are fairly weak interactions, but they are specific and can lead to clear gains in binding affinity [3].

A case study of PDE5 and its inhibitors compared the corresponding H, F, Cl, Br and I analogues binding to the catalytic domain of PDE5 [29]. Comparison of the activities shows iodo being the most potent and fluoro the least active, actually worse than hydrogen (Figure 1.8). ITC studies reveal that binding strength of the halogen bonding between chlorine, bromine, and iodine of inhibitor and the protein is -1.57, -3.09, and -5.59 kJ/mol, respectively.
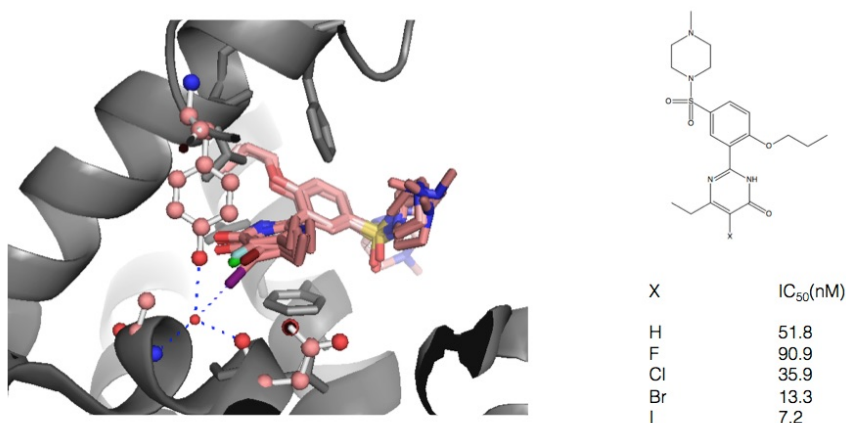
Fig. 1.8 Structures of five inhibitors and their complexes with the catalytic domain of PDE5. Chemical structures of the five inhibitors. PDB codes are 4OEX, 3SHY, 3SHZ, 3SIE, 4OEW.)

1.3 Protein-ligand structure determination

The two most commonly used methods for obtaining structural information on protein-ligand complexes are x-ray crystallography and NMR-spectroscopy. Currently, x-ray crystallography is a superior approach when it comes to providing a detailed picture of the complex at the atomic level. However, the two approaches should be considered as complementary, where the limitations of one technique can be overcome by use of the other.

1.3.1 Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR spectroscopy can provide detailed information about protein-ligand interactions, inaccessible to other techniques. It can be used to study different aspects of biomolecular recognition, such as screening libraries of small molecules, determining 3D structure of the complex, or to determine thermodynamic and kinetic parameters [30]. NMR detects ligand binding through changes in the resonant frequencies called chemical shifts. These changes are highly sensitive to the atom environment, making it possible to determine where the ligand binds and which parts of the molecule are contributing to the binding. The advantage NMR has over other methods is that it quickly delivers information about the binding event, even when the structure of the receptor can not be determined at high resolution [31]. Unlike x-ray crystallography, it allows protein structures to be determined in conditions that are very close to the physiological state (i.e. in solution). The main disadvantages of NMR are that the sample preparation can be much more expensive and that data interpretation is much more complex and time-consuming than x-ray crystallography.

1.3.2   **Protein-ligand crystal structures**

The primary source of structural information for protein-ligand complexes is x-ray crystallography, which captures static snapshots of (sometimes highly) dynamic complexes. Once protein crystals have been obtained, x-ray diffraction measurements can be carried out. This is done either by using in-house equipment or at a synchrotron radiation source. The three-dimensional structure is initially obtained in the form of an experimental electron density map. The map is then interpreted, resulting in an atomic model for the protein structure. Once an atomic model is known, crystal structures of the protein of interest with ligand bound can be prepared and analyzed in a straightforward fashion [32, 33].

X-ray structures are an irreplaceable tool not only in biochemistry, chemistry and related disciplines, but also in computer based methods where they are largely used as starting points on which simple or more complicated calculations are applied. When these structures are used for compound design, a number of assumptions have to be made: 1) the protein structure is correct and known with high accuracy, 2) the experimental conditions under which the crystal was obtained are relevant to the binding event, 3) the ligand structure is accurate and that interactions between the binding partners are correct and well understood [34]. What is often forgotten is that an x-ray structure is a subjective interpretation of an observed electron density map, and as such involves assumptions and sometimes even mistakes [35]. Users who misunderstand this can be misled, especially when studying highly mobile regions of the structure, such as active sites and ligands [36]. Validation of protein structures is relatively straightforward, but small molecules have almost unlimited conformational freedom and chemical variability which are difficult to parameterize, and can be particular to each ligand. This becomes problematic when the structure of the protein is secondary and the real interest is in the ligand. Problems can include: erroneous ligand structures (missing atoms, connectivity issues and incorrect bond orders), geometric constraints (incorrect bond orders and angles), steric clashes between protein and the ligand, conformational errors (cis- or twisted amides, distorted rings, non-planar aromatic groups, etc.), incorrect orientations relative to the binding site, and problems with assigning protonation states and charges [37]. Resolution is a global indicator of quality, but it does not provide any information about the local quality of the protein-ligand model. Another common metric used to measure the overall quality of a model is the R-value, a linear residual measuring the difference between the observed and calculated diffraction data. These two metrics are coupled: a low resolution structure will have higher residuals than a similar well-defined, higher resolution structure [38]. This implies that to assess which parts of the model are strongly supported by the data and which are not, relying exclusively on statistical parameters is not enough. To avoid basing experiments on wrong structural data, users should examine electron density maps

in regions of functional interest. One of the most common metrics for assessing the fit of a model to the local density is the real space-correlation coefficient (RSCC). It ranges from 0 ('bad', electron density is effectively missing) to 1 ('good', model fits the density perfectly) [39].

1.3.2.1   Ligand strain

Potent inhibitors are expected to bind in a relatively low energy conformation [40]. An important question regarding biomolecular recognition is what is the energy price paid by the ligand to adopt this bioactive conformation? The simplest approach to answering this question would be to estimate strain energies using molecular mechanics (MM) via force fields and then calculate the energy difference between the bioactive conformation (the one present in the protein-bound state) and the global energy minimum of the unbound, free ligand. This approach makes an assumption that the crystallographically observed ligand conformation is correct and accurately represented. X-ray derived conformations can be a valid approximation of the bioactive ones, however, what is often overlooked (albeit unintentionally) is that protein-ligand crystal models are interpretations that are several steps removed from the experimental measurements, and as such have lot of uncertainties (the most simple ones regarding bond lengths and angles) introduced during the refinement process, usually dependent on the force field used. The energy of the lowest energy structure (global minimum) can never be verified with absolute accuracy, however, a well-chosen one allows a lower boundary to be placed on the strain energy, because any lower energy structure must increase it. A study by Perola et al. [40] suggested that approximately 60% of the ligands in a set of 150 medicinally relevant complexes bind with strain energies lower than 5 kcal/mol, while 10% of the strain energies were above 9 kcal/mol. A more recent quantum mechanical (QM) study [41] on more than 1200 structures reports this range to be from 0-25 kcal/mol. Several QM/MM [42], and QM [43] methods describing and quantifying ligand strain provide additional evidence about this ubiquitous problem. It has also been pointed out that current force fields do not assess the relative energies of different sized molecules evenly [44]. Different force fields used during structure refinement have different optimal values for bond lengths, which may result in artificially large strain energies. An estimated 25% of recently determined protein-inhibitor complexes have errors that could be removed during crystallographic refinement. These errors are large enough to potentially lead to a wrong interpretation of the binding interactions [45].

Chapter 2

Thermodynamics of drug binding

2.1 Introduction

Understanding the nature of molecular recognition phenomena requires a detailed description of the forces that drive formation of complexes. Since biomolecular recognition is governed by thermodynamics, a quantitative description of the forces that govern molecular associations requires determination of changes of all thermodynamic parameters. Correlating these parameters with the structures of the interacting partners is of great interest to medicinal chemistry since it contributes significantly to structure-based molecular design strategies [7]. Improvements in instrumentation capable of full thermodynamic profiling such as ITC (isothermal titration calorimetry), a technique which has become cheaper and more widely applied in the last decade, have enabled the use of individual thermodynamic binding terms. These provide a useful guide for compound optimisation because they help pinpoint the forces that need optimisation [46].

2.2 Theoretical background

If we consider a simple, reversible binding event

$$P + L \rightleftarrows PL$$

where P is a protein and L a small molecule (ligand), then the change in Gibbs free energy (ΔG) under arbitrary conditions, for PL complex formation is related to standard Gibbs free energy change (ΔG°), under standard conditions (for example, 1M of P, 1M of L at pH 7 and 25 C°) by the following equation which is based on the second law of thermodynamics:

$$\Delta G = \Delta G^\circ + RT ln \frac{[PL]}{[P][L]}$$

At equilibrium and under standard conditions where ΔG=0, this equation becomes

$$\Delta G^{\circ} = -RT ln\frac{[PL]}{[P][L]} = -RT lnK_a = RT lnK_d$$

where R is the gas constant, T is the absolute temperature, K_a is the equilibrium association constant and K_d is the equilibrium dissociation constant. It is more appropriate to write K_d/C_{ref} than K_d. C_{ref} is a standard or reference concentration expressed in the same units as K_d. Therefore, ΔG° depends on the reference concentration. A common value used is 1 molL^{-1}. From now on, C_{ref} is implicit, unless otherwise stated.

Strength of the interactions between the protein and the ligand is usually represented by K_d, K_a or K_i (inhibition constant, the concentration required to produce half maximum inhibition). These values are often referred to interchangeably. Sometimes the IC$_{50}$ (half-maximal inhibitory concentration) value is used instead of the binding constant. This value is defined as the concentration of the ligand required to produce 50% inhibition of protein activity. Unlike K_i, IC$_{50}$ values depend on the concentrations of the enzyme and the substrate used in the enzymatic reaction. In principle, IC$_{50}$ values can be transformed into binding constants using the Cheng–Prusoff equation [47]. As shown above, at equilibrium and under constant pressure and standard conditions, the binding constant can be transformed into Gibbs free energy of binding, ΔG. ΔG is a sum of two energetic terms: the change of enthalpy ΔH and the change of entropy ΔS such that:

$$\Delta G = \Delta H - T\Delta S$$

where T is the absolute temperature. In principle, many combinations of ΔH and ΔS can give rise to the same ΔG value and, therefore, to the same binding affinity to a given target. However, compounds with equal affinities but optimised enthalpically or entropically are not equivalent because the interactions that give rise to favourable binding enthalpies or entropies are different [48].

In a simplified view, the enthalpic contribution to the free energy of binding is related to the specificity and strength of the interactions formed between the protein and the ligand. The entropic contribution to the free energy reflects the dynamics of the whole system. A favourable enthalpic term is obtained through good geometric complementarity between the binding partners. Interactions like hydrogen bonds are stereo specific which implies that favourable enthalpy change is contributing not only to affinity but also to selectivity [49]. Two major terms define entropic contribution: a favourable, but non specific force that is associated with the burial of hydrophobic groups and usually an unfavourable conformational entropy term that reflects the loss of conformational degrees of freedom in

both drug-like molecule and protein. Both the enthalpy and entropy change are dependent on the heat capacity change (ΔC_p):

$$\Delta H = \int_{T_1}^{T_2} \Delta C_p dT$$

$$\Delta S = \int_{T_1}^{T_2} \frac{\Delta C_p}{T} dT$$

A negative value for ΔC_p indicates that the complex has a lower heat capacity. A favourable entropic component is often associated with the release of water molecules from a binding interface, and coupled with a decrease in ΔC_p, it can be used as an indicator of hydrophobic interactions [50, 51].

To achieve high binding affinity (that is, low dissociation constant K_d) both enthalpy and entropy have to contribute constructively. Although simultaneous optimisation of both enthalpy and entropy is desired, in practice this is notoriously difficult to achieve. Even if enthalpic optimisation is achieved, it is often compensated by restriction of mobility of the interacting molecules, which manifests as entropy loss. This phenomenon is referred to as enthalpy-entropy compensation [52].

2.3   Enthalpic considerations

The enthalpic term is a measure of a net change in the number and strength of the protein-ligand interactions on going from the apo to bound state. Therefore, increasing an enthalpic contribution to ΔG seems a sensible strategy for improving affinity and selectivity during the ligand optimisation process. However, when comparing the thermodynamic signatures of the compounds synthesised in medicinal chemistry programs to thermodynamic signatures of the biological ligands, it can be noted that the former have on average a greater favourable entropic component [53]. This is because introducing lipophilic interactions that lack directionality is much easier than controlling the geometries of individual polar groups. Furthermore, attempting to 'build-in' enthalpy by introducing additional polar groups is extremely challenging to achieve, as the breakage of the interactions that water makes causes an energetic penalty that might be hard to overcome. There is also a penalty for desolvating polar groups that is often not compensated by enthalpically favourable bond formation. Even when all these obstacles are overcome, structuring of protein regions induced by newly formed interactions can lead to loss in conformational entropy, resulting in no affinity gains. Although this suggests that a focus on good enthalpic starting points is justified, the examples in the medicinal chemistry literature where the progress from first-in-class to the best-in-class is achieved by improving the enthalpic contribution are anecdotal rather than factual. Even though the thermodynamic profiles of HIV protease,

and HMG-CoA reductase inhibitors show the improvement in the enthalpic contribution from first-in-class (Fluvastatin; Indinavir) to best-in-class (Rosuvastatin; Darunavir) (Figure 2.1), this is no different to detecting a biased coin: flipping a coin 5 times is not enough to enable us to detect bias at 95% confidence level; a coin needs to land the same way at least 8 times out of 9 before we can confidently declare bias (this is further discussed in Section 3). In other words, it is not clear if such entropically or enthalpically driven thermodynamic profiles can be rationally designed.
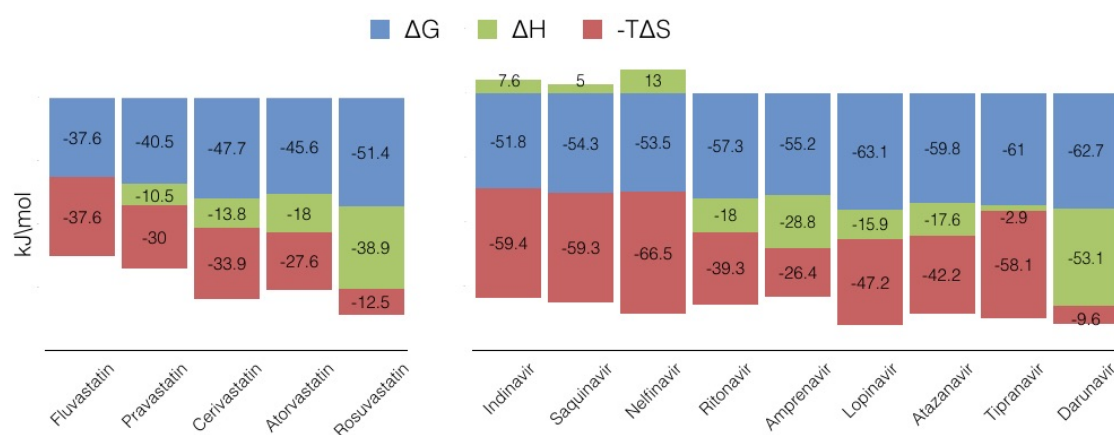
Fig. 2.1 Enthalpic optimisation towards best-in-class compound: thermodynamic profile of HMG-CoA reductase inhibitors; thermodynamic profile of HIV protease inhibitors [54].

2.4 Measuring binding thermodynamics

2.4.1 Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) is a technique capable of directly measuring enthalpy change upon binding. In most ITC instruments, a reference cell is filled with buffer, and a sample cell is filled with the protein into which the ligand is titrated (Figure 2.2). Each injection of the ligand causes heat change that arises from four sources: binding interaction, dilution of both ligand and the macromolecule and a heat effect caused by mixing. The extent of binding is determined by directly measuring heat exchange, whether the heat is being generated or absorbed upon binding: the temperature difference between reference and sample cell is detected by the instrument, triggering a change in the feedback power applied to the sample cell. The heat change is calculated by integrating the heater power over time of the measurement and parameter values (stoichiometry n, K_a and ΔH) estimated by non-linear regression analysis, enabling determination of K_b, ΔG and ΔS [46].
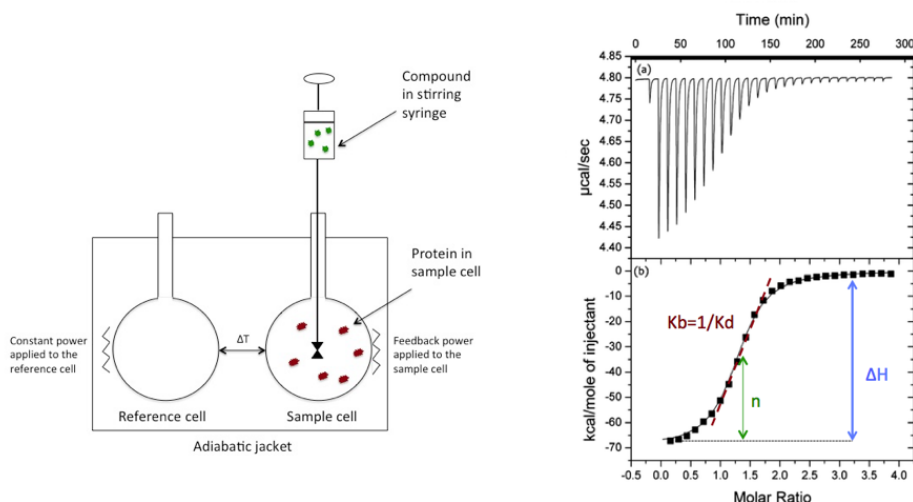
Fig. 2.2 Basic configuration of an ITC instrument. Raw data measured in a typical ITC experiment.

An adequate model has to be applied to fit the titration curve and the results extracted from the ITC experiments should be interpreted carefully. There are four most common models used for fitting ITC data [55, 56]:

- Ligand binding to one molecule with n identical and independent sites
  This is the simplest binding model. It assumes a single independent binding site and a formation of a 1:1 ligand/macromolecule complex. The best fit will generate n, K and ΔH.
- Ligand binding to one macromolecule with m different and independent classes of sites
  This model makes the assumption that there are two binding events that occur with independent thermodynamics and the statistical saturation of each site is dictated by the magnitude of the individual K values. In this case, the best fit will generate n, K and ΔH for each site.
- Ligand binding by the displacement of another ligand in the single binding site of a macromolecule (competitive binding)
  Competitive binding experiments are carried out when for example solubility is an issue, or binding affinities exceed the detection limit of the ITC instrument. Competitive binding studies are carried out using the strong-binding ligand A as the injectant, with the solution in the cell containing the second competitive ligand B as well as the binding protein (or other target molecule). In order to do curve-fitting

on results from a competitive binding experiment, a non-competitive experiment
must first be carried out in the conventional way to determine the binding parameters
for ligand B (n_B, K_B and ΔH_B) itself. These three parameters are used as input
which then allows n_A, K_A and ΔH_A to be determined from results of the competitive
experiment.

- Ligand binding to a macromolecule with two dependent (cooperative) binding sites
  Model assumes that the binding of one ligand influences the binding of the second.
  Association constants will differ by a cooperativity constant α.

By performing the titration experiment at a range of different temperatures, ΔC_p can
be determined [57]:

$$\Delta C_p = \frac{\Delta H_{T_2} - \Delta H_{T_1}}{T_2 - T_1}$$

Each of these parameters reflects the characteristics of the complex minus that of free
partners. In other words, negative ΔG° indicates that the complex is in a lower free energy
state than it is in the sum of the free binding partners, therefore binding is favoured.

2.4.1.1 Protonation effect

All interactions between the ligand and the macromolecule occur in a solvent or buffer
environment and include rearrangement of bonds between the interacting partners as well
as surrounding solvent molecules. A common interaction that takes place in such an
environment is proton transfer between the complex and the bulk solvent. Therefore, calori-
metric experiments result in enthalpies composed of two contributors: reaction enthalpies
and enthalpies of ionisation of buffer. Protein-ligand interactions can be expressed as:

$$P + L + nH^+ \rightleftarrows PL(nH^+; \Delta H)$$
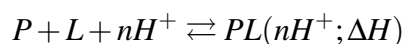
where n represents the number of protons taken up during the process of complex formation.
If the reaction is conducted in a buffer solution, the protons to be taken up are released
from the buffer:

$$BH^+ \rightleftarrows B + H^+; \Delta H_i)$$
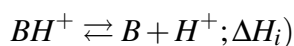
thus, in an actual calorimetric measurements the observed enthalpy change is given as:

$$\Delta H_{obs} = \Delta H + n\Delta H_i$$

This means that to correctly determine the net enthalpy change ΔH, a contribution from the
buffer side on the observed heat effects has to be subtracted from the observed enthalpy
change. To calculate the appropriate correction it is necessary to perform a calorimetric
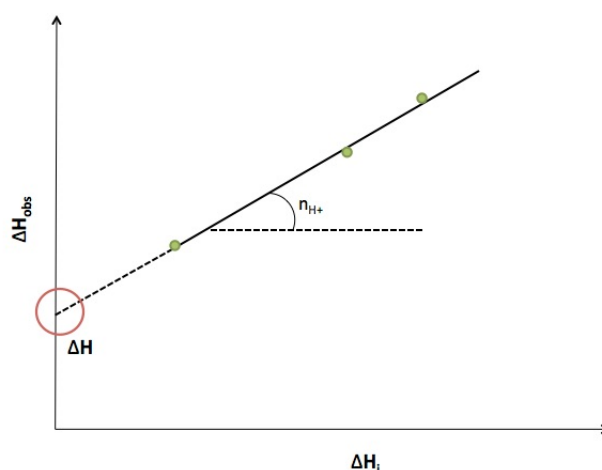
Fig. 2.3 ΔH_{obs} vs ΔH_i:variation of observed binding enthalpy with various buffers

measurement with at least two buffers of different ΔH_i at the same pH. The ΔH_{obs} is then plotted against the ΔH_i, enabling the determination of *n* and ΔH (the slope and the intercept respectively), the actual binding enthalpy (Figure 2.3) [58, 59].

If protonation changes occur uniformly across the congeneric ligand series, or the protonation state is relatively distant to the site where ligands are modified, relative differences between the thermodynamic signatures can still be interpreted conclusively. For example, compensating effects arising from deprotonation of His and protonation of the primary amino function of the ligand were observed in the ITC study of a congeneric series of thrombin inhibitors, so similar behaviour for the analogous inhibitors was assumed [60].

ITC experimental setup, interpretation of thermodynamic signatures, and the assessment of the protonation effect will be discussed in chapters 7 and 8.

2.4.2 Estimation of thermodynamic parameters from changes in protein stability

Binding of a ligand to a protein occurs only if it represents an energetically favourable process, that is, if it leads to the release of free energy. Therefore, the protein-ligand complex has to be more stable than the free partners, with the extent of stabilisation dependent on the magnitude of the binding energy. Therefore, it is possible to estimate the free energy of binding by comparing the differences in stability of the complex and free partners.

The simplest model of protein denaturation is:

$$P \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} U$$
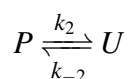
where P is folded protein in apo form, U is the unfolded protein, k_2 and k_{-2} are the rate constants for folding and unfolding. The Gibbs free energy change for the reversible unfolding without the presence of the intermediates is then given by

$$\Delta G_{U-P} = G_U - G_P$$

.

In the presence of the ligand, the unfolding scheme becomes:

$$P'L' \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} P + L \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} U$$
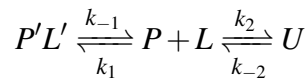
and the observed Gibbs free energy change is now

$$\Delta G_{obs} = \Delta G_{U-P} - \Delta G,$$

from which it is then easy to determine ΔG.

Thermodynamic characterisation of protein stability is usually performed by thermal denaturation. Differential Scanning Calorimetry (DSC) [61] is the method in which the heat capacity of the system is monitored as function of temperature. This method can be used to determine the melting temperature, T_m, change in heat capacity, ΔC_p, and the enthalpy of unfolding, ΔH. In an equilibrium state, the change in these parameters can be used to estimate binding affinity.

2.4.3 Indirect thermodynamic measurements

ITC measures enthalpy change upon binding directly. When performing this measurement is not possible or practical (solubility issues, reagents or instrument availability, etc.), this parameter can be accessed indirectly, by determining van't Hoff enthalpy (ΔH_{vH}) through the temperature dependence of K_d as defined by the van't Hoff relationship

$$ln\frac{K_{d_{ref}}}{K_d} = \frac{\Delta H_{vH_{ref}} - T_{ref}\Delta C_p}{R}\left(\frac{1}{T_{ref}} - \frac{1}{T}\right) + \frac{\Delta C_p}{R}ln\frac{T}{T_{ref}}$$

where K_d is the dissociation constant and ref refers to any arbitrary reference temperature. The potential problem with this approach is the inclusion of the ΔC_p term which allows for the temperature dependence of ΔH. This term is often difficult to estimate in the absence of calorimetric data. Values of ΔC_p and ΔH are correlated and the magnitude of the changes in ΔG with temperature are usually relatively small, leading to discrepancies between calorimetric and van't Hoff enthalpies [46].

2.5   Ligand efficiency metrics-LEMs

For any small molecule to be considered a likely drug candidate, it must satisfy a number of different properties such as size, lipophilicity, shape, number of hydrogen bond donors or acceptors, polarity, toxicological profile etc [62, 63]. It has been observed that these properties for orally bioavailable drug candidates occupy a relatively narrow range known as 'drug-like' space [64]. Ligand efficiency metrics are used to quantify the link between physicochemical properties and potency. They account for the contribution of a risk factor such as molecular weight or lipophilicity to activity [65, 66]. The rationale is that this would allow for a more relevant comparison of compounds than would be possible just by using affinity or potency measures [67]. LEMs are typically defined by scaling or offsetting measured activity by the value of the physicochemical property that is considered to be a risk factor. For example, it is assumed that an increased lipophilicity or molecular size can lead to poor physical properties and compound promiscuity. Affinity scaled by molecular size gives Ligand Efficiency (LE), a metric that aims to compare molecules according to their average binding energy per atom [68]: $LE = \Delta G^\circ / HA$ where ΔG° is standard Gibbs energy and HA number of heavy atoms. Offsetting potency or affinity by lipophilicity gives Ligand Lipophilic Efficiency (LLE) [69]. LLE quantifies the moving of a compound from octanol to its binding site: $LLE = pIC50 - (c)logP$. In other words, it describes the contribution of lipophilicity to potency.

   LEMs are widely accepted and used by the drug discovery community, often ignoring underlying assumptions [70–72]. Both scaling and offsetting assume that the relationship between the activity or potency and the physicochemical property is linear. In the case of offsetting, it is assumed that the relevant line has a unit slope, whereas scaling implies that the relevant line passes through the origin, which suggests that affinity or potency values in all assays are zero for ligands with zero atoms. It is important to remember that values of zero for ΔG° or IC_{50} do not indicate an absence of interactions between protein and the ligand [73]. A zero value simply is a statement that affinity or potency is equal to standard or reference concentration:

$$\Delta G^\circ = RT ln(\frac{Kd}{C^\circ})$$

$$pIC_{50} = -log(\frac{IC_{50}}{M})$$

   In both of these expressions K_d and pIC_{50} have been divided by a concentration because the *log* function is defined only for dimensionless units. Therefore, it is important to remember that molarity is built into both of these definitions and that the choice of 1 M as a reference concentration is arbitrary and historical in nature. Consequences of this are best illustrated using the example given in table 2.1:

Table 2.1 Effect of standard concentration on ligand efficiency

N_{HA}	K_d/M	$C°$/M	LE
10	10^{-3}	1	0.40
20	10^{-6}	1	0.40
30	10^{-9}	1	0.40
10	10^{-3}	0.1	0.27
20	10^{-6}	0.1	0.33
30	10^{-9}	0.1	0.36
10	10^{-3}	10	0.54
20	10^{-6}	10	0.47
30	10^{-9}	10	0.45

This example shows how the perception of efficiency changes when the concentration unit with which affinity or potency is specified varies. When a reference concentration of 1 M is used, the three compounds are all equally ligand efficient. If standard C is 0.1 M, the compounds appear to become more ligand efficient as molecular size increases but the opposite behaviour is observed when standard C is 10 M.

Even though LLE is invariant with the choice of reference state (unlike standard concentration, reference concentration has no thermodynamic significance), the unit slope is arbitrary in the same way. It is valid to question whether slope=1 is a better metric than slope=0.7, for example. There is no reason that normalisation of activity should be restricted to any of these functional forms. Rather than imposing constraints to make assumptions about the trends in the data, the relevant line should be fit to establish the actual trend. This way, LEM can be thought of as an attempt to quantify the extent to which compounds beat (or are beaten by) the actual trend in the activity data, or, in other words, which compounds have higher (or lower) activity than it would be expected given the physicochemical property. Residuals give the difference between observed value and value predicted by the model and do not change with the choice of concentration units and are independent of molecular size.

The driving forces in ligand binding and the relation to LLE will be further explored in chapter 5, where it will be explained how and why this property can be used in computational modelling.

Chapter 3

Kinetics of drug binding

3.1 Introduction

At the beginning of the twentieth century, Paul Ehrlich coined a phrase that soon became a postulate of modern pharmacology and molecular medicine: *corpora non agunt nisi fixate* ("a substance will not work unless it is bound") [74]. In other words, a drug is efficacious as long it remains bound and modulates the activity of its receptor. Target affinity is considered to be a proxy for an *in vivo* efficacy, focusing much of an early stage drug discovery efforts on optimising the selectivity and affinity of the binary complex.

Fighting disease or its symptoms using drugs can be described as a continuous kinetic process, easily divided into two parts: pharmacokinetics (PK) which determines how the concentration of drug changes over time (sometimes described as what the body does to the drug) and pharmacodynamics (PD), a term that describes what the drug does to the body [75].

Binding kinetics (BK) represent how quickly a drug-target complex forms (described by a rate constant of ligand association-k_{on}) and dissociates (described by a rate constant of ligand dissociation-k_{off}). The affinity of interaction is represented by the equilibrium dissociation constant K_d, a value that is dependent on both rate constants, according to the equation:

$$K_d = \frac{k_{off}}{k_{on}}$$

The steps that determine how quickly the complex will form, like diffusion or desolvation, are difficult to affect, making optimisation of the association rate a difficult task. The rate at which the protein-ligand complex forms also depends on the concentration of the ligand. On the other hand, the dissociation rate constant is not affected by pharmacological factors or ligand concentration and is only dependent on the specific interactions between the binding partners. From this equation it can easily be seen that different combinations

of k_{off} and k_{on} can result in the same affinity. In most situations, the duration of biological effect induced by formation of the drug-target complex is related to the lifetime of the complex: the longer the ligand is bound to its receptor, the longer the biological effect lasts [76]. Given a comparable K_d, a ligand's efficacy (ability to produce biological response upon binding) is determined by k_{off}, a constant that is concentration-independent, but is dependent on specific interactions between ligand and protein. This implies that it is the lifetime of the receptor-ligand complex that determines the effect in the cellular context; when the residence time (reciprocal of the dissociation rate constant; $1/k_{off}$) of the complex is long, a significant level of receptor occupancy can be preserved even when the concentration of the ligand in the system is reduced. A good real life example is given by Copeland et al. [77]; they compare residence time to ones residence time at a hotel. The price for a hotel stay is not influenced by how long it takes for the guest to arrive at the hotel but by the time between check in (analogous to the bound state) and check out (analogous to dissociation).

3.2 General binding mechanisms

There are three general binding mechanism that can lead to long residence times. The simplest mechanism is a formation of PL binary complex
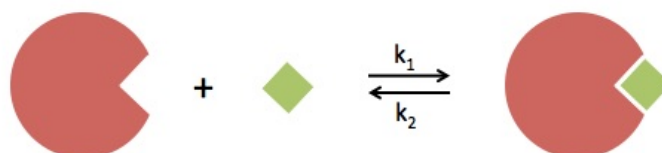
in which k_1 is equivalent to k_{on} and k_2 is equivalent to k_{off}. In the second mechanism, the protein exists in an ensemble of conformations, some of which are capable of binding the ligand. Presence of the ligand will shift the equilibrium between the conformations, favouring the formation of the conformation capable of binding the ligand:
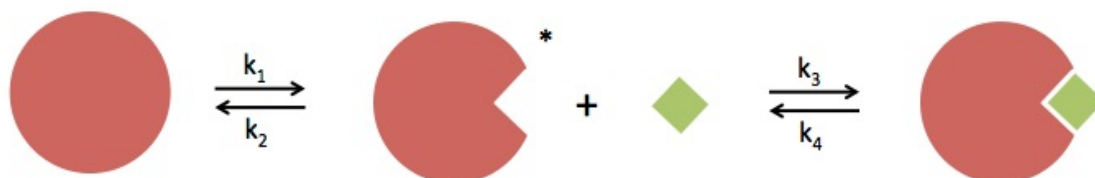
In this model, the rate-limiting step is the interconversion of the two forms of the protein, P and P^*. Once the P^* forms, the ligand binds rapidly. k_1 represents the rate

constant of this interconversion and k_2 is the constant for the reverse process. As only the P^* conformation can bind the ligand, K_d^* is defined as the ratio of k_4 and k_3.

In the third mechanism, the protein exists in a single conformational state with suboptimal complementarity to the ligand, but is still capable of ligand binding. Upon the initial encounter with the ligand, the protein undergoes a conformational readjustment:

In this model, the rate limiting step is the slow isomerisation from the PL state to P^*L. The binding affinity in this model reflects the affinity of the final P^*L state and is defined as $K_d^* = K_d/(1+(k_3/k_4))$.

3.3   Long or short residence time

General binding mechanisms can sometimes be hard to distinguish experimentally. Most drug-like molecules bind to their targets through the induced-fit mechanism (third model) [78]. Regardless of what the association pathway is, the definition of residence time is the same:

$$\tau = \frac{1}{k_{off}}$$

The time it takes for 50% of the protein-ligand complex to decay is the half-life of the complex:

$$t_{1/2} = \frac{ln2}{k_{off}}$$

Both residence times and half-lives are measures of the lifetime of the protein-ligand complex.

One of the crucial considerations during early-stage drug discovery is whether to aim for long or short residence times, mainly because a long residence time can have both beneficial and detrimental outcomes. As a general rule, short residence time is required when a ligand needs to perform routine physiological functions. A prominent example is Memantine, an N-methyl-D-aspartate (NMDA) receptor antagonist, which is an open channel blocker, capable of fast dissociation from the receptor, thus enabling the receptor to quickly return to its resting state [79]. Other examples include Dopamine-2 receptor antagonists, such as Risperidone, a typical antipsychotic, for which it has been established that mechanism-based toxicity can be avoided by means of rapid and competitive

dissociation rates and cyclooxygenase inhibitors Ibuprofen and Naproxen, often used as anti-inflammatory agents. Ibuprofen and Naproxen are rapid reversible inhibitors, whereas their analogue, acetylsalycilic acid (Aspirin) has an irreversible binding profile with concomitant increased risk of bleeding [80]. The majority of the marketed drugs have a long residence time, as this is considered to be more efficient. Long residence time is the best strategy when therapies require long lasting target occupancy (maintaining high blood levels for a long time is the alternative). The foremost example is probably Darunavir - the HIV protease inhibitor most recently approved by the FDA. What differentiates Darunavir from its analogues is an extremely long residence time and this property has been linked to the higher antiviral activity and higher genetic barrier to development of resistance [81, 82]. Darunavir also has the highest enthalpic contribution to the overall free energy of binding and it is intuitive to think that the specific interactions that contribute to enthalpy also contribute to the long residence time: ligand dissociation requires increased mobility for both the protein and ligand and will thus be associated with a favourable entropy component. Therefore, the energy barrier is likely to be almost exclusively enthalpic [83].

Until recently, routine use of kinetic data in drug discovery processes has not been common. However, the measurement of accurate kinetic data for protein-ligand interactions has been enhanced in recent years by the use of sensitive, high-throughput label free instrumentation, making drug-target residence times a valuable parameter in the drug discovery process [84] (Figure 3.1).
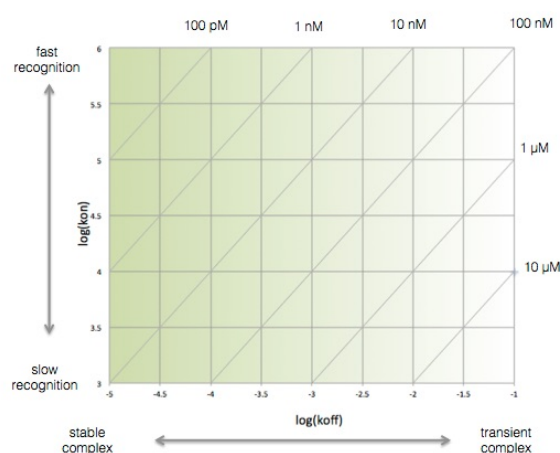
Fig. 3.1 Association rate constant k_{on} plotted against dissociation rate constant k_{off}. Diagonal lines in the plot represent equilibrium dissociation constant (K_d) values.

3.4 Measuring binding kinetics

The increased interest in optimising binding kinetics has been accompanied by an improvement in the related instrumentation capable of providing binding rates. In theory, any technique capable of measuring and differentiating between the amount of bound and free ligand can be used to determine *on* and *off* rates. In practice, measuring binding kinetics accurately and reproducibly can be very challenging. The most widely used techniques include label-free methods such as surface plasmon resonance and techniques using labelling and fluorescence.

3.4.1 Surface plasmon resonance

Kinetic characterisation using surface plasmon resonance (SPR) is one of the most established methods for measuring real-time quantitative binding affinities and kinetics. In this label-free technique, a protein of interest is immobilised to the sensor surface while analyte is free in solution and passed over the surface. Binding of an analyte to an immobilised protein induces a change in the refractive index on the sensor surface: this change is linear to the number of molecules bound. A sensorgram depicts changes in SPR angle in real time, with responses measured in resonance units (RU). When the pulse of analyte (drug) is long enough, steady state can be reached (flat part of the sensorgram curve). At the end of the injection, the analyte is replaced by a continuous buffer flow, and the corresponding decrease in signal now reflects dissociation of analyte from the surface-bound protein (Figure 3.2) [85].

K_d is determined by plotting the response units (RU) at equilibrium against the ligand concentration. Rate constants k_{off} and k_{on} are estimated by regression of association and dissociation gradients measured at different analyte concentrations. The simplest 1:1 binding mechanism is applicable in the vast majority of cases. In this model, one analyte molecule from solution binds to immobilised protein. Analysis of the sensorgram curve in the association phase, in which binding is measured while the analyte solution flows over the ligand surface, allows the determination of the rate of complex formation. R_t and R_{max} correspond to the observed response at time *t* and to the theoretical maximal response that would be observed if an infinite concentration of analyte A was injected [86, 87]:

$$\frac{d[R]}{dt} = k_{on}[A](R_{max} - R_t) - k_{off}R_t$$

$$R_t = \underbrace{\frac{R_{max}[A]}{k_{off} + [A]}}_{\text{Determines the equilibrium level}} \times \underbrace{[1 - e^{-(k_{on}[A] + k_{off})t}]}_{\text{Determines the time to reach the equilibrium}}$$
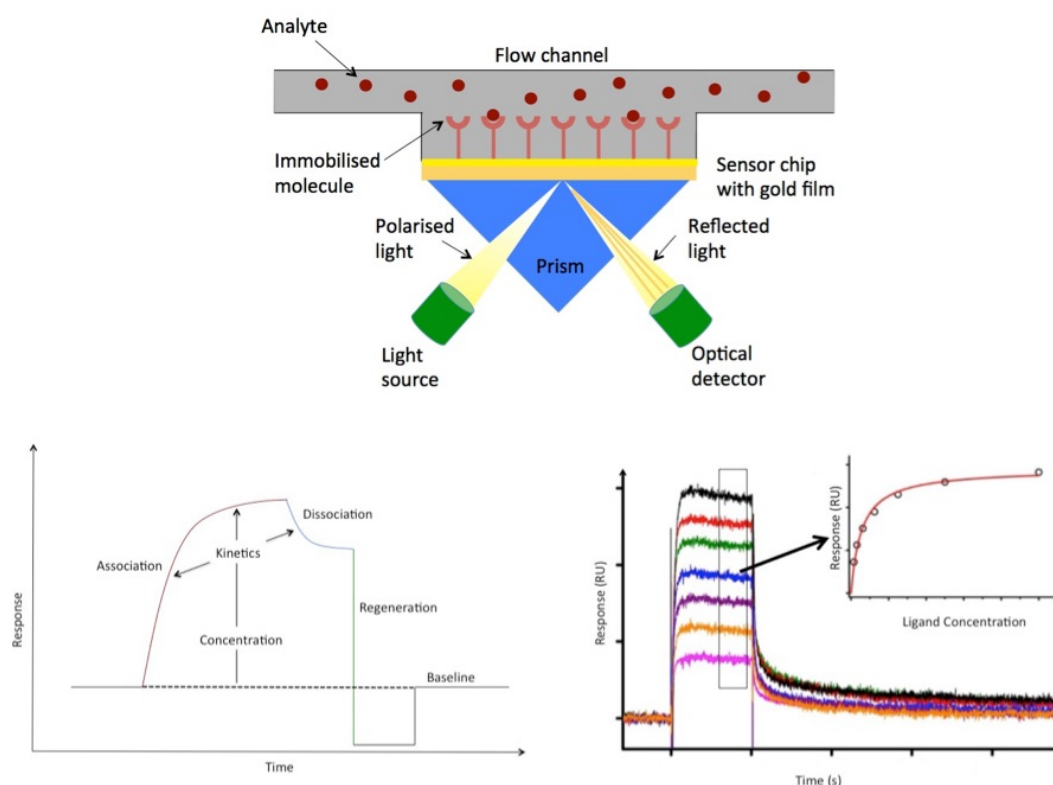
Fig. 3.2 Surface plasmon resonance detects changes in the refractive index in the immediate vicinity of the surface layer of a sensor chip. A typical SPR sensorgram. K_d is determined by plotting the response units (RU) at equilibrium against the ligand concentration.

In the dissociation phase, the concentration of analyte in the flow is suddenly reduced to zero by the injection of the running buffer. The rate of complex dissociation follows simple exponential decay, or first-order kinetics:

$$R_t = R_0 \times e^{-(k_d t)}$$

where R_0 is the signal level at the beginning of dissociation.

When a poor fit is obtained to the data using the simple 1:1 binding model the binding kinetics are considered complex. Because all complex binding models generate equations with two or more exponential terms, it is usually impossible to distinguish between different models by curve fitting alone and additional experiments have to be performed [88]. There are four complex binding models for analysing non-Langmuir interactions: the heterogeneous analyte, heterogeneous ligand, two-state, and bivalent analyte models. It is necessary to provide a biological justification for the use of other models, and conclusions based on analyses with these complex models should be confirmed with additional experiments [89].

SPR has a short assay development time, low material consumption and can be used to determine kinetic parameters in parallel mode. However, the immobilisation of the receptor could have an effect on the binding properties. The method also has relatively low throughput due to the slow regeneration step. The dynamic range for rate constants determined by SPR is limited: the limits for k_{on} are between 10^3 and 10^7 M^{-1}s^{-1}, and for k_{off} between 10^{-6} and 10^{-1} s^{-1}. It is almost impossible to determine the rate constants for very weak interactions where quick association is caused by high analyte concentration. These complexes usually have k_{off} below the limit that can be determined.

The experimental setup, interpretation of kinetic signatures, and the assessment of the residence time will be discussed in chapters 7 and 8.

3.4.2 Radioligand binding

Radioligand binding is a preferred technique for measuring binding kinetics of G-protein coupled receptors [90, 91]. There are two main approaches to measure kinetics using this method. The first approach is direct radiolabeling where receptors are pre-incubated with the radiolabeled ligand to reach equillibrium. Further binding of the radioligand to receptor is then blocked by adding an excess amount of assay buffer or saturating concentration of unlabeled ligand to prevent reassociation. After initiating dissociation, off rate is measured to determine how rapidly the ligand dissociates from the receptors. The association is determined in a separate experiment at different radioligand concentrations, or by performing an experiment at single a concentration when k_{off} is known. The second approach is indirect. The kinetic properties of a compound are determined through competition displacement by a radioligand of known affinity [92]. The high-throughput variant is a dual-point competition association assay which measures radioligand binding at two different time points in the absence or presence of unlabeled competitors. If the unlabeled competitor dissociates faster than the radioligand, the competition curve will monotonically reach its equilibrium. If however, the radioligand dissociates faster, the association curve will have two phases [91].

Even though radioligand binding assays are capable of determining kinetic rate constants, the experiment is expensive and generates radioactive waste. Alternative approaches are fluorescence methods, such as time resolved fluorescence resonance energy transfer (TR-FRET) [93] or fluorescence anisotropy [94].

Chapter 4

Computational approaches to molecular recognition

4.1 Introduction

Computational chemistry can be described as chemistry performed using computers rather than chemicals. It is usually used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. Much like the experiment, calculations can be applied in order to collect the data or simply as a way of looking at the unusual.

Although performing calculations is a lot cheaper than experiment, the cost of calculations rapidly increases with the molecular size. Based on the underlying theory and applicability domain, computational chemistry can be divided into three main approaches, as summarised in table 4.1.

Over the last two decades powerful molecular modelling tools have been developed which are capable of accurately predicting structures, energetics, reactivities and other properties of molecules. These developments are largely due to the dramatic increase in computer speed and the design of efficient algorithms. Accurate calculations are still computationally very demanding and it is often advantageous to calculate the differences in binding energies throughout the ligand series, rather than absolute values [33, 95]. This information can then be used to guide molecular design and synthesis. Ideally, the computational approaches would be as accurate as experiment, with binding energies correct to within 1 kcal/mol. This goal has not yet been achieved. There are many reasons why accurate prediction of free energy of binding is difficult. These estimates are small differences between large numbers and the interactions between binding partners depend on the details of molecular conformation and both ligand and protein have many degrees of freedom, which makes exploration of conformational space a challenging task [96].

Table 4.1 Computational chemistry methods

Method	Advantages	Disadvantages
Molecular Mechanics	• relies on force-field with embedded empirical parameters • applicable for large systems computationally least intensive	• force-field dependent • requires experimental data • restricted to equilibrium structures
Semi empirical	• requires experimentally derived empirical parameters • less demanding than ab initio methods • suitable for studying transition states	• computationally expensive • less rigorous than ab initio methods • requires experimental data
Ab initio	• does not require experimental data • suitable for studying transition states • useful for determination of properties that are inaccessible experimentally	• computationally very expensive • applicable only for smaller systems

Furthermore, contributions from both enthalpic and entropic changes to the overall binding free energy are global parameters that reflect the overall heat effect caused by energy exchange between all the species in the reaction mixture [97]. To deconvolute these contributions to individual terms is a very challenging task. Despite all the difficulties, computational modelling has a significant potential to make accurate predictions and guide molecular design.

4.2 Approaches to calculating affinity

4.2.1 Docking and scoring

Docking methods aim to identify the predominant binding mode(s) of a ligand with a protein of known three-dimensional structure. This conformation is then used to assign a binding energy or a binding score. The method must be computationally rapid so thousands of molecules can be docked and ranked per hour. Scoring functions are used to rank different binding modes and can be empirical, force field based, or knowledge based [98, 99]. In an empirical scoring function, binding energy is decomposed into several

terms, such as hydrogen bonds, ionic interactions, hydrophobic effect and binding entropy, which are then summed up to give a final score. The force-field-based scoring functions calculate the sum of van der Waals and electrostatic energy terms. Knowledge-based scoring functions use the assumption that the more favourable an interaction is, the more frequently they will appear in the available structures of protein-ligand complexes. The score is calculated by rewarding preferred contacts and penalising repulsive interactions between each atom in the ligand and protein within a given cutoff.

Table 4.2 Examples of Docking and Scoring approaches

Scoring Function

	Description	Examples
Empirical	sum of individual contributions to binding energy	LUDI [100], ChemScore [101], GlideScore [102]
Force field based	sum of van der Waals and electrostatic energy term	Dock [103], Gold [104], Autodock [105]
Knowledge based	derived from crystal-structure data	PMF [106], DrugScore [98]

Ligand Conformational Space Search

Systematic	uniform sampling of search space
Stochastic	random changes to a single ligand or a population of ligands
• Monte Carlo	Prodock [107], ICM [108], MCDOCK [109], DockVision [110]
• Genetic Algorithm	GOLD [104], AutoDock [105], DIVALI [111], DARWIN [112]

In theory, search space for protein-ligand binding should consist of all possible unbound ligand and protein conformations, all possible ligand conformations within a given protein conformation, and all possible protein conformations with all possible ligand conformations. In practice, it is impossible to exhaustively explore this search space with the current technology. Therefore, to satisfy both speed and search space coverage criteria, two major approaches have been developed: systematic and stochastic. Systematic searching is deterministic and includes sampling at predefined intervals, while stochastic searching samples the conformational space by performing random changes to a single ligand or a population

of ligands. There are two basic types of methods based on random algorithms: Monte Carlo methods and Genetic Algorithm methods [113, 114]. In Monte Carlo methods, an initial configuration is refined by taking random steps which are accepted or rejected, based on a Boltzmann probability function, until a certain number of steps have been tried. The genetic algorithm approach is inspired by biological evolution. This approach starts from an initial population of different conformations of the ligand with respect to the protein. The score of each pose acts as the fitness function used to select individuals for the next iteration, until the final population is reached. Most common examples of docking and scoring approaches are given in table 4.2

Based on how broadly the method explores the search space, the approaches fall into two categories: local or global. Local search methods aim to find the nearest or local energy minimum to the starting conformation, whereas global methods search for the best or global energy minimum within the defined search space. Hybrid global–local search methods have been shown to perform even better than global methods alone, being more efficient and able to find lower energies [115].

4.2.2 Molecular mechanics

There are several methods for computing the potential energy of an isolated molecule or a complex in a given conformation. Empirical methods substitute fitted parameters of electronic structures in a few energy terms that are easily calculated. This is molecular mechanics or force field approach and the underlying principle here is that cumulative physical forces can be used to describe molecular geometries and energies. The building blocks are atoms and electrons are not considered as individual particles. A molecule is thought of as a collection of masses centered around the nuclei (atoms) connected by springs (bonds). A potential energy is constructed as a function of atomic positions, using a large body of experimentally determined, parameterised data. The most commonly used examples include CHARMM [116], AMBER [117], GROMOS [118] and OPLS [119].

Force field energy can be written as a sum of terms, each describing the energy required for distorting a molecule in a specific fashion:

$$E_{FF} = E_{str} + E_{bend} + E_{tors} + E_{vdW} + E_{el} + E_{cross}$$

The first two terms treat bond stretching and angle bending, the third term is the energy for rotation around a bond, E_{vdW} and E_{el} describe the non-bonded interactions. E_{cross} describes the coupling between the first three terms (Figure 4.1) [120].
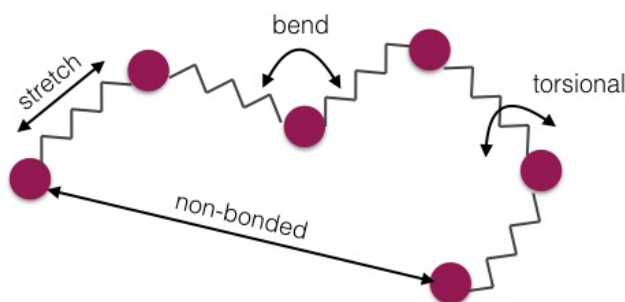
Fig. 4.1 Force field energy terms

4.2.3 Endpoint methods

End-point methods involve sampling of the ligand and protein in bound and free (unbound) state and then taking the difference between the absolute energies of the two states to estimate the binding free energy [121].

The simplest example is the Linear Interaction Energy (LIE) approach [122, 123]. LIE involves running two molecular dynamics simulations: one for the ligand in solution and the other for the ligand in the protein binding site. The binding free energy is estimated based on the difference in the energetics of the interactions between ligand and receptor in the "bound" versus "free" states (i.e. the endpoints):

$$\Delta G_{bind} \approx \beta(\langle E_{elec}^{bound}\rangle - \langle E_{elec}^{free}\rangle) + \alpha(\langle E_{vdW}^{bound}\rangle - \langle E_{vdW}^{free}\rangle) + \gamma$$

The main assumption, deriving from linear response considerations, is that only (convergent) averages of the interaction energies between the ligand and its surroundings need to be evaluated [124]. The angle brackets indicate ensemble averages and their superscripts indicate from which simulation the average is computed. α, β and γ are the coefficients that account for the internal energy of the solvent and the protein and are dependent upon the force field, polarity of the small molecule, and the hydrophobicity of the binding pocket [125].

In second end-point approach, molecular mechanics energies are combined with the Poisson–Boltzmann (MM-PBSA) or generalized Born (MM-GBSA) and surface area continuum solvation to calculate binding free energies. Three molecular dynamics simulations are carried out: free ligand, apo protein and their complex [126, 127]. Binding free energy is then calculated as a difference between the complex, free ligand and the unbound receptor:

$$\Delta G_{bind} = \langle G_{PL} \rangle - \langle G_P \rangle - \langle G_L \rangle,$$

The free energy of each state is approximated from the sum of six terms:

$$G = E_{bnd} + E_{elec} + E_{vdW} + G_{pol} + G_{np} - TS_{MM}$$

where the first three terms are the molecular mechanics energy (bonded interactions, electrostatic and van der Waals interactions), G_{pol} is polar solvation energy from PB or GB, G_{np} is nonpolar continuum-solvation energy contributions and TS_{MM} is the solute entropic contribution.

LIE and MM/PBSA have been compared several times, but the performance varied strongly with the tested system [128–131]. MM/PBSA often overestimates differences in binding affinities, giving a favourable correlation coefficient, but a poor median absolute deviation, MAD [132].

4.2.4 Pathway methods

In pathway methods the ligand is gradually separated from the protein. The initial and final states are linked by a pathway of small steps for which the difference in free energy can be calculated. This class includes methods in which the ligand/protein interactions are gradually switched off, such as thermodynamic integration (TI) and free-energy perturbation (FEP). Both techniques use thermodynamic cycles to estimate free energy changes, $\Delta\Delta G$, between two ligands. One ligand is perturbed into another with the help of a coupling parameter in both the solvent and protein environment. This method is termed computational alchemy [133]. A free energy cycle shows that the difference between the free energies of binding $\Delta G_{bind}A$ and $\Delta G_{bind}B$ equals the difference between the binding free energies of the two ligands, ΔG_2 and ΔG_1 (Figure 4.2). Calculating $\Delta G_{bind}A$ and $\Delta G_{bind}B$ involves simulation on a physical pathway (horizontal lines) and can be very difficult, as it involves defining a reaction coordinate. Since the free energy depends only on the state and not the pathway, this difference can be calculated along the non-physical pathway (vertical lines). Ligand A is transformed to ligand B, both in the solution and in the context of the protein. ΔG_1 results from transforming ligand A to ligand B in solution, while ΔG_2 results from the same transformation, this time in the binding site. Thus, relative change in free energy of binding between two related ligands can be calculated

$$\Delta\Delta G_{bind} = \Delta G_{bind}B - \Delta G_{bind}A = \Delta G_2 - \Delta G_1$$

.
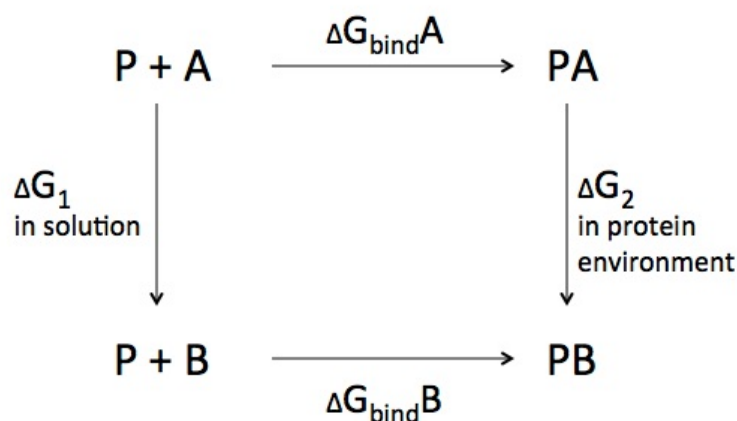
Fig. 4.2 Thermodynamic cycle

A widely used approach to calculating free energy changes associated with an alchemical transformation is thermodynamic integration, TI. In this approach, the energy of the system is a function of a parameter λ. λ varies from 0 in the initial state to 1 in the final state. Usual procedure includes simulations at several λ values and computing the numerical integral at the each step. At each intermediate step λ_i, a short "equilibration" run is first performed (since the point of equilibrium has changed) and then a "production" run where the integral is computed. The number of windows between λ=0 and λ=1 has to be decided in advance, but the data can then be divided into equilibration and production parts later. It is possible to accumulate runs so no data are wasted.

In another approach, free energy perturbation (FEP) simulation of the system is performed at λ and $\lambda + \Delta$ steps. This way, an energy change associated with each step is obtained, and the small free energy difference between the two states can be determined. The number of windows and the equilibration time have to be decided in advance. The windows are created serially, so if the equilibration time is insufficient, the simulation has to be repeated. Both approaches are conceptually shown in figure 4.3

The double decoupling method involves simulation of two alchemical processes [134]. In the first one, the interactions between the protein and the ligand are switched off, while the ligand is restrained in the binding site. In the second step, the restraint is removed.

All of these methods are widely used to evaluate the differences in the free energy of binding between a series of closely related ligands [135–140]. However, all of these techniques suffer from the same drawbacks. Insufficient sampling, limited accuracy of the current force fields and the fact that these methods do still require vast computational resources are the main factors that hamper routine use of these approaches. The most common parameter used to test validity of the force field description of the protein is the structural conformational preference, which may not be sufficiently accurate for the use in

$$\Delta F_{AB} = \pm kTln\langle e^{-[H(\lambda+\delta\lambda)-H(\lambda)]\backslash kT}\rangle_{\lambda}d\lambda \qquad\qquad \Delta F_{AB} = \int_{\lambda_A}^{\lambda_B}\langle\frac{\partial H(\lambda)}{\partial\lambda}\rangle_{\lambda}d\lambda$$
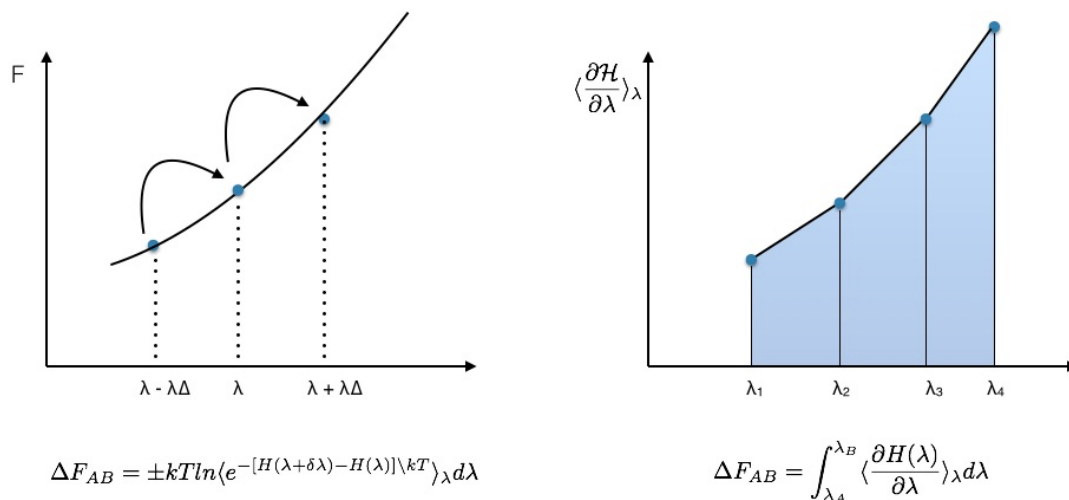
Fig. 4.3 Left: FEP, the difference in free energy is estimated from the relative probability of sampling a configuration appropriate to state B in state A. Right: TI, the free energy difference is determined by integrating along a reversible path from A to B.

binding calculations. Another possible source of errors is ligand parameterisation, as only a few force fields have tools for determining compatible functional groups [141].

Many approaches have been developed to overcome the sampling problem. Molecular dynamics are most often applied with enhanced sampling techniques that allow for rare events to be observed on a feasible timescale. Steered molecular dynamics (SMD) [142], umbrella sampling [143] and metadynamics [144] are techniques where the ligand and the protein are physically separated from each other. In SMD, the ligand is dragged out from the protein using a moving restraining potential, and umbrella sampling includes independent sampling of a a finite number of windows along the path from the bound to the unbound state. In metadynamics, an external history-dependent potential is added in a few degrees of freedom (collective variables) that prevents the system from sampling already observed events.

These methods have been successfully applied to estimating free energies of ligand binding [145–147], however, they do require knowing the binding mode in advance which may limit their applicability.

4.3   Computational methods for binding kinetics

Most of the computational approaches to describing molecular recognition have been focused on calculating affinities and free energies. Unlike free energy calculations, where valuable insights can be obtained by studying only reaction endpoints, estimation of bind-

ing kinetics relies on the reconstruction of the binding pathway. Two main approaches have been developed [148]. Brownian dynamics is a technique in which the solvent molecules are coarse-grained, which enables sampling rare events on reasonable timescales. This technique has recently been used to provide quantitative estimates. In particular, the residence time of HSP90 and MDM2 inhibitors was computed along with the characterisation of the drug-receptor encounter complexes [149].

The approximations and assumptions (rigid body dynamics, reduced point-charge interactions, implicit solvent, and a relatively large timestep) made when using Brownian dynamics to simulate molecular binding can also introduce inaccuracies. Therefore, to accurately recover experimental observables related to a binding process, additional models to approximate physical effects due to solvation shells and polarisation, solvent entropic effects, and solute internal degrees of freedom are often needed [150].

The second approach involves molecular dynamics. Ligand binding and unbinding occurs on millisecond time scales, which can hardly be achieved in a reasonable computer time. Enhanced-sampling methods are generally used to address this shortcoming and in some recent applications it has been possible to reconstruct the kinetics of binding by combining them with other theoretical approaches [151, 152].

All these approaches modify the system by applying forces that might alter the dynamics of the system. The advances in software algorithms, parallelisation of codes enabling high performance computing and the development of specialised hardware infrastructures has enabled the use of unbiased brute force simulations. In 2008, a special-purpose machine for molecular dynamics simulation (ANTON)[153] was developed by D.E.Shaw Research that enables 1 millisecond scales to be reached [154].

4.4   The basics of quantum mechanics

4.4.1   Introduction

Classical mechanics, built upon Newton's second law ($F = ma$) accurately describe most systems that can be easily observed: movements of the objects of "normal" size (larger than an atom, smaller than a planet), at a "normal" temperature (close to room temperature), moving through space at a "normal" speed (anything significantly less than the speed of light). In other words, classical physics is deterministic, meaning that it is possible to predict the trajectory of a particle and therefore location of the particle at a certain time. When the velocity of a particle is comparable to the speed of light, relativistic effects become important. Small light particles do not behave in a way that is consistent with the Newton's equations: it is common to think of the particle as being distributed through space like a wave, and it is then possible to calculate the probability of a particle being at a

certain place at a certain time. The interpretation is no longer deterministic, it becomes probabilistic, and the mathematical representation that replaces the classical concept of trajectory is called a wavefunction ψ. Atoms and molecules behave essentially as classic particles, while electrons can only be described by quantum mechanics.

The following sections briefly describe the basics of quantum mechanics and approximations that have to be made in order to develop these concepts into techniques that can be used to study physical, chemical and biological phenomena, and is based on different molecular modelling and computational chemistry textbooks [120, 155].

4.4.2   Atomic units

Quantum mechanics is concerned with electrons, protons and neutrons. The atomic units have been chosen such that the fundamental electron properties are all equal to one atomic unit. Mass, charge, etc. can then be expressed as multiples of these constants. The atomic units of mass, charge, length and energy are given below:

1 mass unit equals rest mass of electron, $m_e = 9.1094 \times 10^{-31}$ kg

1 unit of charge equals the elementary charge on an electron, $e = 1.6022 \times 10^{-19}$ C

1 unit of length (1 Bohr) is $a_0 = 4\pi\hbar\varepsilon_0/m_e e^2 = 5.2918 \times 10^{-11}$ m

1 unit of energy (1 Hartree) is $E_h = \hbar/m_e a_o^2 = 4.3597 \times 10^{-18}$ J

4.5   Schrödinger equation

In 1926, the Austrian physicist Erwin Schrödinger proposed the equation that describes how the wavefunction of a physical system evolves over time. The time-independent form of this equation for a particle of mass m, under the influence of an external field V is:

$$\{-\frac{\hbar^2}{2m}\nabla^2 + V\}\psi(r) = E\psi(r)$$

where $\hbar^2$ is a modification of Planck's constant $\hbar = \frac{h}{2\pi} = 1.05457 \times 10^{-34}$Js, E is energy and ∇^2 ('del squared') is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

In the general case, the Schrödinger equation is written

$$\hat{H}\psi = E\psi$$

where $\hat{H}$ is the Hamiltonian operator, which describes the kinetic and potential energy of the system:

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V$$

or,

$$\hat{H} = -\underbrace{\sum_{i=1}^{electrons}\frac{\nabla_i^2}{2}}_{T_e} - \underbrace{\sum_{A=1}^{nuclei}\frac{\nabla_A^2}{2m_A}}_{T_n} - \underbrace{\sum_{i=1}^{electrons}\sum_{A=1}^{nuclei}\frac{Z_A}{r_{iA}}}_{V_{en}} + \underbrace{\sum_{i=1}^{electrons}\sum_{j>1}^{electrons}\frac{1}{r_{ij}}}_{V_{ee}} + \underbrace{\sum_{A=1}^{nuclei}\sum_{B>A}^{nuclei}\frac{Z_A Z_B}{R_{AB}}}_{V_{nn}}$$

where T_e is the operator for the kinetic energy of the electrons, T_n is the operator for the kinetic energy of the nuclei, V_{ee} is the operator for the Coulomb attraction between electrons and nuclei, V_{ee} is electron repulsion and V_{nn} is repulsion between nuclei. Schematic representation of two atoms with the electrons whose coordinates are given by the vectors r_i and r_j is given in figure 4.4.
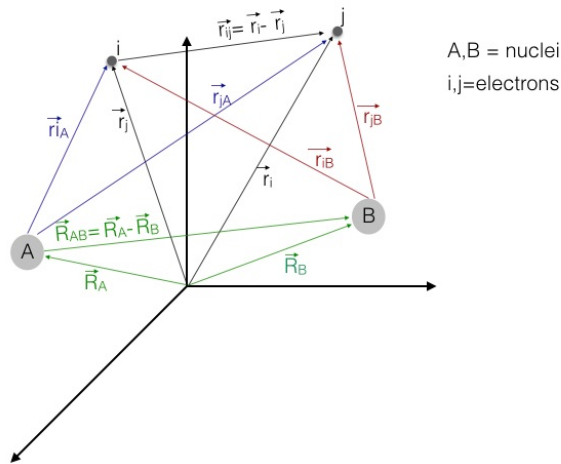
Fig. 4.4 Schematic representation of two atoms with the electrons whose coordinates are given by the vectors r_i and r_j.

4.6   Born-Oppenheimer approximation

It is a general principle of quantum mechanics that there is an operator for every physical observable. If the wavefunction that describes a system is an eigenfunction of an operator, then the value of the associated observable is extracted from the eigenfunction by operating on the eigenfunction with the appropriate operator[1]. To solve the Schrödinger equation,

it is necessary to find values of E and the function ψ such that when the wavefunction is operated upon by the Hamiltonian, it returns the wavefunction multiplied by the energy. ψ is an eigenfunction and E the eigenvalue.

Because the mass of the nuclei is much greater than the mass of electrons, only the electrons are considered as independent variables of $\hat{H}$, and the position of the nuclei is assumed to be stationary. The movement of electrons and nuclei are therefore separated. This is called the Born-Oppenheimer approximation, and can be written in the following form:

$$\psi_{molecule} = \psi_{electrons}\psi_{nuclei}$$

The Hamiltonian for a molecule with stationary nuclei ($T_n = 0$, $V_{nn} = constant$) is thus :

$$\hat{H} = -\sum_{i=1}^{electrons} \frac{\nabla_i^2}{2} - \sum_{i=1}^{electrons}\sum_{A=1}^{nuclei} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{electrons}\sum_{j>i}^{electrons} \frac{1}{r_{ij}}$$

The first term is the kinetic energy of the electrons, the second term is the attraction of the electrons to nuclei and the third term represents repulsion between electrons. Because of the large mass difference between electrons and nuclei, nuclei are stationary from the electronic point of view. The expression therefore ignores the coupling between the nuclear and electronic velocities but provides the potential energy surface on which the nuclei moves. In other words, upon invoking the Born-Oppenheimer approximation, the wave equation describes the probability of finding the electrons, given the positions of the nuclei.

A quantum mechanics treatment requires no experimental information as input, and, in theory, it should be the tool of choice for the reliable description of complex chemical processes. The first step in solving the Schrödinger equation for an arbitrary molecule is to set the specific Hamiltonian $\hat{H}$ of the target system. To do that, parts of the Hamiltonian that are specific for the system in hand need to be known: number of the electrons in the system and the external potential that is completely determined through the positions and charges of the nuclei. All the other operators are independent of the particular molecule. The next step is to find the eigenfunction ψ and the corresponding eigenvalue E. Once ψ is determined, all properties of interest can be obtained by applying the appropriate operators to the wavefunction. In practice, the Schrödinger equation can only accurately be solved for the hydrogen atom, meaning that a range of approximations have to be applied, which in turn reduce the accuracy of the obtained results.

Two basic QM approaches are used in practice and they both rely on the Born-Oppenheimer approximation in which the electrons move in the field of the fixed nuclei,

---

[1]An eigenfunction of an operator $\hat{A}$ is a function f such that the application of $\hat{A}$ on f gives f again, times a constant.

but differ based on the different types of approximations they use: *semi-empirical* and *ab initio*.

4.7 *Ab initio* methods

Ab initio is a Latin term meaning "from the beginning", or "from first principles" which implies that this approach requires no empirical parameters as an input. These methods do still rely on experimental data, but in a more subtle fashion, by guiding the selection of the model. Many different models exist and which one will be used is chosen by running calibration studies and comparing the performance against experimental data.

4.7.1 Hartree-Fock approximation

The Schrödinger equation can exactly be solved for only a few problems (one being the hydrogen atom) and the full quantum solution of a system of any nontrivial size is very difficult. The simplest type of *ab initio* electronic calculation is Hartree-Fock. The basic idea behind this approach is that electrons behave as independent particles: each electron interacts with the average field of all the other electrons and the many-electron wavefunction takes the form of a determinant of single-electron wavefunctions (Slater determinant). This approximation gives an average effect of repulsion, but not the explicit repulsion interaction and the resulting energies tend to be too high. In case a complete basis of single-electron wavefunctions is found, it is possible (in principle) to express the exact many-electron wavefunction as a linear combination of all possible determinants made of these wavefunctions (convergence to ground state).

The Hartree-Fock method starts with an initial guess for the orbital coefficients. This function is then used to calculate energy and a new set of orbital coefficients, which can then be used to obtain a new set, and so on. The procedure continues iteratively until the energies and orbital coefficients remain constant from one iteration to the next. This procedure is called the Self-Consistent Field procedure (SCF).

The most significant drawback of Hartree-Fock theory is that it fails to account for electron-electron correlation. The simplest way to tackle this problem is Møller-Plesset Perturbation Theory, usually of second (MP2), third (MP3) or fourth (MP4) order. The Hamiltonian operator H is expressed as sum of a 'zeroth order' Hamiltonian H_0 and a perturbation V: $H = H_0 + \lambda V$, where V is the small perturbation and λ is a dimensionless parameter.

4.7.2 Basis set

In *ab initio* methods, molecular orbitals are approximated by a linear combination of atomic orbitals. A basis set is a mathematical description of the orbitals within a system used to perform the theoretical calculation. Larger basis sets approximate more accurately the orbitals by imposing fewer restrictions on the locations of the electrons in space. An infinite basis set would lead to the calculation of the exact molecular energy, but this calculation would take an infinite amount of time. In practice, a calculation has converged to the exact solution when all the approximations are sufficiently small in magnitude so that the finite set of basis functions tends toward the limit of a complete set. The convergence is usually not monotonic, and sometimes the smallest calculation gives the best result for some properties.

There are two types of basis functions commonly used in electronic structure calculations: Slater type orbitals (STO) and Gaussian type orbitals (GTO). STOs decay exponentially with the distance from nuclei, while accurately describing the long-range overlap between the atoms. Unfortunately, some of the integrals in STOs are difficult or even impossible to solve. It is therefore common to approximate STOs with a linear combination of Gaussian orbitals, which leads to a 4-5 orders of magnitude more rapid calculations and huge computational savings. This is due to "Gaussian Product Theorem" which states that the product of two GTOs on two different atoms is a finite sum of Gaussians centered on a point along the axis connecting them. There are many levels of basis sets comprised of GTOs and they differ depending upon how many basis functions are used to characterise a given electron in an atom in the molecule. The smallest of these is called the minimal basis set and it describes only the most basic aspects of the orbitals. It contains the minimal number of functions that are required to represent all the electrons on each atom.

4.7.2.1 Split-valence

It is common to represent bonding orbitals by more than one basis function, and since the inner-shell electrons are not as vital to the calculation, they are described with a single orbital. A basis set that doubles the number of functions in a minimal basis set is called a double zeta basis (triple if it triples, quadruple if it quadruples and so on). These are so called "split-valence" basis sets. The notation is of form $X - YZg$, where X represents the number of primitive Gaussians in the inner shell, Y and Z indicate that the valence orbitals are composed of two basis functions, one for contracted, other for extended valence orbitals. Unlike the minimal basis sets, where the density is fixed and unable to adjust to different molecular environments, split-valence basis sets allow the density to adjust in space appropriate to the particular molecular environment.

4.7.2.2   Polarised sets

The most common addition to minimal basis set is the addition of polarisation functions. For example, using the minimal basis set, the only basis function located on a hydrogen atom would be the one approximating the 1s orbital. The electron cloud in an isolated hydrogen atom is symmetrical, but when the hydrogen atom is present in a molecule, electrons are attracted towards the other nuclei. When polarisation is added, a p-function is also added to the basis set, allowing molecular orbitals involving hydrogen atoms to be more asymmetric about the nucleus, leading to more accurate representations of bonding between atoms. In a similar fashion, d-type functions can be added to a basis set with valence p orbital, and f functions to a basis set with d-type orbitals. The use of polarisation functions is denoted by an asterisk (*). One asterisk (*) indicates that a polarisation function has been added on the heavy (non-hydrogen) atoms, while two asterisks (**) indicate the use of a polarisation function on hydrogen and helium atoms as well.

4.7.2.3   Diffuse sets

Another addition to the basis set which is useful for describing anions, molecules with lone pairs, transition states, excited states, or any other molecule that have a significant electron density at the distances further from the nuclei than the most ground state neutral molecules, is a diffuse function, which is denoted by a plus (+) sign. These are very shallow Gaussian basis functions, which more accurately represent the "tail" portion of the atomic orbitals, far from the atomic nuclei. Two plus signs (++) indicate that diffuse functions have been added to light atoms as well.

For example, the 6-31+G* basis set uses a single STO composed of 6 Gaussians for inner shell electrons, and two STOs for valence electrons: the contracted (inner) STO is composed of three GTOs, the extended (outer) STO is composed of a single GTO, a set of d orbitals is added on the heavy atoms (denoted by *) and an additional set of diffuse Gaussians is added on the heavy atoms (denoted by +).

The horizontal axes of the diagram in figure 4.5 relates to the extent to which the motions of electrons are independent from one another. At the extreme right are the models where there is no separation of electron motions, i.e., electron motions are fully correlated. At the extreme left are the Hartree-Fock models that do not take correlation energy into account and practical correlated models are somewhere in the middle.

The vertical axis relates to the basis set. At the top is the complete basis set that involves all possible functions, while at the bottom is the basis set containing the fewest possible functions. The region at the top right (fully correlated and complete basis set) corresponds to solving the exact Schrödinger equation.
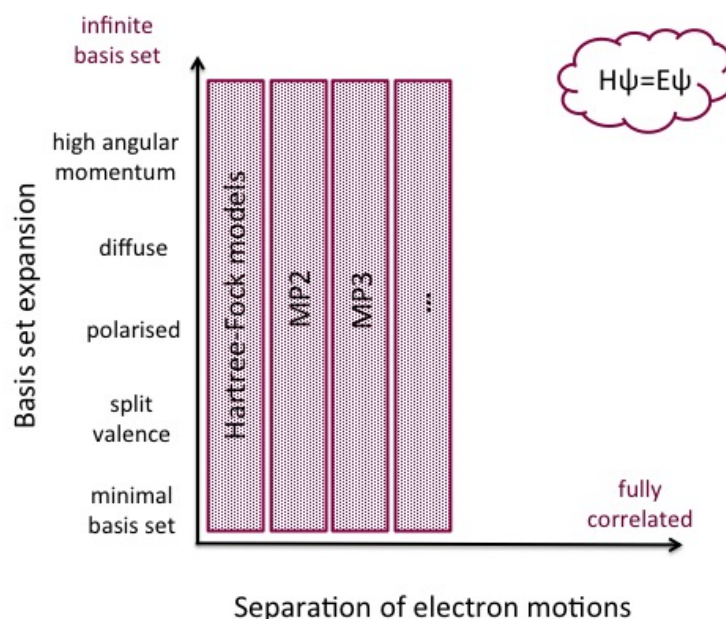
Fig. 4.5 Pople diagram describing the convergence of the quantum-mechanical nonrelativistic electronic energy with the size of the basis set and the level of electron correlation.

4.7.3 Density functional theory (DFT)

Density functional theory (DFT), a variant of *ab initio* is not based on solving the Schrödinger equation. In this approach, the many-electron wavefunction is completely bypassed in favour of the electron density, which is a function of three Cartesian cooridnates. Unlike the wavefunction which becomes more complicated as the number of electrons increases, the determination of electron density is independent of the number of electrons. According to the the Hohenberg-Kohn Theorems, the ground state energy of the system depends uniquely on the electron density, i.e. the total energy is a functional of the electron density that depends on only 3 spatial coordinates. Electron density is expressed as a linear combination of basis functions similar in mathematical form to Hartree-Fock orbitals. A determinant is then formed from these functions and it is used to compute energy. The correct ground state density for a system is the one that minimizes the total energy through the functional $E[n(x,y,z)]$, so it is possible to apply the variational principle to minimise the energy with respect to the electron density. The problem with this technique is that no one knows what the exact energy functional is. The simplest approximation is the local-

density approximation (LDA), that assumes that the density of the molecule is uniform throughout the molecule - the functional depends only on the (local) density at a given point. Modern methods using generalised gradient approximations (GGAs) account for the non-uniformity of the electron density by taking into account the gradient of the density to produce a more accurate functional. An extension of GGAs is Meta-GGAs, where the functional depends on density, its gradient, and its second derivative, such as functionals from the Minnesota family , and Hybrid DFT functional that mix in Hartree-Fock exchange. Most popular examples are B3LYP (hybrid GGA) [156–158] and M06 [159], which is a hybrid meta-GGA functional and the dominant method of choice in this work.

4.7.4   Semi-empirical methods

Semi-empirical methods are applicable to larger molecules and may give accurate results when applied to study molecules that are similar to ones used for parameterisation. In the approach, some integrals are either omitted or parameterised based on experimental data, such as ionisation energies of atoms or dipole moments.

Existing semi-empirical methods differ based on the approximations made when evaluating one- and two- electron integrals and by the parameterisation approach. The most frequently used methods (MNDO [160], AM1 [161], PM3 [162]) are all based on the Neglect of Differential Diatomic Overlap (NDDO) [163] integral approximation. This is a class of Zero Differential Overlap (ZDO) methods, in which all two-electron integrals involving two-centre charge distributions are neglected [164].

4.7.5   Quantum mechanical approaches to calculating binding affinity

A systematic approach in which *ab initio* methods improve the accuracy with which molecular recognition is described presents an appealing complementary approach to existing MM based methods [165–167]. Advances in parallel computing and sophisticated QM algorithms in conjunction with plunging prices of CPU time have led to diversification in the evolution of computational methods: with an end of Moore's law approaching (according to Intel, number crunching power is now doubling only every 2.5 years), the future of computing is driven by the progress in new algorithms, cloud computing services that enable harvesting of additional computing and storage resources for a fraction of the cost to purchase new computers, and new computing architectures- specialised chips optimised for specific jobs (one example being Anton, a supercomputer built for molecular dynamics simulations). These impressive improvements enable QM to progress from being

a sidekick to less computationally demanding (but also less accurate methods), to a more central role in *in silico* compound design.

In recent years, significant progress has been made in approaches that use mainly QM to estimate binding (or decompose binding affinity). So far, QM calculations of binding affinities that include only the portion of the protein that is directly involved in binding have been performed with a range of methods, including semi-empirical, HF and DFT, with basis sets up to 6-31G**. Most of the examples have been mentioned in a recent comprehensive review by Ryde and Söderhjelm [167] and prominent examples together with some additional ones are discussed here.

One of the earliest examples is Peräklyä and Pakannen's work in 1994, where the authors looked at binding energies of L-arabinose, D-fucose and D-galactose to L-arabinose binding protein [168]. Three contributions were taken into account: direct ligand active site energy calculated at the HF/6-31G* level of theory, solvation of the ligand and long-range electrostatic interactions both calculated at MM level, using a distance dependent dielectric constant. The system included ligand and nine surrounding residues and the correlation to experimental data, R^2, was 0.37. In 1995, they used a similar approach to study binding of seven ligands by p-hydroxybenzoate hydroxylase (PHBH). The calculations included ligand and six surrounding residues and the complex was optimised at the HF/3-21G level, resulting in R^2=0.7 [169].

In 2004, Nikitina et al. used semi-empirical QM on eight protein-ligand complexes to estimate the enthalpic contributions to binding free energy. In their approach they included all the residues with at least one heavy atom within 5 Å of the ligand and performed optimisation at the PM3 level of theory, with fixed backbone atoms. This study also looked at two different approaches to including water molecules that were bonded to both protein and the ligand [170]. This work was followed up by a study in which mixed implicit and explicit solvation models were included. Water molecules were added to all potential hydrogen bonding sites in both the protein and the ligand. The calculated enthalpies agreed with the experimental enthalpies within 2 kcal/mol [171].

In 2005, Raha and Merz used semiempirical QM to design a scoring function capable of calculating electrostatic interactions and solvation free energy expected during protein-ligand complexation. This approach was successful in replicating the binding free energy trends in a diverse set consisting of 165 protein-ligand complexes and also had a predictive power in calculating the binding free energy both from the experimentally solved structures and from docked poses. For 165 noncovalent complexes, the interaction energies were within 2.5 kcal/mol without any fitting [172].

In the same year Villar et al. looked at how AM1 and PM3 derived interaction energies and geometries between small molecules that simulate typical ligand-protein interactions compare to higher level techniques MP2 and B3LYP. AM1 gave much better agreement

with higher level techniques than PM3 and was further tested on the 5 ligand-protein complexes from the PDB. All protein residues within 5 Å from the ligand were treated with AM1. This gave reasonable geometries but no comparison with experimental energies was performed [173].

Leach et al. studied two series of compounds that inhibit dimerisation of iNOS. Their system included backbone atoms of six residues forming the binding site, the Cys interacting with the iron, and iron porphyrin with the various sidechains on the heme porphyrin ligand removed. The systems were treated with the range of QM methods, with M06HF performing the best (R^2 of 0.94 for LLE) [174]. Svensson et al. used DFT to explore the relationship between the measured and predicted affinities between a set of positively charged amidine and guanidine cores binding to the β- site of APP cleaving enzyme (BACE-1) [175]. Using a similar approach, Roos et al. looked at the relative binding affinities of a matched series of mineralocorticoid receptor (MR) agonists, using a range of DFT methods, with M062X performing the best (R^2 of 0.76 for LLE) [176].

A lot of work has been done with the *ab inito* fragment molecular orbital (FMO) approach. In this approach, a system is divided into fragments which are then treated with QM. By summing up the pair interaction energies (PIE) calculated for individual fragments, it is possible to get estimates of K_i. FMO has been used to predict the inhibitor potency of 28 CDK2 inhibitors [177] where the fragments were evaluated at the MP2/6-31G* level of theory, to assess binding energies of the FK506-binding protein and 10 ligand molecules [178], for lead optimisation of interleukin-2 Inducible T-Cell Kinase (ITK) inhibitors [179], combined with hierarchical GPCR modelling protocol (HGMP) to investigate 18 class A GPCR-ligand x-ray structures [180] and to explore the interactions and selectivity of the human orexin-2 receptor and its nonpeptidic agonist [181].

Research aims of this project

Predicting the potency of a compound remains an unsolved problem. For any compound to become a likely drug candidate, it needs to satisfy a number of different ADMET properties and have a good toxicological profile. Above all, the compound needs to bind tightly and selectively to a specific location on the target, and elicit the desired functional response of the target molecule.

Many compounds are synthesised and tested, most of which turn out to be inactive against the desired target, failing thus at the very first hurdle. Computational approaches to predicting many of the properties of the compounds are improving and provide a benefit by avoiding the expenses of making and testing compounds that are unlikely to become drugs. In parallel with these computational advances, what is understood by potency as measured by experimentalists has evolved. Drug action begins with the interaction between the compound and its target. This molecular recognition triggers a series of molecular events that ultimately result in a useful pharmacological response that corrects the pathophysiology. The molecular details of this response are dependent on the thermodynamics and binding kinetics. Therefore, addressing both thermodynamics and kinetics aspects of ligand binding can potentially reduce the numbers of compounds that need to be evaluated, and provide new opportunities for rational lead optimisation.

An overall aim of this project is to propose new ideas that contribute beneficially to the development of computational methods for interpreting potency in terms of thermodynamic or kinetic parameters.

The work described in this thesis presents a novel approach in which quantum mechanics is used to predict the enthalpic component of ligand binding, by constructing 'theoceptors'- theoretical receptors constructed by computing the optimal geometry of the ligand and the binding site. The introductory chapters 1 to 4 describe the theoretical background and give an overview of the state-of-the-art, while chapter 5 describes how this approach can be used to provide useful insights into interactions between proteins and ligands and turn these insights into new ideas for fragment series (see example of lactate dehydrogenase A inhibitors).

During the course of the project, an opportunity arose for expanding the application domain of the theoceptor approach. To test whether the method is also suitable for

correctly selecting preferred protein conformations and ligand positions where there is some movement of the protein backbone, a series of multiconformational structural refinements with the theoceptor derived geometries was performed. The results present compelling case to support the broader use of the methodology. Chapter 6 details this work. This work was supported by EPSRC UK National Service for Computational Chemistry Software (NSCCS) at Imperial College London.

This project included an industrial placement in the AstraZeneca's Structural and Biophysical Sciences area, where thermodynamic and kinetic signatures of 3-phosphoglycerate dehydrogenase and *Mycobacterium tuberculosis* Enoyl-ACP reductase inhibitors were generated. These results together with some of the issues and uncertainties associated with the two biophysical techniques used (surface plasmon resonance and isothermal titration calorimetry) are discussed in chapters 7 and 8.

In addition to this, an ongoing collaboration with MedChemica Limited (a local start-up company) was established. In chapter 9, some probabilistic insights into the field of matched molecular pair analysis are provided. This work was awarded a CINF Scholarship for Scientific Excellence at the 2016 American Chemical Society meeting in San Diego.

Part II

Experimental section

Chapter 5

Theoceptors for Lactate dehydrogenase A (LDHA)

5.1   Introduction

In most normal cells, ATP is generated by oxidative phosphorylation, while cancer cells predominantly produce energy by converting glucose into lactate through glycolysis, a phenomenon first discovered by Otto Warburg [182]. One molecular target in the glycolytic pathway is lactate dehydrogenase A (LDHA), an enzyme that catalyses the reversible conversion of pyruvate to lactate, with the concomitant conversion of NADH to NAD^+ (Figure 5.1). Several classes of cancers are characterised by elevated levels of lactate, and the enzyme itself is overexpressed in human tumours, making LDHA a promising target for cancer chemotherapies [183–187].
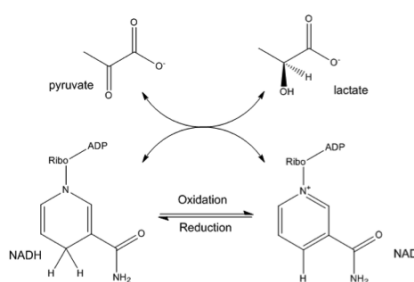
Fig. 5.1 Reaction catalysed by lactate dehydrogenase

Binding calculations can be very computationally demanding. The first principles nature of QM calculations enables systematic improvements to the accuracy with which biomolecular recognition is described. This methodical approach to solving chemical and biological problems is quite appealing, but practical use is hindered by the computational cost of these calculations. In this chapter, an alternative structurally "exclusive" approach

that includes only the parts of the protein that are involved in ligand recognition and neglects the remainder is proposed.

The proposed approach is reductionist and begins with the notion that a necessary first step to identifying interacting partners is to understand the interactions between molecules at an atomic level. By using *ab initio* methods, new discoveries require only the protein geometry. The computational method (steps 1 to 5) is presented in Figure 5.2. The steps involved are: 1) A binding site in a known protein structure (from x-ray crystallographic or NMR experiments) is selected 2) The protein residues immediately flanking the binding site are identified. The remainder of the protein is deleted. Tautomeric and protonation states of the protein must be assigned. 3) The $C\alpha$ and $C\beta$ atoms are identified. These will be fixed in space during the quantum mechanical calculations. 4) Quantum mechanical geometry optimisation causes small changes to both the ligand and protein geometry. Large changes indicate an inappropriate structure or protonation state. Structural water molecules can be constrained or allowed to move. 5) The ligand is optimised in isolation. Solution phase quantum mechanical optimisation is performed.
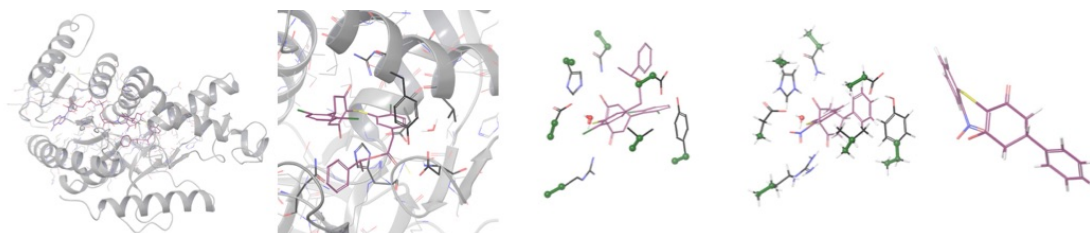
Fig. 5.2 Steps involved in constructing the theoceptor. Binding site in the crystal structure is selected. Residues directly involved in binding are identified. The Cα and Cβ atoms are identified and fixed in space to mimic the scaffolding effect of the remainder of the protein. Geometry of the complex is optimised. Ligand is optimised in isolation
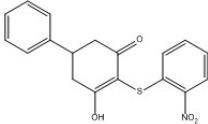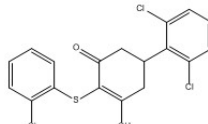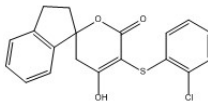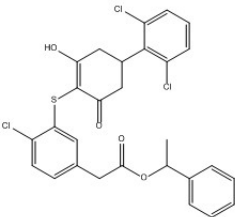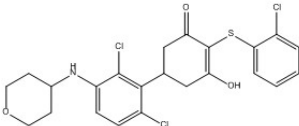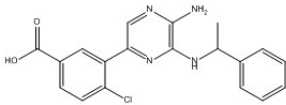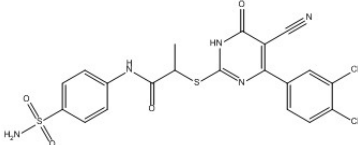
This chapter details the efforts in computational modelling of different series of LDHA inhibitors and shows how quantum mechanical calculations can provide useful insights into interactions between protein and ligands, with the ultimate goal of suggesting a new fragment series with the more favourable predicted potency.
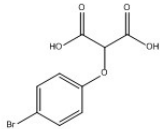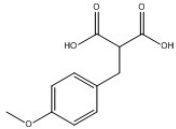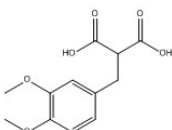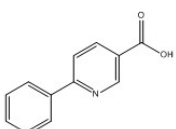
5.2 Compound selection

The set of compounds included in this computational study included 3,6-disubstituted dihydropyrones, 3-hydroxycyclohex-2-enones, 2-thio-6-oxo-1,6- dihydropyrimidines, and amino-pyrazines from a high-throughput screen (HTS) of the Genentech/Roche corporate compound collection [188–192]; small, negatively charged pyridine derivative [193]; and three compounds sharing malonate as a common substituent originating from AstraZeneca

fragment screening campaigns [194] (Table 5.1). The compounds were selected based on the availability of structural and bioactivity data.

Table 5.1 Compounds selected for computational studies. For structures **1-7**, LLE was calculated as pIC50- clogP, and for structures **8-11** as pKd –clogP, logP was calculated using ChemAxon logP predictor [195]

		pIC_{50}	pK_d (SPR)	pK_d (NMR)	LLE
1		5.76	5.46	NC	1.98
2		6.06	5.74	NC	0.41
3		6.44	NC	NC	1.96
4		8.22	NC	NC	0.64
5		7.22	NC	NC	2.13
6		5.3	5.3	NC	1.49
7		6.12	5.29	NC	3.16

8		<3.3	3.67	3.33	1.52
9		<3.3	2.96	3.63	1.98
10		<3.3	3.55	3	1.53
11		<2.7	2.63	NC	0.19

* NC=not calculated

5.2.1 Crystal structure evaluation

Crystal structures of **8** and **9** are available in complex with rat LDHA (PDB codes 4AJE and 4AJI), as attempts to generate a soakable crystal system using human LDHA were unsuccessful [194], and structure of **11** is available in complex with rabbit LDHA (PDB code 4I8X), as that system gave well-diffracting crystals [193]. Structures **1-7** are all available in complex with human LDHA [188–192].

RSCC values for the 10 ligands selected are given in table 5.2. Density maps contoured at 1 σ of the ligands with lower RSCC values (ligands **3, 6** and **7**) are shown in figure 5.3.

Table 5.2 RSCC values of ligands **1-8,10-11**.

	PDB	Chain (µM)	RSCC ligand
1	4QO7	A	0.958
2	4QO8	A	0.921
3	4RLS	D	0.883
4	4R68	B	0.979
5	4R69	D	0.943
6	4M49	A	0.816
7	4JNK	D	0.911
8	4AJE	B	0.988
10	4AJI	B	0.963
11	4I8X	B	0.913
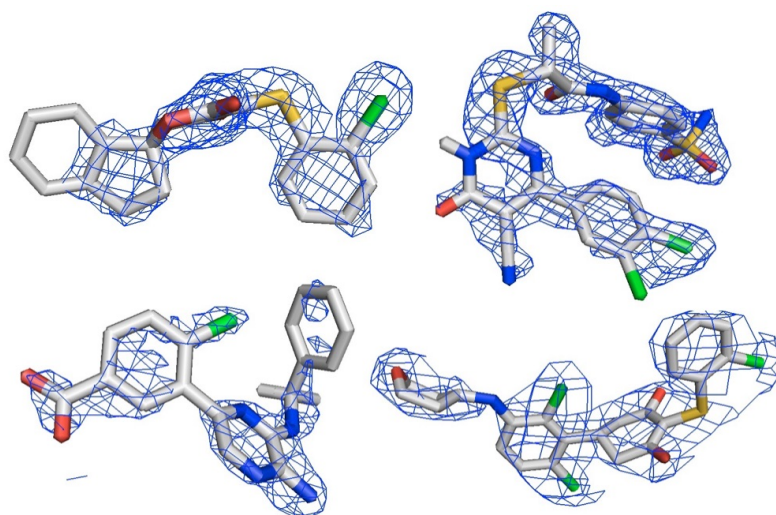
Fig. 5.3 Density maps $2mF_o\text{-}DF_c$ of ligands **3,7** (first row) and **6, 5** (second row) contoured at 1 σ.

5.3   Docking

The set of structures was studied using AutoDockVina [196]. Docking was performed into a representative structure from each rat, rabbit and human system. Three structures (**6, 11** and **8**) were processed with Schrödinger's protein preparation wizard [197] to ensure reasonable starting points for docking: hydrogen atoms were added after deleting any original ones, followed by adjustment of bond orders for amino acid residues and the ligand. The protonation and tautomeric states of Asp, Glu, Arg, Lys and His were adjusted to match a pH of 7.4. All water molecules and cofactors were removed. Structures were overlaid and a user-defined docking search space, which is explored for possible ligand binding conformations (grid box), was centred on the 2-aminopyrazine-based ligand from 4M49

and extended 30 grid points in each direction. Ligands were extracted from the crystal structures of the complexes and polar ligand hydrogens were added with AutoDock Tools (ADT 1.5.4) [198]. Two types of ligand geometries were provided as input: geometries extracted from protein crystal structures and geometries generated by application of Corina [199] to 2D structures. Correct bond orders and protonation states were assigned upon visual inspection.

The geometries obtained for compounds **1-8** and **10-11** by AutoDock Vina were compared to their crystal structures and results are summarised in Table 5.3. The entries coloured red are the ones where the native binding mode was not correctly reproduced (rmsd<=1.5 Å).

Table 5.3 Docking scores for the highest ranking pose obtained for docking each compound into each protein structure. RMSDs for ligands in the best ranked docking pose when superimposed to those in crystal structures. For structures **1-7**, LLE was calculated as pIC_{50}- clogP, and for structures **8-11** as pK_d –clogP.

	pIC_{50}	pK_d	LLE	Human Score	Human RMSD	Rat Score	Rat RMSD	Rabbit Score	Rabbit RMSD	Corina Score	Corina RMSD
1	5.76	5.46	1.98	-6.5	0.54	-6.0	1.08	-7.1	1.14	-7.5	4.31
2	6.06	5.74	0.41	-6.7	0.33	-7.6	2.08	-6.9	2.06	-7.0	5.14
3	6.44	NC	1.96	-7.4	0.54	-8.6	2.04	-7.5	2.05	-7.8	3.75
4	8.22	NC	0.64	-7.3	1.17	-6.5	2.50	-7.8	1.23	-7.3	4.89
5	7.22	NC	2.13	-8.0	2.36	-7.8	0.49	-7.5	0.37	-8.1	5.54
6	5.3	5.3	1.49	-8.2	2.50	-7.6	2.04	-7.8	1.80	-7.5	5.26
7	6.12	5.29	3.16	-7.5	1.69	-8.1	3.10	-8.4	3.73	-8.0	6.8
8	<3.3	3.32	1.52	-5.2	1.54	-5.9	1.18	-5.7	1.44	-5.4	2.43
10	<3.3	3	1.98	-5.7	2.14	-5.8	0.86	-5.8	1.05	-5.6	4.18
11	<3.3	2.63	1.53	-6.3	1.14	-6.2	0.44	-6.6	0.46	-6.2	1.66

The ligands come from very different series and do not share a common substructure hence, the distances between the centre of mass of each ligand, Arg168 Cβ and Asn137 Cγ, (two residues participating in hydrogen bond formation), were assessed (Table 5.4, Figure 5.4). It is interesting to note that none of the dockings ranked the most potent and one of the weakest binding ligands, ligands **4** and **8** correctly.

The geometric parameters in Table 5.4 show that poses Vina ranks the best are not the native ones. In most of the cases, the native binding pose was found but it was not ranked the best. In terms of orientation of the ligand, redocking of the native conformation or docking of Corina-generated conformations into the human structure outperformed docking into rabbit or rat crystal systems.

Surprisingly, the best correlation between the docking score and potency or binding affinity is achieved when docking Corina generated structures into the human receptor
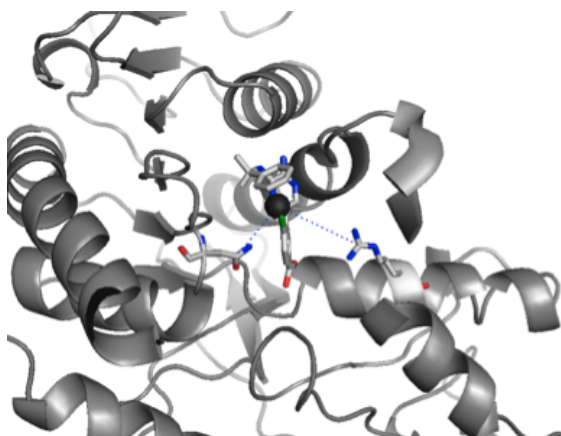
Fig. 5.4 The centre of the docking grid box was placed on the centre of ligand **6**. Geometry assessment: distance between centre of the mass of the ligand and Arg168 Cζ and Asn137 Cγ. In case of the ligand **6**, both of these centres coincide and are shown as black ball.

(R^2 of 0.6), even though it performed the worst on both RMSD and geometry assessment (Table 5.5).

As it has been discussed previously (Chapter 2, Ligand efficiency metrics-LEMs), it has been observed throughout the years that general partitioning effect caused by increased lipophilicity is linked to poor physical-chemical and ADME[1] properties. It is a common practice to examine ligand lipophilic efficiency (LLE, defined as pIC_{50}- clogP for structures **1-7**, and for structures **8-11** as pK_d –clogP), a metric that accounts for these effects [65]. Therefore, LLE is a useful property to improve during a lead optimisation process, and it is of value to examine if any of structure-activity relationship (SAR) modelling affects LLE.

Docking results were compared to LLE and no correlation was observed - correlation coefficients range from 0.08 for redocking into human crystal system, to 0.21 for docking of Corina-generated ligands (Table 5.5).

---

[1]absorption, distribution, metabolism, and excretion

Table 5.4 Distances from the centre of mass of the ligand in the crystal structure and the best ranking poses to Cζ Arg168 and Cγ Asn137. Entries coloured green are within 1 Å relative to crystal structure.

		1	2	3	4	5	6	7	8	10	11
Crystal	Arg168 Cζ	6.0	6.8	5.2	7.9	7.3	9.7	11.9	7.1	7.4	7.9
	Asn137 Cγ	6.1	5.7	6.1	4.9	7.3	6.7	4.1	4.9	5.2	5.1
Human	Arg168 Cζ	8.8	9.6	9.0	9.8	12.7	18.1	9.0	7.2	8.6	24.6
	Asn137 Cγ	4.9	6.2	6.0	5.3	8.3	11.9	5.4	5.7	4.3	16.3
Rat	Arg168 Cζ	6.3	5.2	5.0	9.5	6.1	5.4	5.8	9.6	9.0	5.8
	Asn137 Cγ	16.2	15.3	15.2	5.5	16.0	15.5	15.8	4.7	4.5	15.8
Rabbit	Arg168 Cζ	8.2	13.6	5.4	9.5	11.4	9.9	15.5	9.0	7.3	7.8
	Asn137 Cγ	5.5	7.8	15.6	5.5	6.6	3.6	9.3	4.8	4.8	4.7
Corina	Arg168 Cζ	10.7	10.0	20.5	9.8	17.1	10.2	15.5	7.2	8.5	24.5
	Asn137 Cγ	5.2	6.6	14.6	5.0	10.9	3.4	9.6	4.6	4.5	16.2

Table 5.5 Summary statistics describing the relationship between docking scores and pIC$_{50}$/K$_d$ or LLE. For structures **1-7**, LLE was calculated as pIC$_{50}$- clogP, and for structures **8-11** as pK$_d$ -clogP

	pIC$_{50}$/K$_d$		LLE	
	R^2	RMSE	R^2	RMSE
Human	0.48	0.89	0.08	3.01
Rat	0.27	1.57	0.16	2.3
Rabbit	0.48	0.77	0.15	1.94
Corina	0.6	0.69	0.21	1.75

5.4   Structure processing and QM optimisation

Ten crystal structures (the complexes of **1-8, 10** and **11**) were split into their constituent chains and superimposed. Noticeable differences in binding site residues were present for: 1) Arg105 - this highly mobile residue was observed protruding into binding sites in rat and rabbit structures, as well as human structures in chains where there is no ligand present. In human structures with the ligand present, this residue is pointing away towards Glu191 2) Tyr238 - sidechain points away from the binding site in chains where there is no ligand present (Figure 5.5).

No cofactor was present in crystal structures of **8, 10** and **11** and no electron density corresponding to cofactor structure could be observed. Crystal structures of remaining
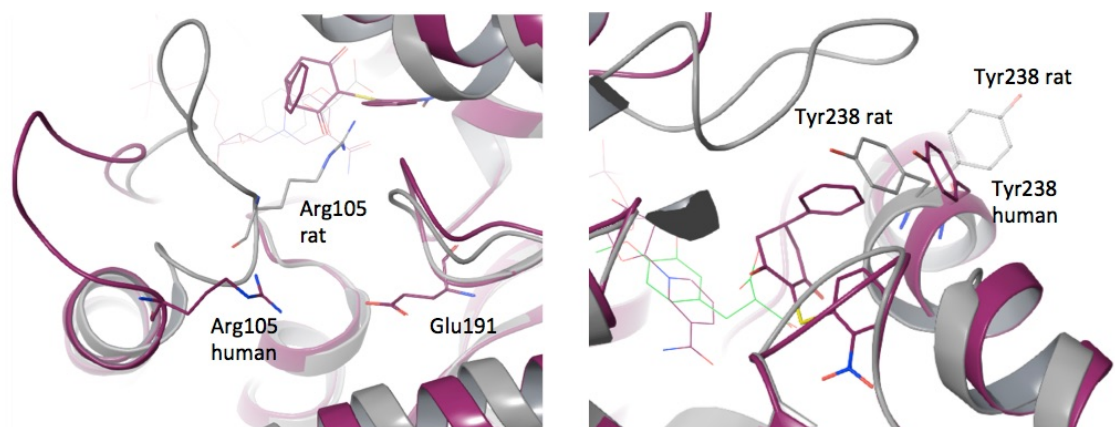
Fig. 5.5 Arg105 and Tyr238 were observed protruding into binding sites in rat and rabbit structures and human structures in chains where there is no ligand present. Human structure is shown in purple, rat in grey.

complexes (**1**- **7**) had cofactor modelled into well defined corresponding electron density. Selected structures are given in Table 5.2. The structure of **9** was obtained by analogy to that of **10**. All the ligands ionise at physiological pH.

Residues that were identified as the ones directly involved in ligand binding were selected as a part of the theoceptor: Arg168, His192, Val234, Asp165, Asp194, Tyr238, and Asn137. Protonation states were assigned manually and the net charge of the residues was 0. Asp165 forms a hydrogen bond with His192 and is stabilised by an additional hydrogen bond with one water molecule. Preliminary studies in which the theoceptor did not include this water molecule showed that this hydrogen bond is highly dependent on the starting Asp165 conformation. When the water molecule is present, Asp165 forms a hydrogen bond with water oxygen, which keeps the sidechain conformation in optimal position to interact with positively charged His192 sidechain. Therefore, this water molecule is considered structural as it is necessary for plasticity of the binding site and it was kept as a part of the theoceptor. All interactions with the ligand are through protein sidechains: backbone, cofactor and the remainder of the protein were deleted. To account for the scaffolding effect of the rest of the protein, Cα, Cβ and water oxygen atoms were fixed in space using the fixed Cartesian redundant coordinate feature in Gaussian09, as shown in Figure 5.6.

The process of ligand binding to its protein target can be seen as a partition problem where ligand is transferred from one medium (free in water) to another (binding site). Dielectric constant (ε) value for the interior of the protein is relatively low (about 6-7) [200], compared to dielectric constant of water, which is 80 at 20 C$^\circ$. Therefore, the interior of the protein resembles the gas phase more than it resembles water (Figure 5.7). LLE can be interpreted as proportional to free energy of binding for an inhibitor partitioning
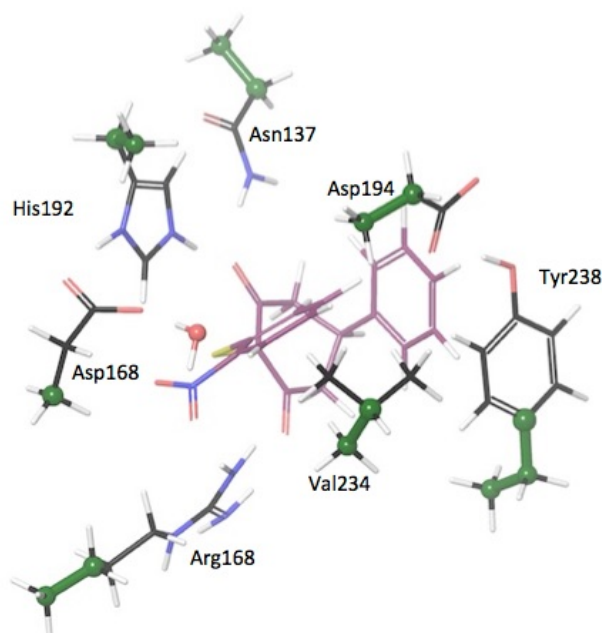
Fig. 5.6 Theoceptor **1**. Cα and Cβ were fixed during the QM optimisation

between 1-octanol and the enzyme, rather than between water and the enzyme. Dielectric constant of 1-octanol is 10.3 [201], making it more similar to gas phase than water. From a computational perspective, this has removed the aqueous environment, which is hard to model, and suggests LLE values may be better reproduced using this approach than potency or affinity.
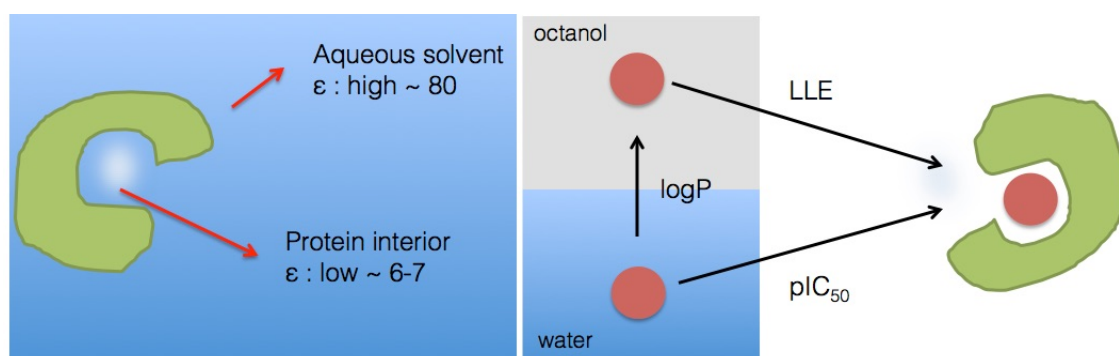
Fig. 5.7 For proteins in aqueous solutions, hydrophobic pockets inside the protein have a lower dielectric constant than an outer solvent. LLE can be interpreted as proportional to free energy of binding for an inhibitor partitioning between 1-octanol and the enzyme, rather than between water and the enzyme

The geometries were optimised in vacuo at the M06 level of theory with the 6-31+G** basis set. To include the effect aqueous environment has on a small molecules, free ligand

Table 5.6 LDHA theoceptor complexation energies

	Predicted affinity (kcal/mol)	LLE
1	-23.2	1.99
2	-13.6	0.41
3	-13.8	1.96
4	-15.0	0.64
5	-17.7	2.13
6	-10.3	1.48
7	-24.9	3.16
8	-15.6	1.53
9	-18.9	1.99
10	-19.7	1.53
11	-10.4	0.19
11m	-6.4	0.19
malonate	-39.5	2.63
oxamate	-62.6	5.66

structures were optimised with solvation incorporated, using the PCM water model, at the same level of theory.

The affinity was defined as the energy difference between lowest energy of the complex, free ligand and the apo structure: $\Delta E = E_{complex} - E_{ligand} - E_{apo}$. The calculated values are given in table 5.6. LLE plotted against computed complexation energies for all the compounds is show in figure 5.8.
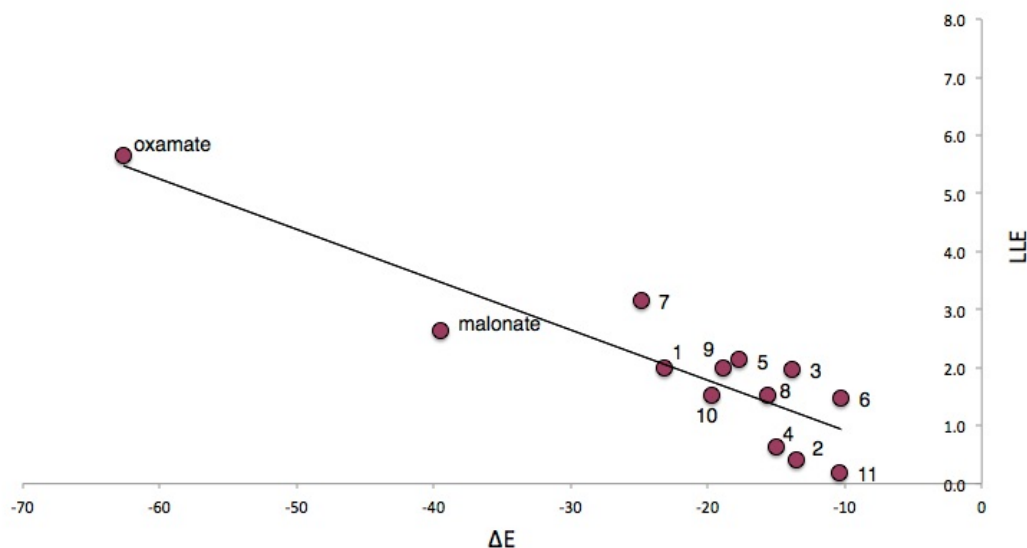
Fig. 5.8 LLE plotted against computed complexation energies for compounds **1-11**, oxamate and malonate

5.5 Beyond electron density maps - conformational, structural and stereochemical preferences

The tetrahydropyran ring of ligand **5** is crystallographically modelled in a boat conformation, a conformation that has torsional and steric strain and is usually of higher energy than the chair conformation. When this tetrahydropyran ring was modelled as a chair, the free conformation was 9 kcal/mol lower than the boat. The RSCC of the ligand was 0.94, which implies that the ligand fits the density well- however looking at the density contoured at 1σ (Figure 5.3), it is clear that both chair and boat would fit equally well. It is interesting to note that the conformation has not been changed after a short minimisation using the OPLS_2005 force field. Optimising the theoceptor with ligand **5** in a chair conformation resulted in more than -10 kcal/mol energy lowering compared to the starting crystal-derived structure (Figure 5.9).

A RSCC for ligand **3** of 0.833 suggests that the ligand does not fit the density map accurately, which was confirmed after visual inspection of the emap- the density around the indane moiety is missing (Figure 5.3). The first step to model this compound was to determine the preferred position of the oxygen atom in the dihydropyrone ring, because this can easily be misplaced when interpreting electron density maps. Changing the position of oxygen atom inverts the configuration of the stereocentre. The energy difference of the theoceptor optimised with the ligand stereocentre in (S) configuration (as it was crystallographically interpreted) and the one where the ligand stereocentre was in (R)
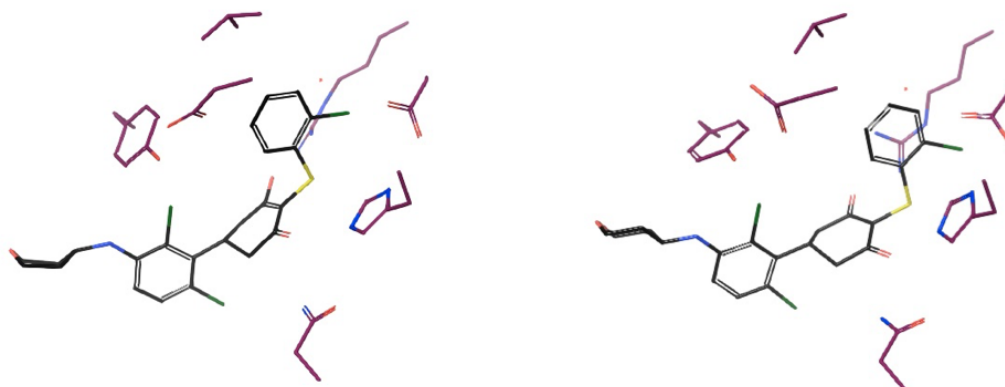
Fig. 5.9 Geometry optimisation of theoceptors with compound **5**. The tetrahydropyran ring
was crystallographically modelled in boat conformation. Modelling it as chair resulted in
more than -10 kcal/mol energy gain. Hydrogen atoms are omitted for clarity.

configuration was less than 0.2 kcal/mol, indicating that the dihydropyrone oxygen is not
interacting with the theoceptor (Figure 5.10).
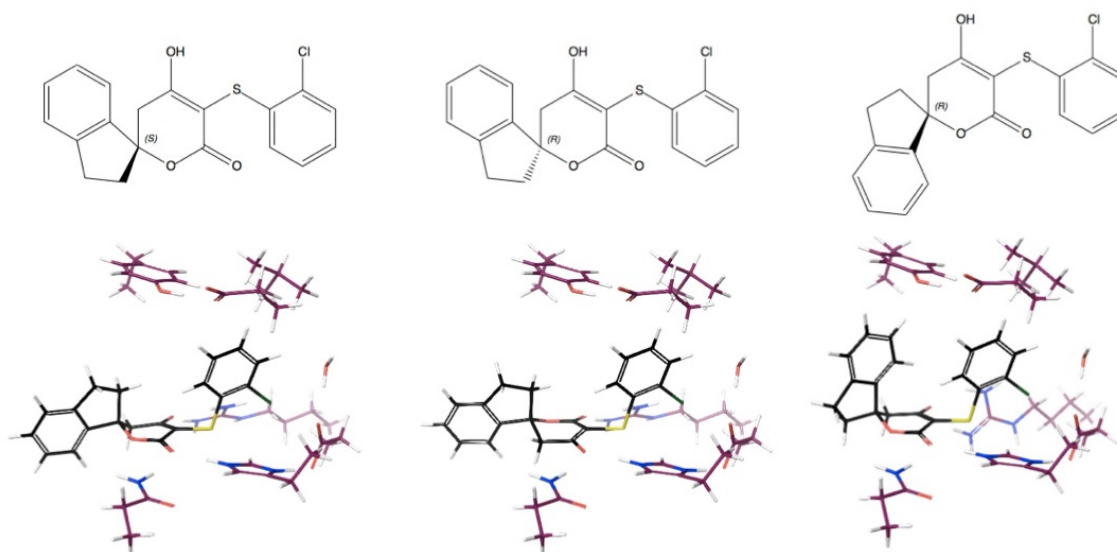
Fig. 5.10 Exploring stereochemical preferences of ligand **3**.

The missing electron density around the indane moiety suggests that the phenyl sub-
stituent can adopt pseudo-axial or pseudo-equatorial orientation, relative to the dihydropy-
rone ring. Pseudo-axial conformation with the ligand (R) stereocentre configuration had
-8 kcal/mol lower energy than the starting crystallographically observed structure. This
energy gain is originating from additional edge to face stacking between the ligand and

tyrosine, as well as the intramolecular interactions between the two terminal rings. (Figure 5.10).

Similar to **3**, we next looked at the position of nitrogen in the pyridine carboxylic acid moiety of ligand **11** (Figure 5.11). The energy difference between the ligand with nitrogen oriented as in the crystal structure and the mirror structure in which the pyridine is rotated through ~180°(**11m**) was 4 kcal/mol. This 4 kcal/mol gain preference for **11** is due to an interaction of the pyridine nitrogen with a positively charged environment formed by the sidechains of His192 and Asn137. In **11m** this nitrogen is pointing away from the theoceptor.
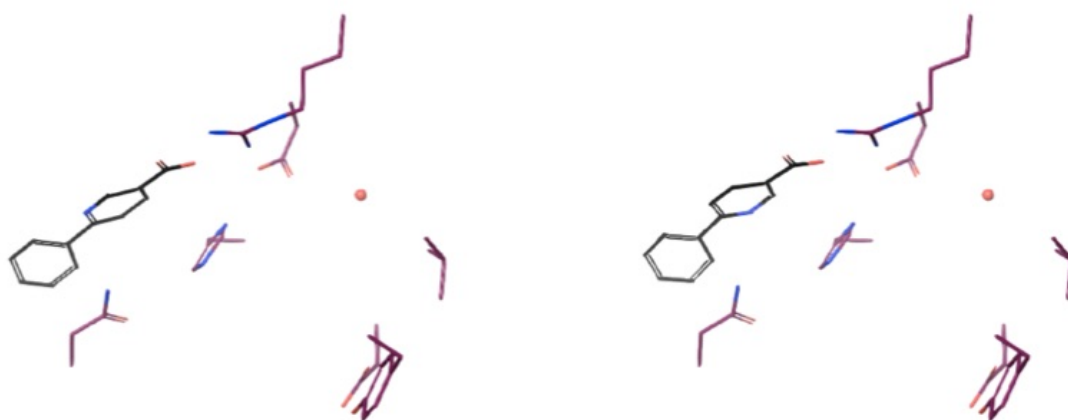
Fig. 5.11 Theoceptor **11m** and **11**. There is a 4 kcal/mol energy preference for ligand **11**, caused by pyridin nitrogen interacting with the positively charged portion of the theoceptor.

The phenyl substituent in ligand **1** was observed crystallographically in an axial position relative to the 3-hydroxycyclohexenone ring. Somewhat surprisingly, axial and equatorial ligand conformations were computed to be energetically indistinguishable in the unbound state. Substituents larger than hydrogen generally suffer greater steric crowding when they are oriented axial rather than equatorial. Consequently, substituted cyclohexanes will preferentially adopt conformations in which large substituents assume equatorial orientation. In 3-hydroxycyclohexenone ring there are no axial hydrogens located on the same side of the ring, and, consequently, no steric crowding. However, there is a clear preference in the bound state: the theoceptor with the ligand in an equatorial conformation is 4 kcal/mol higher in energy. This difference originates from protein-ligand edge to face stacking present when the ligand is in an axial conformation (Figure 5.12).

Fig. 5.12 Axial and equatorial orientations of ligand **1**. When the ligand is in an axial orientation, there is an energy contribution originating from protein-ligand edge to face stacking.
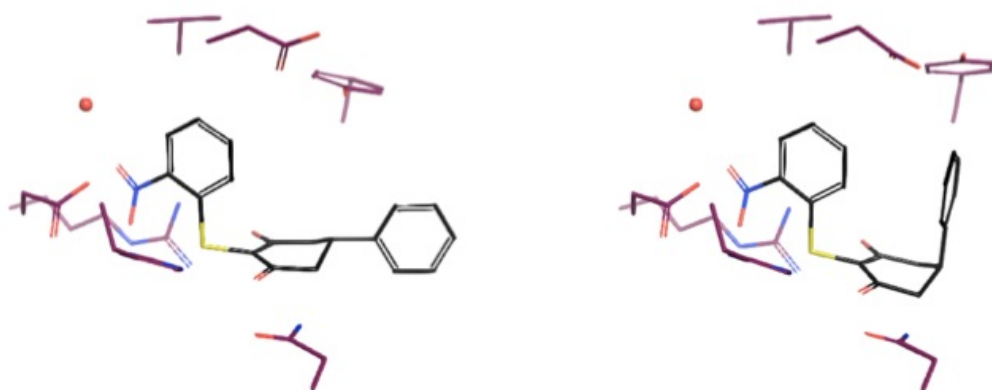
5.6 Finding the ligand global minimum energy conformation

Although the ligands are often expected to bind to the protein in a low energy conformational form, this is not always the case [40, 202]. Apart from the protein induced strain, additional complicating factors are inaccuracies in interpreting electron density maps that end up with the ligand conformations far from the global energy minimum, such as was the case with ligand **5**. To overcome this limitation, additional ligand conformations were generated using a stochastic search, as implemented in MOE (Molecular Operating Environment) [203] and these conformations were then optimised in solution.

One of the strong indicators of how far away the ligand is from the global minimum is the complexation energy: due to the relationship $\Delta E = E_{complex} - E_{ligand} - E_{apo}$, deviation from the line of the best fit can be caused if the geometry of the complex or the ligand are deviating from their lowest energy structure. Which one of these two parameters is not optimal is indicated by the direction of the offset. The initial complexation energy of theoceptor with ligand **4** was -20.0 kcal/mol, causing an offset from the line of the best fit. The direction of the offset indicated that the ligand conformation is in fact trapped in one of its local minima. This was confirmed by optimising MOE derived structures: QM optimisation energies of almost all additionally generated structures were lower in energy and the structure determined to be the likely global minimum (at this level of theory) was 5 kcal/mol lower than the energy of the optimised PDB extracted structure (Figure 5.13).

Initial optimisation of the theoceptor with ligand **7** resulted in a complexation energy of -16.5 kcal/mol, causing an offset from the line of the best fit that implies that the complex
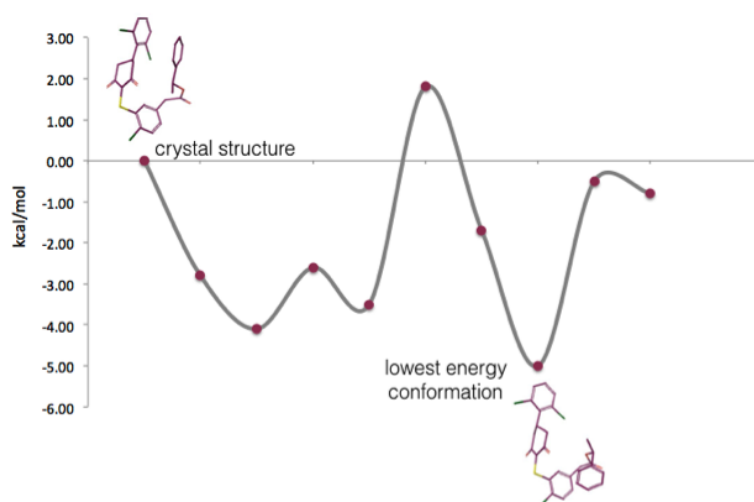
Fig. 5.13 Schematic representation of structure **4** conformational search. Each point represents a local minimum. The lowest energy structure (global minimum) is 5 kcal/mol lower in energy than the starting one.

geometry is not in its lowest conformational form. Ligand **7** forms some of its interactions with the protein through its sulfonamide moiety and residues Asp140, Glu191 and Arg105, that were not included as the part of the theoceptor. The lack of these additional interactions was first believed to be the origin of the positioning away from the line of the best fit (Figure 5.14).
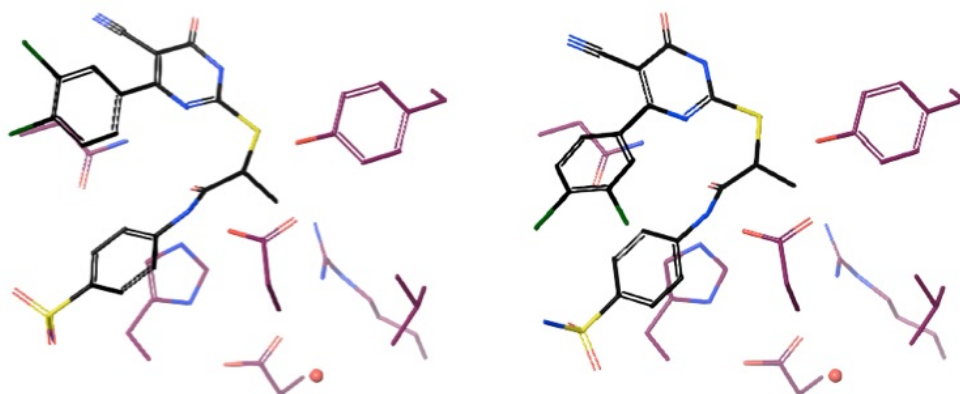
Fig. 5.14 Left: ligand **7** forms some of its interactions with the protein through its sulfon-amide moiety and residues Asp140, Glu191 and Arg105, that were not included as the part of the theoceptor. Right: there is a 9 kcal/mol energy difference between the theoceptor optimised with the crystallographically observed ligand conformation and theoceptor with global minimum.

However, all generated ligand **7** conformations were lower energy states, with the lowest energy conformation being -3.9 kcal/mol away from the crystal derived one (Figure 5.15).
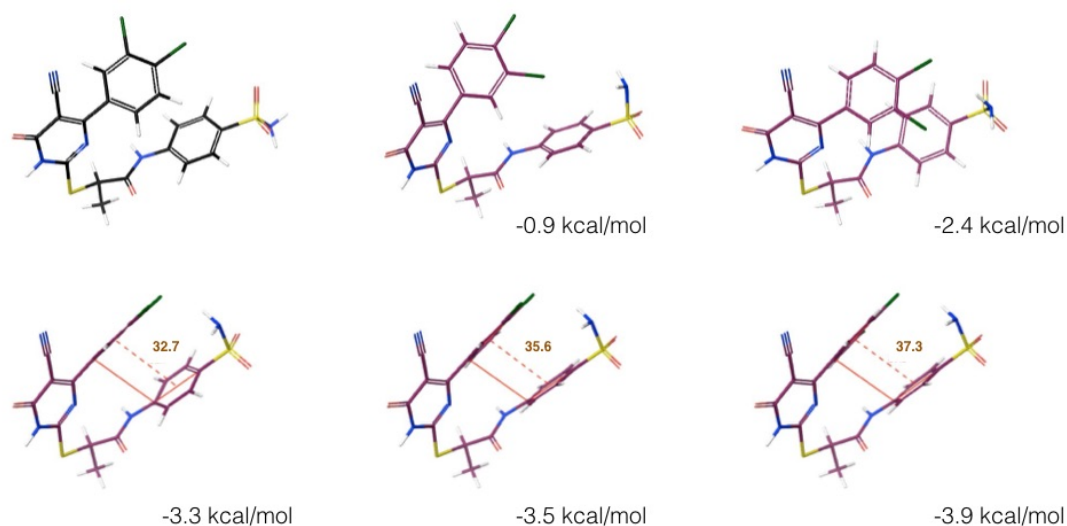
Fig. 5.15 All additionally generated ligand **7** conformations were lower in energy than the crystallographically observed one (top left, shown as black sticks). The angle between the two ring planes is shown.

In this conformation, the dichlorophenyl group is stacking with the sulfonamide substituted terminal phenyl group and the amide is pointing towards the aromatic ring, which was not observed in the crystal structure. This rotation is supported by the observed electron density map (Figure 5.3). Optimising the complex with this ligand conformation resulted in a complexation energy -9 kcal/mol lower than the theoceptor optimised with the crystallographically observed ligand conformation, minimising the offset and suggesting that the energetic contribution sulfonamide moiety makes with the residues that are missing in the theoceptor description is not as high.

Crystallographically observed structures of ligands **2** and **6** were already close to conformational global minimum: conformational search did not result in lower energy structures. For example, all additionally generated ligand **6** conformations were up to 5 kcal/mol higher in energy than crystal-derived one, and up to 8.4 kcal higher in case of ligand **2**.

5.7 Malonate derivatives - assessing protonation state and tautomeric preferences

Initial positioning of the malonate derivatives was carried out by superimposing the malonate moiety on the 3-hydroxy cyclohexenone ring of ligand **1**, so that it was in a close proximity to the basic side chain of Arg168. The correct protonation state and conformation had to be determined. The malonate moiety of compound **9** was first modelled in its dianionic form, with the resulting complex energy of +44 kcal/mol. This arises due to the loss of solvation energy that is not compensated by sufficient interactions in the receptor. These findings suggest that the malonate moiety may bind in a mono deprotonated form.

To assess the effect that the malonate moiety conformation has on the overall theoceptor energy, we tested three different ligand conformations for malonate derivative **9** in the theoceptor (Figure 5.16).

-18.9 kcal/mol   -8.9 kcal/mol   -3.6 kcal/mol
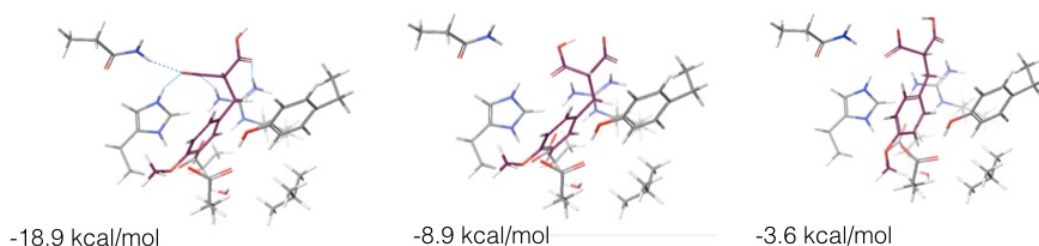
Fig. 5.16 Positioning of ligand **9**. The lowest energy is achieved when malonate forms 4 hydrogen bonds with the surrounding residues.

The lowest energy theoceptor was the one where the malonate moiety made four hydrogen bonds with the nearby sidechains (-18.9 kcal/mol). The energy difference of 15 kcal/mol between the highest and lowest energy theoceptors emphasises the effect subtle conformational changes of the interacting groups can have on the overall binding energy, particularly when strongly interacting groups such as anions are involved. Ligands **8** and **10** were initially positioned based on theoceptor **9**, so that malonate forms four hydrogen bonds with the theoceptor sidechains. Optimised structures are shown in Figure 5.17.
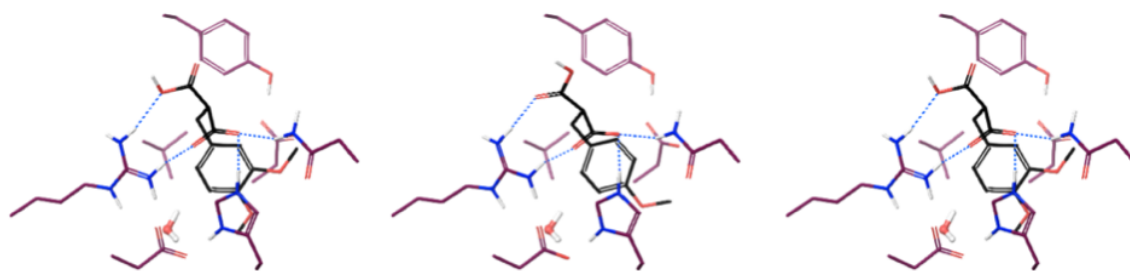
Fig. 5.17 Theoceptors for compounds **8, 9** and **10**. The malonate moiety forms four hydrogen bonds with the protein sidechains.

5.8 Malonate and oxamate - binding in the presence of cofactor

A total of 8 different malonate tautomeric monoanions were optimised in solution and all were found to be energetically indistinguishable (within 0.5 kcal/mol range). The correct orientation of malonate in the binding cavity could not be confidently determined: optimisation of 8 different theoceptors, each with a different malonate starting position resulted in complexes with energies ranging from -8 to +16 kcal/mol (Figure 5.18).

These findings, in conjunction with ambiguity concerning the state of the protein during the experimental measurements[2], suggested that malonate may bind in the presence of cofactor NADH. Calculations employed the oxidised and reduced form of nicotinamide with the ribose present but with the phosphate moiety truncated to a methyl group[3] (Figure 5.19).

The binding energy to the theoceptor with co-factor in its reduced form was 5.4 kcal/mol, suggesting that it is not the correct form. The binding energy to the theoceptor with co-factor in its oxidized form is almost -40 kcal/mol (Table 5.8), which is consistent with the high ligand efficiency observed for this simple compound (Table 5.7).

The optimisation of the malonate theoceptor with the oxidised form of the theoceptor was then repeated, this time with cofactor reduced to just nicotinamide with a methyl group attached and the theoceptor binding energy was found to be approximately -47 kcal/mol.

These calculations achieved energies that were not changing substantially but which did not formally satisfy the optimisation criteria in Gaussian09. In such situations the

---

[2]R. Ward, personal communication

[3]To compute the energy of binding required an updated receptor energy, which involved the normal theoceptor with the co-factor also present. All the Cα and Cβ, and water oxygen atoms were fixed in space, together with the N-C bond between nicotinamide and ribose again reflecting the scaffolding effect of the remainder of the co-factor molecule in the bound state.

-3.9 kcal/mol   +8.1 kcal/mol   +3.1 kcal/mol

+13.9 kcal/mol   +16.0 kcal/mol   -8.2 kcal/mol

-1.0 kcal/mol   -5.5 kcal/mol
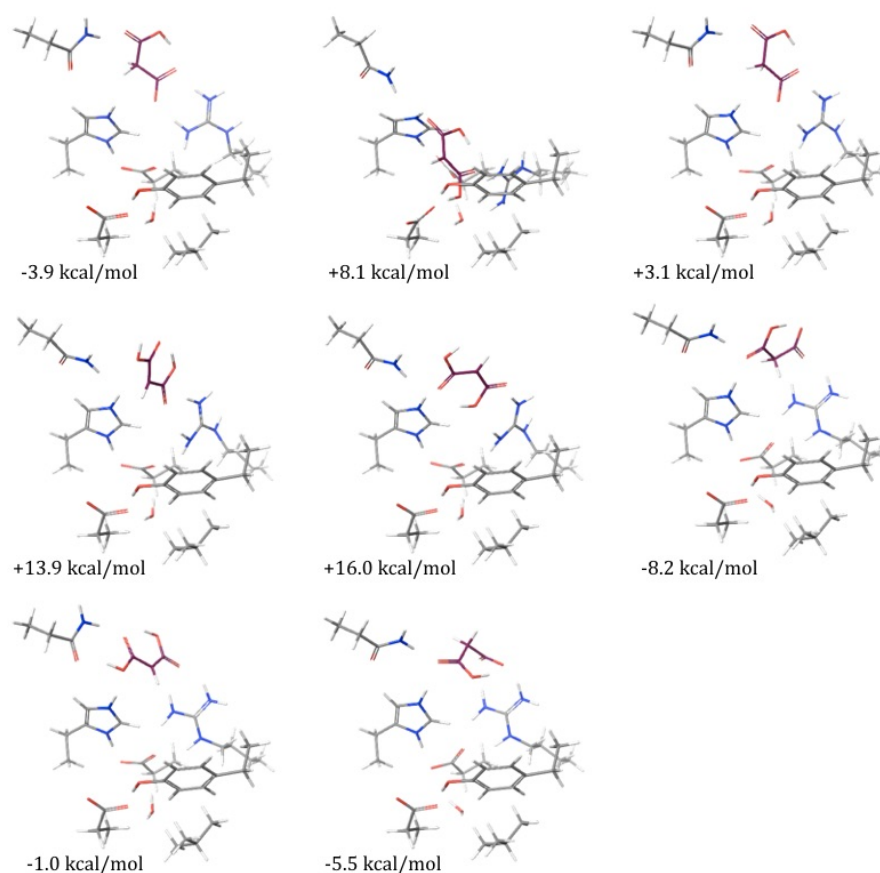
Fig. 5.18 Malonate theoceptors. Findings suggest cofactor is necessary for binding.

potential energy surface was expected to be quite "flat" and the geometry and energy were simply oscillating in a trivial fashion .

The next step was to assess binding of oxamate, a molecule that is isoelectronic and isosteric to pyruvate and as such is an excellent substrate mimic. As was the case with malonate, binding is facilitated by the presence of a positively charged portion of co-factor: the energy difference between the theoceptor with and without co-factor bound was more than 50 kcal/mol (Figure 5.20).

These findings are very interesting from the drug designer's point of view: small negatively charged molecules that mimic the substrate and bind in the presence of cofactor (rather than competing with it) have the tendency to be more ligand lipophilic efficient.
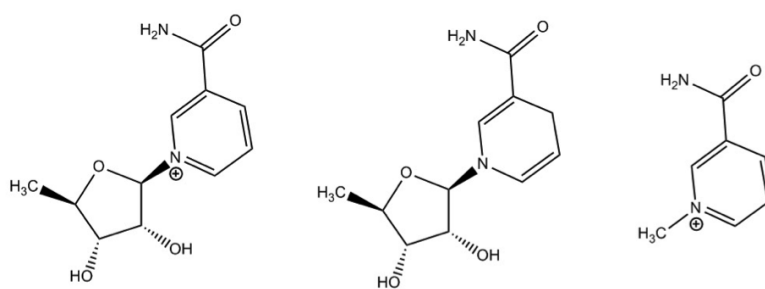
Fig. 5.19 Malonate theoceptors were first optimised in the presence of the oxidized and reduced forms of nicotinamide with the phosphate truncated to a methyl group. Optimisation was then performed with cofactor reduced only to nicotinamide with a methyl group attached.

Table 5.7 Chemical structure of malonate and oxamate. LLE was calculated as pK_d- clogP
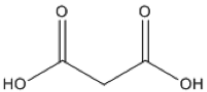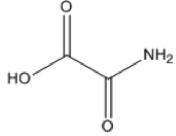
	malonate	oxamate
pK_d	2.3	4.58
clogP	-0.33	-1.07
LLE	2.63	5.66

Table 5.8 Complexation energies for malonate and oxamate in the presence of oxidised and reduced portion of cofactor.

| | Nicotinamide-ribose | | Nicotinamide |
	Reduced (kcal/mol)	Oxidised (kcal/mol)	Reduced (kcal/mol)
malonate	5.4	-39.5	-46.9
oxamate			-62.6
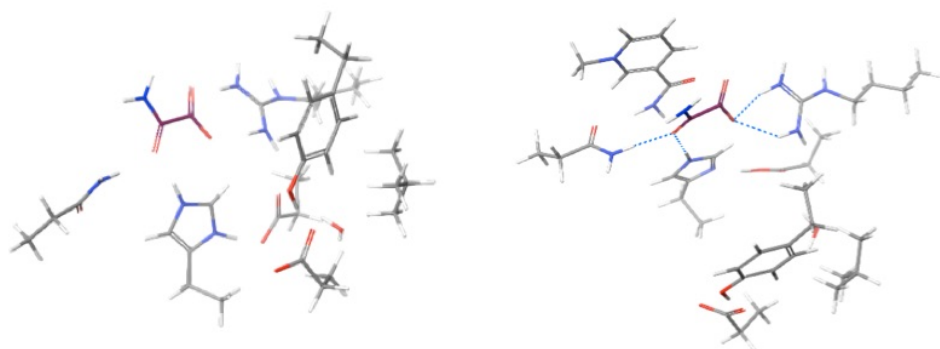
Fig. 5.20 Theoceptor for oxamate without cofactor and with truncated cofactor present. Complexation energies were -5.0 kcal/mol and -62.6 kcal/mol respectively.

5.8.1 Suggested compounds

With these findings in mind, the next step was to design compounds that would bind in the presence of cofactor and be more ligand lipophilic efficient. To test the hypothesis that small, negatively charged compounds are more efficient, the first compound was a fragment that is neutral at physiological pH. This fragment was positioned by modifying the malonate bound structure (Figure 5.21). Complexation energy of the theoceptor with compound **1** was -7.3 kcal/mol, which resulted in relatively low predicted LLE of 0.68. LLE was calculated from the equation of the line of the best fit in figure 5.8, and pK_d was then calculated from the LLE=pK_d-clogP relationship. When the theoceptor was optimised with the negatively charged fragment **2**, the predicted LLE was 5.04, which is in line with what would be expected given the oxamate/malonate binding. In compound **2_Me**, the methyl substituent is moved to position 5 of the isoxazolone ring, in order to explore if this exit vector could be used for fragment growing. Predicted LLE was 5.68, slightly better than predicted LLE for fragment **2**, where the methyl substituent is at position 4. Compound **3** is an oxadiazole ring with methyl substituent at position 3. This exit vector is slightly less supported, with predicted LLE of 5.39.

Compound 1
clogP=-1.01
ΔE=-7.3 kcal/mol
LLE=0.68
pKd=-0.33

Compound 2
clogP=0.04
ΔE=-57.5 kcal/mol
LLE=5.04
pKd=5.08
pKa=6.47

Compound 2_Me
clogP=0.04
ΔE=-64.9 kcal/mol
LLE=5.68
pKd=5.72
pKa=6.47

Compound 3
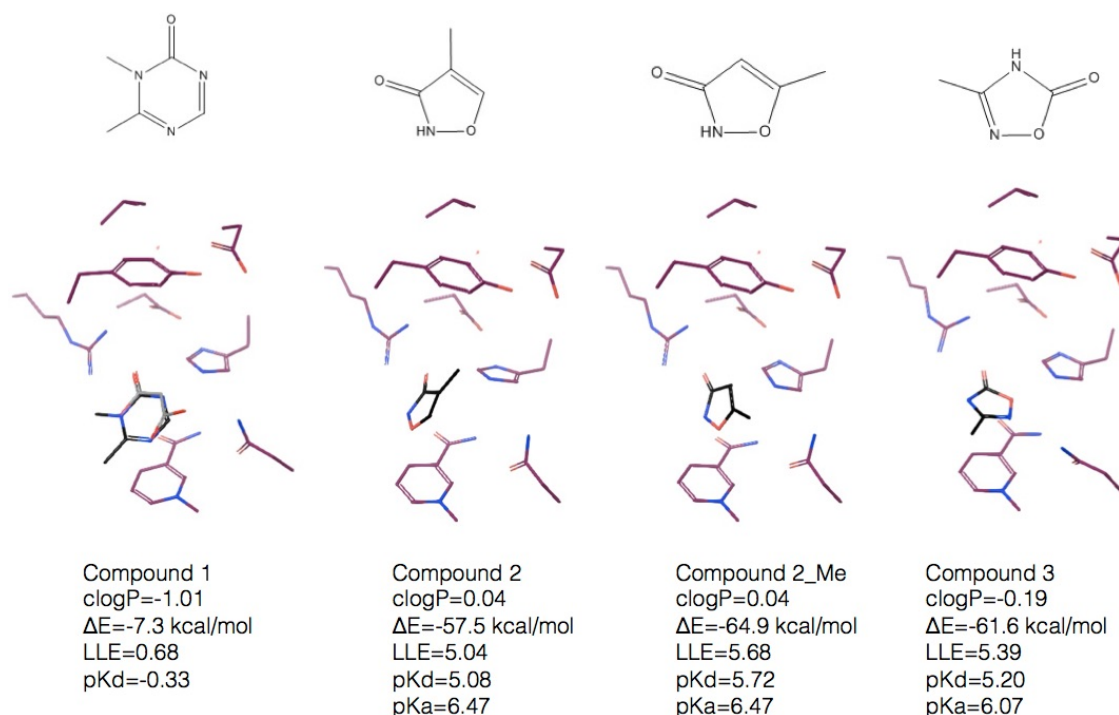clogP=-0.19
ΔE=-61.6 kcal/mol
LLE=5.39
pKd=5.20
pKa=6.07

Fig. 5.21 Structure, binding mode and predicted LLE of suggested compounds **1-3**. Compound **1** is shown with malonate superimosed.

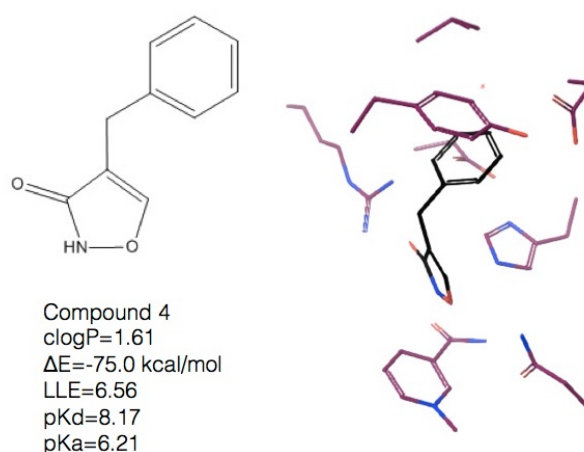Finally, the fragment was grown through addition of benzene to compound **2** (Figure 5.22).

Fig. 5.22 Structure, binding mode and predicted LLE of suggested compound **4**

The predicted LLE of this compound (compound **4**) was 6.56, which would be expected given the hypothesis that the smaller compounds that are negatively charged at the physiological pH and bind in the presence of the cofactor would experience higher LLE. The isoxazolone O acts as a hydrogen bond acceptor to Asn137 sidechain NH_2, while the oxo substituent forms a forked hydrogen bond with the sidechain of Arg168. The positively charged portion of the cofactor and the negatively charged portion of the fragment are within 4 Å of each other, contributing to the overall complexation energy.

5.8.2   Conclusions

This chapter details the efforts in computational modelling of different series of LDHA inhibitors and shows how quantum mechanical calculations can provide useful insight into interactions between protein and ligands. It demonstrates how this deeper understanding of nature of interactions between the binding partners can inspire new ideas- new fragment series is suggested based on the insights gathered from systematic analysis of the interactions between protein and ligands originating from a distinct fragment series.

It also shows how QM calculations can be used for determining conformational, tautomeric and stereochemical preferences and help elucidate uncertainties caused by poor crystallographical models.

Unfortunately, there was no opportunity to test these compounds experimentally, and the hope is that once they are published, our predictions will be verified (or disputed) by one of the interested parties.

Chapter 6

Binding of a homologous series of alkylbenzenes with lysozyme

6.1 Introduction

Synthesising the congeneric series of ligands is a standard process in ligand optimisation. The underlying assumption is that any minor change in the ligand would be accommodated with equivalently small, smooth amino acid sidechain movement. A fascinating recent study by Merski et al. showed that a homologous series of ligands bind to lysozyme with surprising changes in the protein structure and the positioning of the constant part of the ligands [204]. The F-helix adopts different discrete states in the same crystal: closed, intermediate and open, with proportions of each of the three conformations varying with the ligand. This suggests that even small ligand changes such as an addition of methyl group can cause conformational shifts between an ensemble of low-energy states, making prediction of the preferred geometry for proteins binding small molecules a significant challenge. A large number of such easily accessible states is available to most binding sites, so finding ways of distinguishing them from one another is important.

The problem has recently been studied using the FEP/REST[1] approach [205]. It was shown that even while using enhanced sampling techniques, the predicted relative binding free energies are sensitive to the initial protein conformational state. This work also highlights the importance of sufficient sampling of protein conformational changes and demonstrates how inclusion of three key protein residues in the "hot" region of the FEP/REST simulation improves the sampling and resolves this sensitivity, given enough simulation time.

---

[1]free energy perturbation/replica exchange with solute tempering, REST is an enhanced sampling approach

This chapter aims to investigate whether quantum mechanical calculations are able to correctly identify which ligands would bind preferentially to which protein conformations and therefore provide a solution (of the "approaches that sample among precalculated states" class) to the computational challenge posed by Merski et al.

6.2 Discrete protein conformational states

Structures of lysozyme in complex with benzene, ethylbenzene, *n*-propylbenzene, *sec*-butylbenzene, *n*-butylbenzene, *n*-pentylbenzene and *n*-hexylbenzene were determined at high resolution (1.60-1.80 Å) and refined as a multiconformer model (PDB codes: 4W52 – 4W59) [204]. *n*-propylbenzene, *sec*-butylbenzene, *n*-butylbenzene, *n*-pentylbenzene and *n*-hexylbenzene were observed in two major conformations, based on the orientation of the flexible F-helix loop, between the residues 107 and 115 (Figure 6.1). In the apo and benzene bound structure, the closed conformation was the dominant one. This conformation occupied approximately 90% of the observed electron density, while the remaining 10% was unassigned. In the toluene bound structure, the binding cavity starts to open to the bulk solvent. This is the intermediate conformation and it occupies approximately 20% of the observed density. This conformation becomes more dominant as the alkyl sidechain of the ligand grows through the series. In ethyl-, *n*-propyl- and *sec*-butylbenzene structures the closed and intermediate conformations appear in about equal proportions. In the *n*-butylbenzene bound complex, a third, open conformation starts to appear. This conformation coexists with closed and intermediate, occupying different proportions of the observed density. The three discrete states are characterised by a hydrogen-bonding pattern between the backbone residues of the F-helix. This pattern changes as the binding cavity expands, with the open conformation having the lowest number of these intramolecular bonds.
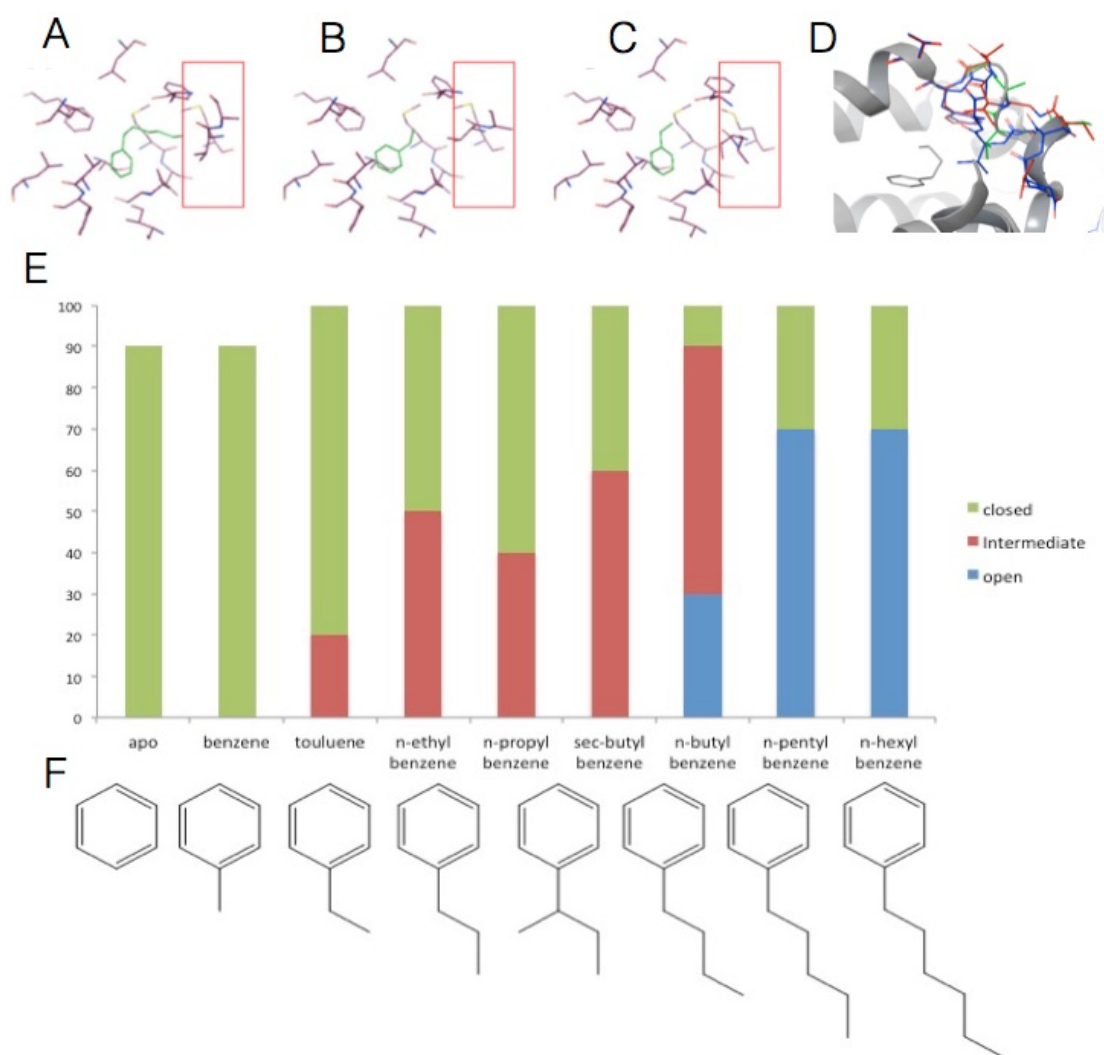
Fig. 6.1 Three protein conformations observed in the binding of congeneric ligands to lysozyme. The residues that undergo the largest changes are highlighted in the red box. The preferred binding conformation of the ethylbenzene (PDB code 4W54) is shown for the closed and intermediate protein conformations (A and B). Open protein conformation (C) is shown with *n*-hexylbenzene bound (PDB code 4W59). All three conformations overlayed (D). Proportions of each of the three conformations varying with the ligand (E). Full congeneric series (F).

6.3 Theoceptor approach

6.3.1 Theoceptor construction

Eight crystal structures were split into their constituent chains and superimposed. Residues that were within 5 Å of the *n*-hexylbenzene in 4W59 were selected as a part of the theoceptor: Ile78, Leu84, Val87, Tyr88, Leu91, Ala99, Met102, Val103, Met106, Vall11, Ala112, Phe114, Leu121, Leu133, Phe153. The shape of the binding cavity was well captured with these fifteen residues (228 atoms in total). Therefore, the two residues that were also within the 5 Å of the ligand (Gly107 and Leu118) were excluded from the selection to decrease the number of degrees of freedom and speed up the calculations. Protonation states were assigned manually and the net charge of the theoceptor was 0. Backbone atoms and the remainder of the protein were deleted. To account for the scaffolding effect of the rest of the protein, Cα and Cβ atoms were fixed in space using the fixed Cartesian redundant coordinate feature in Gaussian09. The geometries were optimised *in vacuo* at the M06 level of theory with the 6-31+G* basis set. The free ligand structures were optimised at the same level of theory but with solvation incorporated, using the polarisable continuum (PCM) water model.

Vibrational corrections, solvation of the receptor and the complex are essentially the same across all the structures, and they were treated as a constant. The affinity was then defined as the energy difference between the lowest energy of the complex, free ligand and a constant:

$$\Delta G = \Delta E_{elec} + (\Delta E_{vib} + \Delta E_{solv})$$

$$= \Delta E_{elec} + \Delta \underbrace{E_{vib} - E_{solv_{complex}} - E_{solv_{receptor}}}_{\text{constant}} - E_{solv_{ligand}}$$

$$= \Delta E_{elec} - E_{solv_{ligand}} - const$$

Constants were determined for open, closed and intermediate conformation. Benzene closed conformation was then set to zero. Calculated complexation energies are reported relative to benzene closed conformation.

6.3.2 Complexation energies

For the smallest ligand (benzene), the calculations accurately predict the most stable conformation to be the closed one (Figure 6.2). The energy of this conformation is the lowest of the three, which is in agreement with the refinement. This value is set to zero and all the other energies are relative to this one. For toluene, calculations show ambiguous results- all three conformations seem to be energetically equal, even though the

Table 6.1 The calculated complexation energies, relative to the benzene closed conformation. Energies are in units kcal/mol.

	closed	intermediate	open
benzene	0.0	2.3	3.6
toluene	-1.2	-1.2	-1.0
ethylbenzene	-4.0	-3.7	-3.3
n-propylbenzene	-6.7	-5.6	-4.8
sec-butylbenzene	-5.6	-7.7	-4.7
n-butylbenzene	-3.9	-7.0	-6.7
n-pentylbenzene	-11.1	-7.9	-9.9
n-hexylbenzene	-12.4	-9.2	-12.9
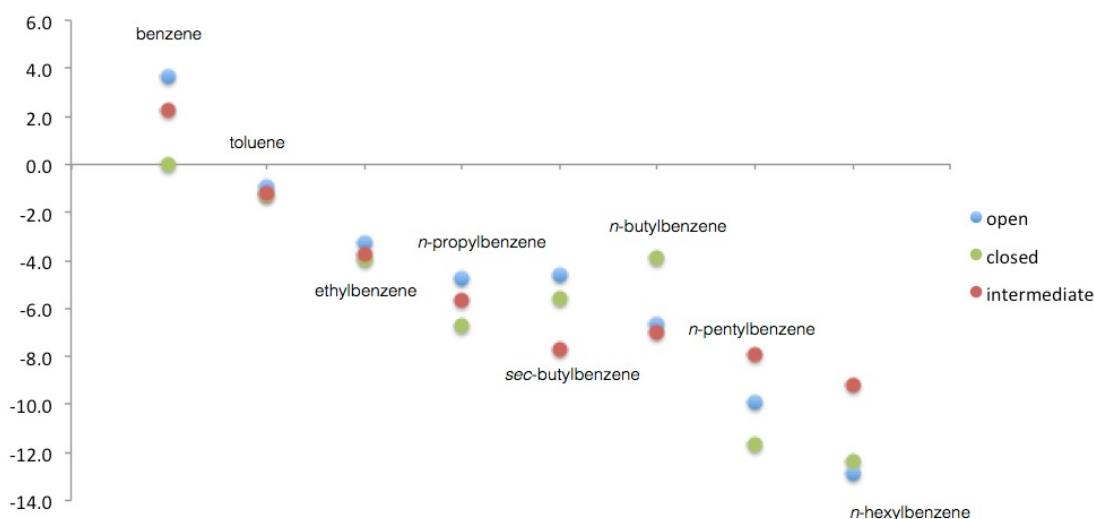
Fig. 6.2 The calculated complexation energies.

crystallographical refinement suggests that the closed conformation is much more abundant than the intermediate, while the open should not be present at all. Ethylbenzene is refined as a mixture of closed and intermediate conformations in equal proportions: this is partially captured in the theoceptor calculations where the conformations are correctly ranked (closed -4.0 kcal/mol, intermediate -3.7 kcal/mol), but the open conformation that is not present according to the original refinement is only 0.4 kcal/mol higher in energy than the intermediate one. *n*-propylbenzene calculations correctly predict the closed conformation to be the dominant one, followed by the intermediate. *sec*-butylbenzene is also calculated to be a mixture of the intermediate and closed, but this time the intermediate conformation is energetically more stable. *n*-butylbenzene is a mixture of all three conformations with

intermediate being the dominant. For *n*-pentylbenzene, QM calculations predict the closed conformation to be the lowest energy one, which is in disagreement with the structural refinement that suggests the open conformation should be prevalent. This is somewhat surprising, given that the QM model accurately pinpoints the open *n*-hexylbenzene as the most stable one.

n-pentyl- and *n*-hexylbenzene were crystallographically modelled in two conformations (open and closed) with the occupancies rounded to the nearest 10%. In these two structures, the closed conformation was modelled without the ligand and is considered to be in fact in an apo state. *n*-butylbenzene, a structure where all three loops are present was also modelled without the ligand in the closed state. However, calculations suggest that the ligand binds in the closed conformation as well and occupies a low energy state. This is further investigated in the following section.

6.4 Multiconformational refinement

Given that the optimised theoceptor structures give the coordinates of the preferred protein conformations and ligand positions, the structures of the complexes were refined with these additional ligand and F-loop conformations. Because the work of Merski et.al used an earlier version of the software, the available PDB structures were first refined in the 1.10.1 version of PHENIX [206], and ligand parameters were generated using eLBOW (PHENIX package) [207] to ensure that the refinement statistics can be compared with the new models. Structures refined with the newer version of the software were in perfect agreement with the original ones, with the R_{work} to R_{free} ratios for original and re-refined structures within ± 0.05, as shown in table 6.2. The absolute differences in R-factors between the refinements in the two versions of the software is most likely caused by the changes in the bulk solvent corrections between the two versions [208].

Table 6.2 Refinement parameters. R_{work} and R_{free} from the work of Merski et al, 're-refinement' are the same structures refined with the newer version of the software, 'theoceptor' denotes the refinement based on the theoceptor calculations.

		Merski et.al		re-refinement		theoceptor	
	PDB	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}
apo	4W51	0.175	0.191	0.160	0.168	0.160	0.168
benzene	4W52	0.165	0.182	0.155	0.179	0.155	0.179
toluene	4W53	0.174	0.205	0.148	0.166	0.149	0.164
ethylbenzene	4W54	0.165	0.198	0.138	0.173	0.138	0.173
n-propylbenzene	4W55	0.166	0.187	0.140	0.160	0.140	0.160
sec-butylbenzene	4W56	0.171	0.188	0.141	0.158	0.141	0.158
n-butylbenzene	4W57	0.170	0.185	0.143	0.164	0.144	0.164
n-pentylbenzene	4W58	0.161	0.184	0.146	0.168	0.141	0.162
n-hexylbenzene	4W59	0.176	0.187	0.145	0.148	0.144	0.148

6.4.1 Apo and benzene bound structure

Both theoceptor calculations and structural refinement suggest that the closed conformation is the dominant one in apo and benzene bound structures. Looking at the electron density maps of the structures refined with the F loop as a single, closed conformation, no patches of positive or negative electron density are observed (Figure 6.3, left hand-side). When the structures are refined with all three loop conformations, patches of positive (green) and negative (red) electron density around the protein can be observed, indicating atom misplacement. This is accompanied with higher R_{free} values (R_{free} increased to 0.173). The patches grow as the occupancy of the alternative loops is increased. Therefore, as it was in the Merski et al. refinement, the closed conformation occupies roughly 90% of the observed electron density, while 10% remains unmodelled.
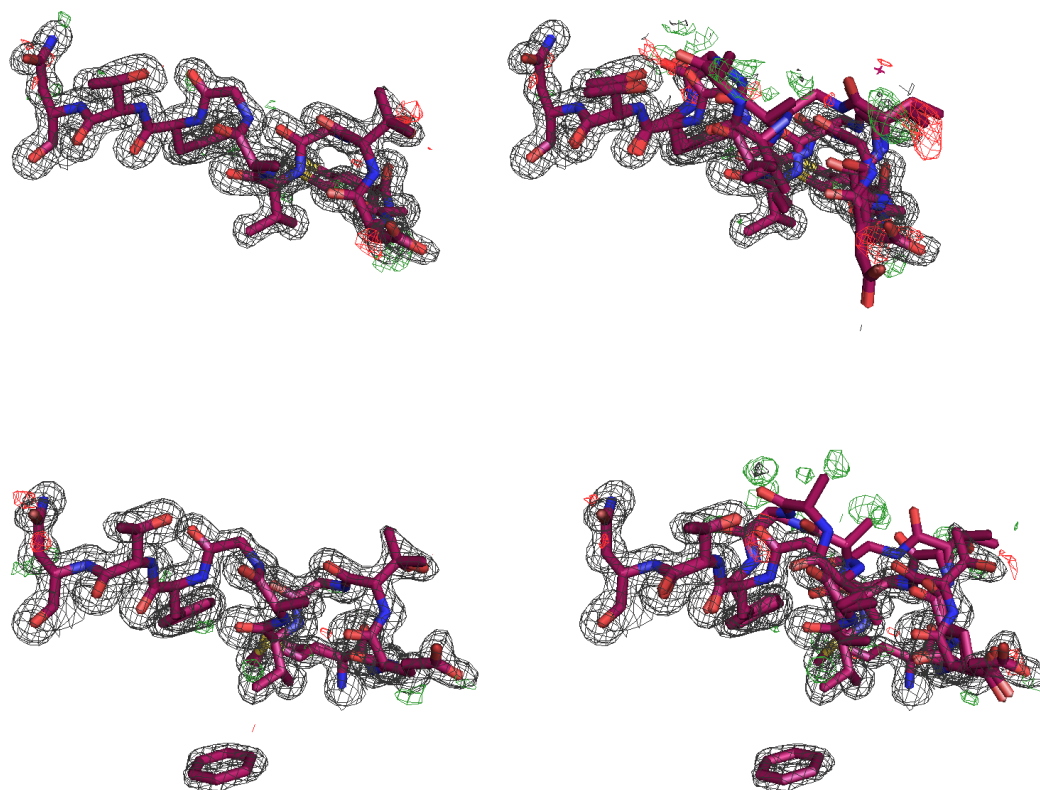
Fig. 6.3 Electron density maps for structural refinement of apo and benzene bound structures. On the left are the re-refined structures where the F loop is in a closed conformation and on the right are the structures with all three loop conformations. Negative electron density (red mesh, mF_o-DF_c,$\pm\sigma$) suggest that it is highly unlikely that any conformation other than closed is present in a proportions that can be confidently determined.

6.4.2   Toluene, ethyl- and *n*-propylbenzene bound structures

In the work of Merski et al., toluene was modelled with the closed conformation occupying roughly 80% of the observed density, while the remaining 20% was assigned to the intermediate conformation. QM calculations suggest closed and intermediate conformations are energetically equal, and open to be only slightly less favourable (+ 0.2 kcal/mol). When refined with all three loops, no additional patches of negative or positive electron density appear (Figure 6.4). The presence of a third loop is also supported by the refinement statistics: R_{free} slightly decreases from 0.166 to 0.164. The proportions of conformations are assigned as follows: 70% open, 20% intermediate and remaining 10% open.

Ethyl- and *n*-propylbenzene bound structures were observed as a mixture of closed and intermediate conformations. In the work of Merski et al., a 10% occupancy cutoff for modelling alternative conformations was applied. Looking at the electron density maps of

the structures with the two F loop conformations (Figure 6.4, left hand-side), small patches of positive (green) electron density can be observed. These structures were refined with the F loop in all possible conformations applying 5% occupancy cutoff criterion. The addition of the third, open loop did not affect the R_{free} value in the case of the *n*-propylbenzene bound structures. In the ethylbenzene bound structure, there was a small increase (from 0.173 to 0.178). Looking at the electron densities of these newly refined structures (Figure 6.4 right hand-side), it can be observed that there are no patches of negative (red mesh) electron density around the protein sidechain that would suggest that the atoms have been modelled in the wrong place. However, the occupancy of this additional loop seems to be below 10%. In the case of the *n*-propylbenzene bound structure, the closed conformation remains at 60%. The intermediate conformation occupies 30% when the third loop is present and the 10% difference is assigned to the open conformation.

The closed conformation as the dominant one is also confirmed by the theoceptor energies. Theoceptors in closed conformation were lowest in energy, followed by the intermediate conformation. This difference is not as prominent in the ethylbenzene bound structure, but becomes more obvious as the ligand increases in size: intermediate and open conformation can hardly be energetically distinguished from one another in the ethylbenzene complex, while in the *n*-propylbenzene complex, there is a 1 kcal/mol difference between the states.
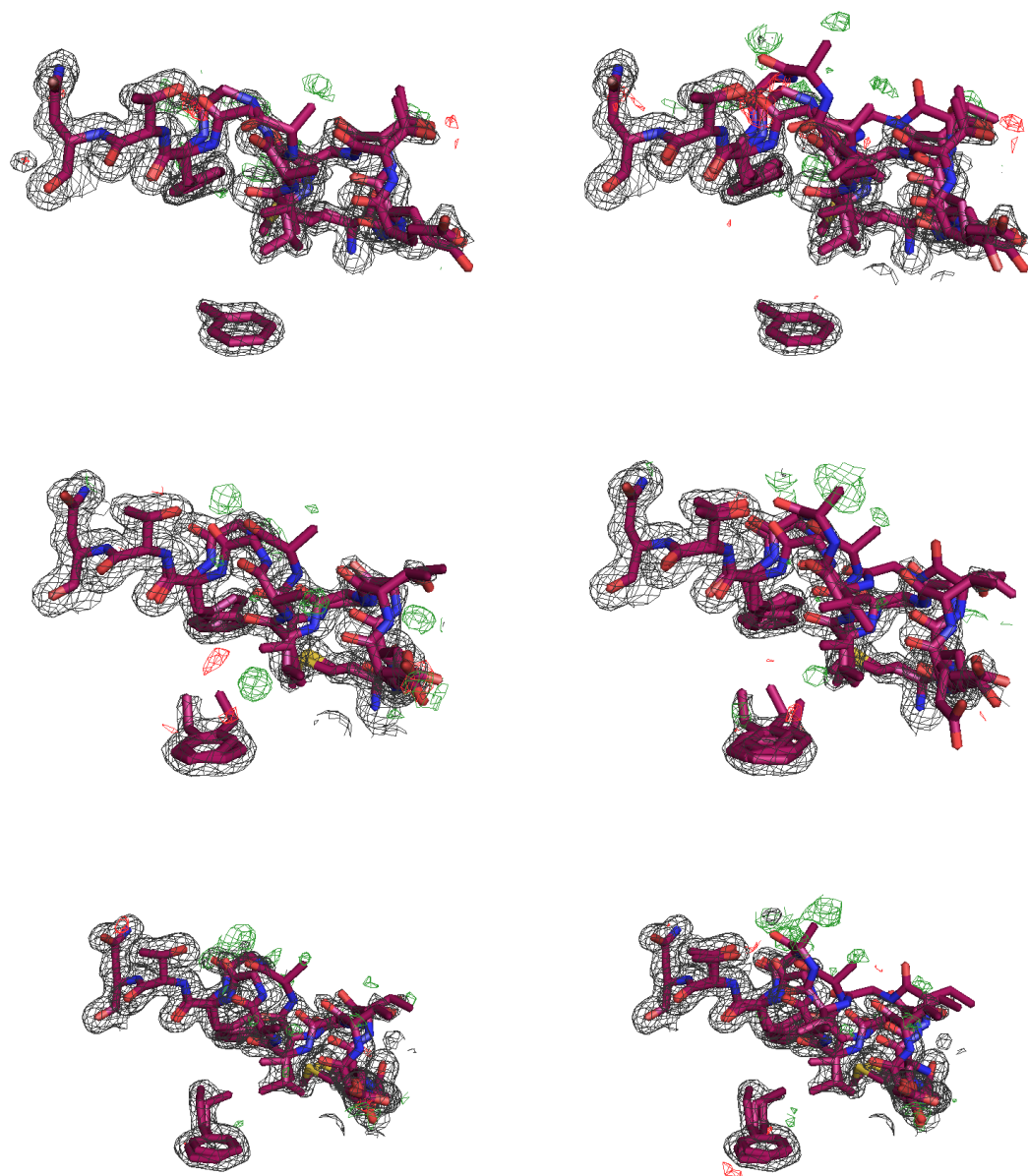
Fig. 6.4 Electron density maps for structural refinement of toluene, ethyl- and *n*-propylbenzene bound structures. On the left are the re-refined structures where the F loop is in the closed and intermediate conformation and on the right are the structures with all three loop conformations. Positive electron density (green mesh, mF_o-DF_c, $\pm 1\sigma$) indicates areas where the current model fails to place sufficient electrons. Negative electron density (red mesh, mF_o-DF_c, $\pm 1\sigma$) indicates areas where the current model places too many electrons.

6.4.3   *sec*-butylbenzene bound structures

sec-butylbenzene was crystallographically modelled by Merski et al as the *R*-isomer with the F loop in intermediate and closed conformations in a 60 to 40 ratio. Looking at the electron density map in Figure 6.5, two small patches of negative electron density can be observed around the ligand aromatic ring and alkyl sidechain.
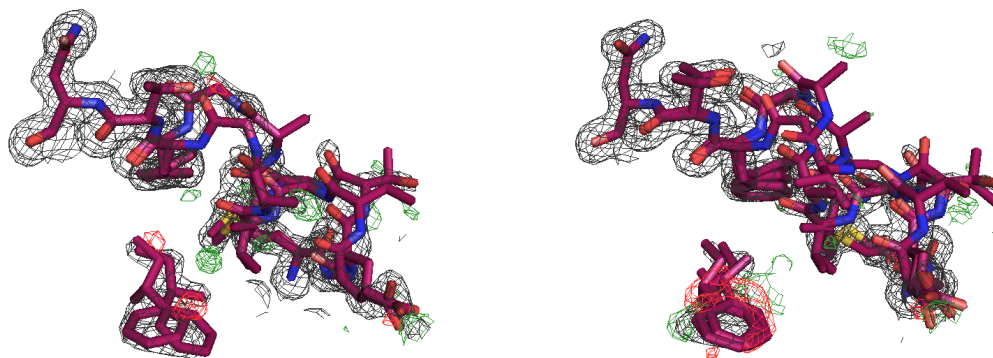

Fig. 6.5 Left: *sec*-butylbenzene was modelled as *R*-isomer with the F loop in intermediate and closed conformations in a 60 to 40 ratio. Right: *sec*-butylbenzene refined with all three loops and both *R* and *S* ligands.

When the theoceptor was optimised with the ligand structures extracted from the PDB files, the closed conformation was calculated to be the most stable one (-5.6 kcal/mol, intermediate -4.6 kcal/mol and open -4.7 kcal/mol). All three complexes were again optimised, this time with *R*-isomers. In closed and open conformations, there is a preference for the *R*-isomer (*R* closed -5.6 kcal/mol, *S* closed -3.9 kcal/mol; *R* open -4.7 kcal/mol, *S* open -1.3 kcal/mol). This preference is even more prominent in intermediate structures: theoceptor with the *S*-isomer was 5 kcal/mol higher in energy than the one optimised with the *R*-isomer (+1.1 kcal/mol compared to -4.6 kcal/mol).

Looking at the electron density around the ligand, it can be observed that is it not as sharp and well defined as it is for other members of the series, especially around the aromatic ring, a portion of the ligand present in each member. This suggests that the ligand can occupy different conformations in the binding cavity. This was further investigated using QM (Figure 6.6).

Additional calculations were first performed for the intermediate structures, as based on refinement, this one would be expected to represent the most stable state. A structure optimised with the ligand as observed in the open state was 2 kcal/mol higher in energy than the starting intermediate (-2.6 kcal/mol compared to -4.6 kcal/mol) while several other attempts where the starting ligand conformation was created by comparison to the

density performed even worse-these structures were 9.7 kcal/mol and 9.3 kcal/mol and 2 kcal/mol higher in energy than the one extracted from the PDB.
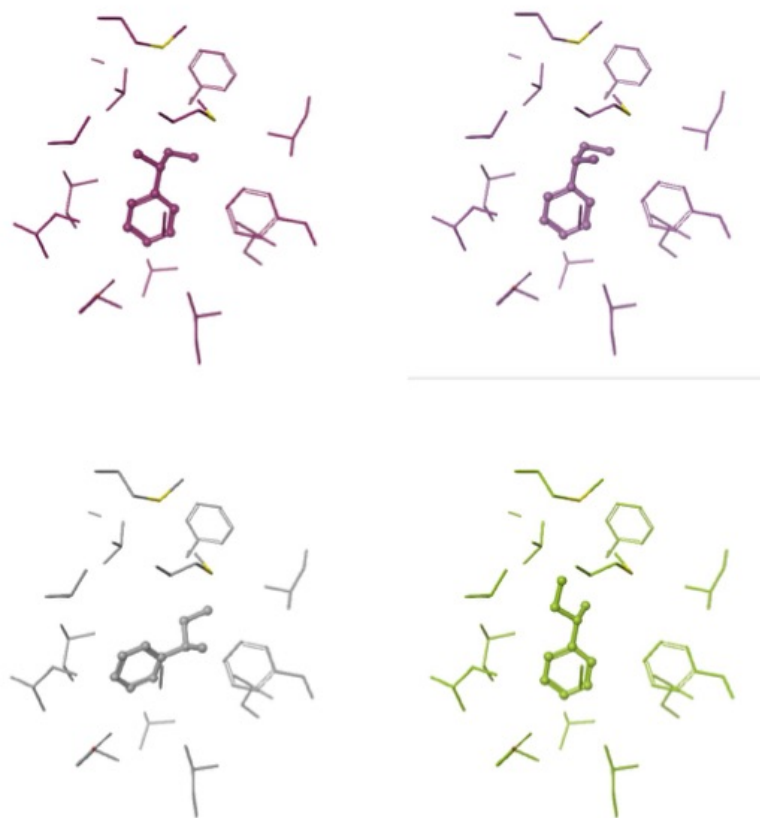

Fig. 6.6 The intermediate complex of *sec*-butylbenzene was optimised with different ligand conformations. Top left: the theoceptor with ligand *R*-isomer ligand structure as deposited in the PDB (purple), the *S*-isomer (violet) was 5 kcal/mol higher in energy compared to *R*. Second row: additional ligand structures that resulted in higher complexation energies. Structure shown in grey is 2 kcal/mol higher, and structure shown in green is 9.7 kcal/mol higher in energy than the one with *R*-isomer ligand structure as deposited in the PDB (shown in purple).

The lowest energy structure was determined to be -3.1 kcal/mol lower than the PDB one (-7.7 kcal/mol compared to -4.6 kcal/mol). Somewhat surprisingly, no stereochemical preference could be observed- the *S*-isomer was only 0.3 kcal/mol higher in energy (-7.4 kcal/mol) (Figure 6.7).

When the structure was refined with all three loops and both *R* and *S* ligands from the lowest energy structures (Figure 6.5 right hand side), patches of both negative and positive electron density around the ligand indicating atom misplacement can be observed. This is accompanied by the increase in R_{free} value, from 0.158 to 0.165. The occupancies are
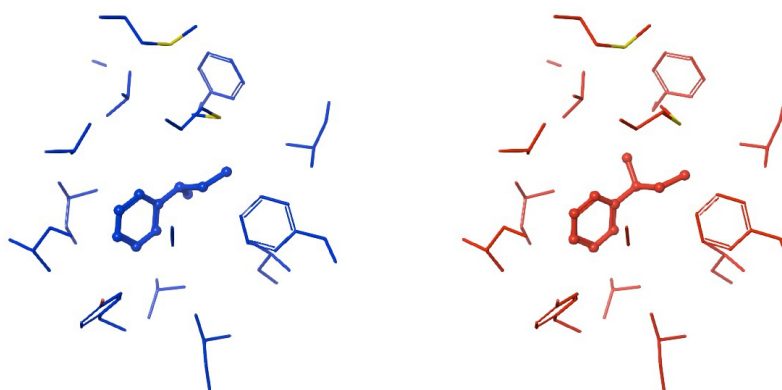
Fig. 6.7 *R* and *S sec*-butylbenzene isomers. These complexes are -3.1 (*R*-isomer, shown in blue) and -2.8 kcal/mol (*S*-isomer, shown in red) lower in energy than the theoceptor with the PDB ligand structure.

assigned as follows: 20% to both *S* and *R* intermediate and closed, and 20% to *R* and 0% *S* open.

The *S* ligand in the open conformation was then excluded from the refinement and the resulting occupancies were set to 25% for *S* and *R* intermediate, 20% *S* closed, 10% *R* closed and 20% *R* open. R_{free} increased to 0.164. Again, the model was a poor fit to the observed electron density.

When refined with just the two loops (closed and intermediate) with the ligands in *R* and *S* conformation, R_{free} slightly increases to 0.161. More importantly, there are no additional patches of negative electron density present (Figure 6.8). The occupancies were assigned as follows: *S* intermediate 30%, *R* intermediate 25%, *R* closed 45% and *S* closed 0%. What is interesting here is that the stereochemical preference is present in closed conformation, but not in intermediate, which was suggested by the theoceptor calculations. Different ligand conformations that are supported by the density and the fact that stereochemical preference seems to exist in one of the protein conformations but not in the other makes this refinement the most interesting and challenging one.
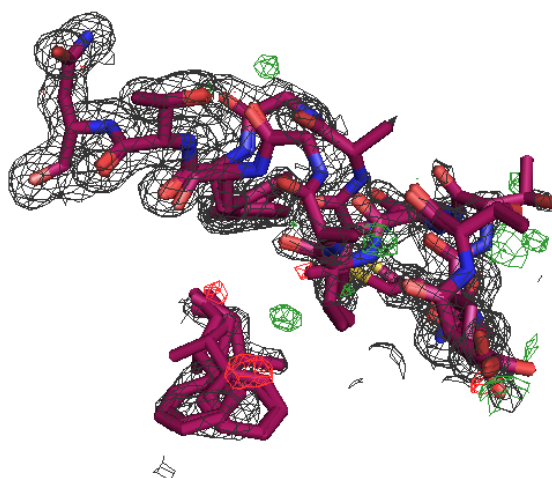
Fig. 6.8 *sec*-butylbenzene refined with closed and intermediate loop and *R* and *S* ligand isomers.

6.4.4 *n*-butyl- *n*-pentyl- and *n*-hexylbenzene bound structures

All three loop conformations are present in *n*-butylbenzene. In the work of Merski et al., the ligand was modelled in intermediate and closed protein conformation, but not in the open.

Fig. 6.9 Electron density maps for structural refinement of *n*-butylbenzene bound structure. Left is the structure where F loop is in all three conformations, with the ligand modelled in intermediate and open conformation, as it was in the work of Merski et al. On the right is the structures with all three loop conformations, as well as all three ligand conformations. Added closed ligand conformation is shown in orange.

When the refinement was repeated, this time with all three ligand structures, no change in the refinement statistics was observed. Visual inspection of the 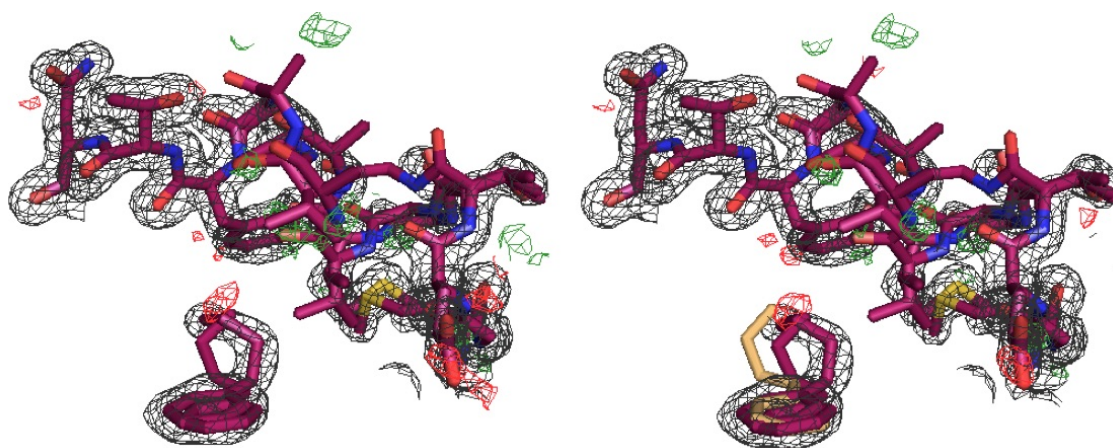electron density maps showed that this third ligand conformation can be tolerated, as no patches of negative electron density appeared (Figure 6.9).

In the work of Merski et al, the *n*-pentyl bound structure was modelled as a mixture of open (70%) and closed (30%) conformation. Ligand was modelled in open, but not in the closed conformation. When looking at the density of this refinement (Figure 6.10), it can be seen that when the F loop is in closed conformation, it protrudes in what is ligand density in open conformation. No alternative ligand conformation is suggested.

Fig. 6.10 The *n*-pentylbenzene bound structure was refined as 70% open (left) and 30% closed conformation, as it was in the work of Merski et al. The closed conformation is considered to be unoccupied by the ligand (right).

When refined as three loops and using theoceptor derived ligand geometries, there is a slight decrease in R_{free} from 0.168 to 0.166 when occupancies are set to 60%, 30% and 10% for open, closed and intermediate conformations. Looking at the electron density for this refinement (Figure 6.11), small patches of negative electron density around the ligand alkyl chain and benzene ring can be observed.

When the intermediate conformation was excluded from the refinement, R_{free} decreased to 0.162, however, some negative electron density around the benzene ring can still be observed (Figure 6.12), as the ring is not fitting the density well.
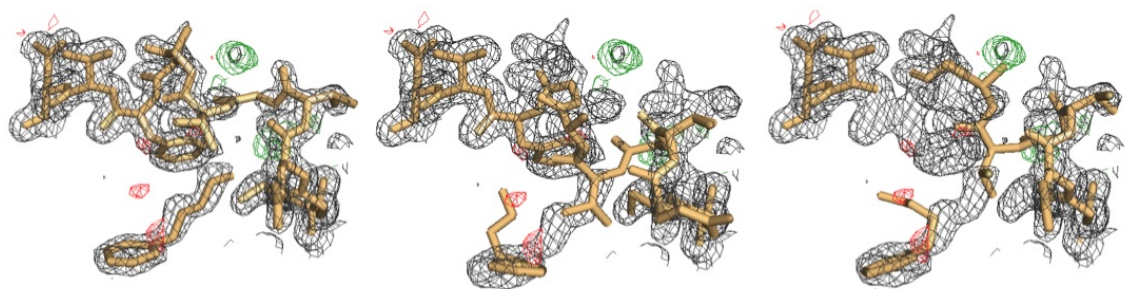
Fig. 6.11 When *n*-pentylbenzene was refined as three loops and using theoceptor derived ligand geometries, occupancies were set to 60%, 30% and 10% for open, closed and intermediate conformations. There is a patch of negative electron density around the ligand.

Fig. 6.12 Deletion of the intermediate loop ligand conformation resulted in an improvement in the refinement statistics. The proportion of open and closed conformations changed to 65% for open and 35% for closed. Patches of negative electron density are present around the benzene ring.

Similarly, the *n*-hexyl bound structure was modelled as a mixture of open (70%) and closed (30%) conformations. The ligand was modelled in open, but not in the closed conformation. When looking at the density of this refinement (Figure 6.13), it can be seen that when the F loop is in closed conformation, it protrudes in the ligand density. Again, no alternative ligand conformation is suggested.

Addition of the intermediate loop and theoceptor derived closed and intermediate ligand conformations resulted in no change in the refinement statistics, R_{free} remained the same, and the proportion of open and closed conformations was 60% open, 35% closed and the small remaining portion was assigned to the intermediate conformation. Like in the *n*-pentylbenzene refinement, patches of negative electron density are present around the ligand (Figure 6.14).
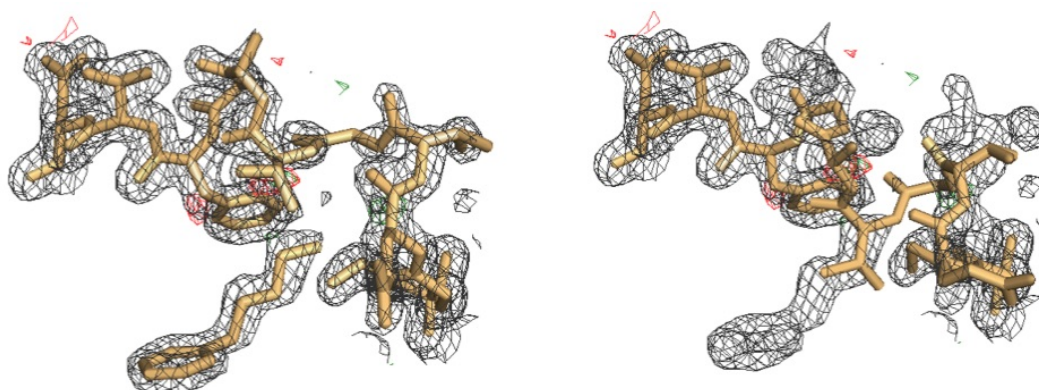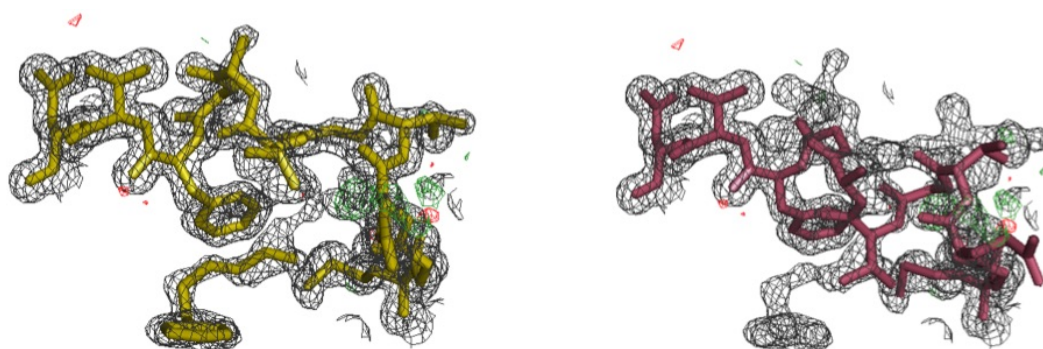
Fig. 6.13 The *n*-hexylbenzene bound structure was refined as 70% open (left) and 30% closed conformation, as it was in the work of Merski et al. The closed conformation is considered to be unoccupied by ligand (right).
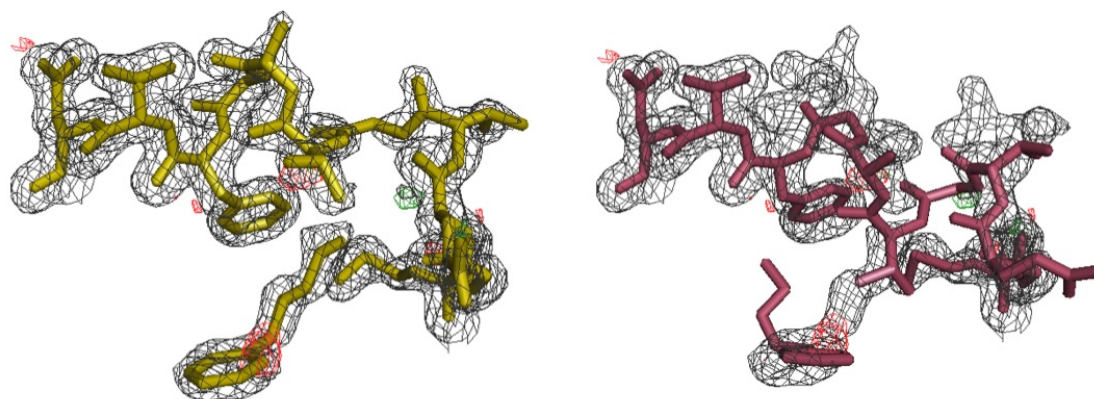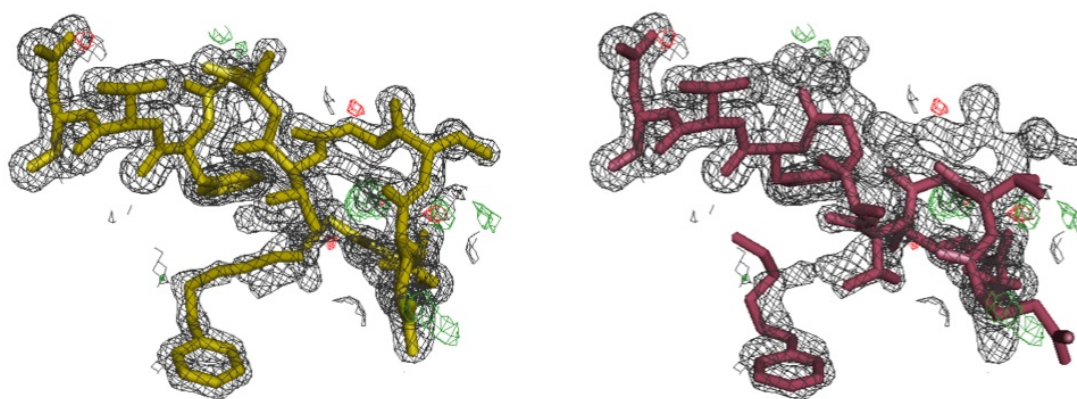

Fig. 6.14 Deletion of the intermediate loop ligand conformation resulted in no change in the refinement statistics. The proportion of open and closed conformations changed to 65% for open and 35% for closed. Patches of negative electron density are not present.

When the intermediate conformation is removed from the refinement, the patches disappear (Figure 6.15), not affecting the refinement statistics. This implies that the *n*-hexylbenzene complex is indeed in open and closed conformation, but also that the ligand can be present in the closed conformation. The proportion of open and closed conformations was 65% and 35%.

Fig. 6.15 Addition of the intermediate loop and theoceptor derived closed and intermediate ligand conformations resulted in no change in the refinement statistics. The proportion of open and closed conformations changed to 60% for open, 35% for closed and the small remaining portion was assigned to the intermediate conformation.

6.4.4.1 Is the closed conformation an apo state?

Whether or not the ligand is actually present in the closed conformation, or these are in fact apo states is not clear from the refinement. QM calculations imply ligand not only can be present in t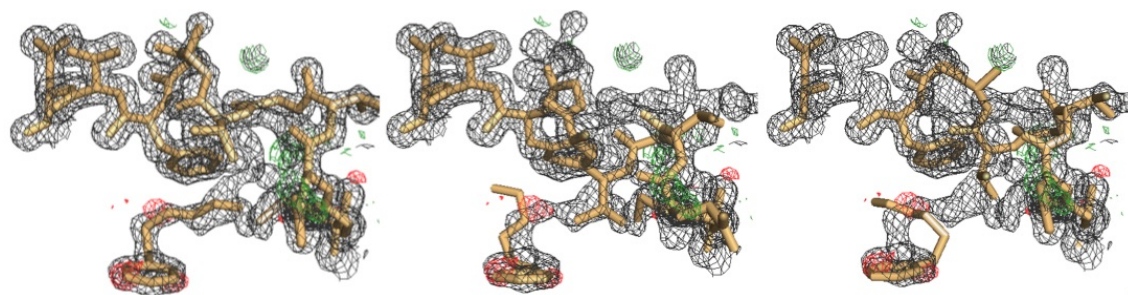he closed conformation, but can also occupy a low energy state. Some small patches of electron density indicating atom misplacement do appear when alternative ligand conformations are added to the *n*-pentylbenzene refinement, which suggest that the ligand conformation might not be optimal. It is interesting to note that in the FEP/REST study [205], simulations with the ligands in the closed structures were successfully run. All this indicates that at this point, it can not be excluded that the ligand is present in the closed conformation. In the case of *n*-butyl and *n*-hexylbenzene structures, the conformation of the ligand that binds to the closed theoceptor suggested by the QM calculations seem to fit the density well. However, in the case of the *n*-pentylbenzene, this conformation is not optimal. It is tempting to speculate whether the correct conformation can be found by optimising the complex with different ligand starting structures.

6.4.5 Partial occupancies of theoceptor derived structures

To examine how these energies are distributed over the three possible conformations[2], energies were converted to Boltzmann populations:

$$\frac{n_i}{N} = \frac{e^{(-E_i/kT)}}{\sum\limits_{n=j} e^{(-E_j/kT)}}$$

where n_i is number of molecules in the ith energy state, N is the total number of molecules, E_i is the energy of the ith energy state (open, closed and intermediate in this case), k is the Boltzmann constant and T is temperature (298 K).

The previous section showed that the theoceptor approach performs well when it comes to selecting the most stable of the three possible conformations. Comparing occupancies (Table 6.3) to what would be expected given the refinement (Figure 6.16) makes it possible to assess how successful the approach is in assigning relative distributions.

Table 6.3 Theoceptor occupancies

	theoceptor			refinement		
	open	intermediate	closed	open	intermediate	closed
benzene	<5	<5	95	<5	<5	90
toluene	20	40	40	10	20	70
n-ethylbenzene	15	35	50	<5	50	50
n-propylbenzene	<5	15	85	10	30	60
sec-butylbenzene	<5	95	<5	<5	60	40
n-butylbenzene	40	60	<5	30	60	10
n-pentylbenzene	10	<5	90	65	<5	35
n-hexylbenzene	70	<5	30	65	<5	35

For benzene, the approach works very well: the benzene bound structure was refined to occupy 90% of the density, while the remaining 10% was left unmodelled, similar to theoceptor calculations that predict 95% of the benzene ensemble to be in the closed state. For toluene, theoceptors predict a mixture of all three conformations, with intermediate and closed in equal proportions (40%) and the remaining 20% is assigned to open. This is in a slight disagreement with the refinement where 70% is assigned to closed, 20% to intermediate and the remainder to the open conformation. In the case of the ethylbenzene, 50% is assigned to closed, the same as it is in the refinement. 35% is assigned to intermediate (50% in the refinement) and the remaining 15% to open (less than 5% in the refinement). n-propylbenzene slightly underpredicts the intermediate conformation (15% compared to 30% in structural refinement) in favour of closed conformation (85% compared to 60% in the refinement).

In the case of the sec-butylbenzene, calculations seem to underpredict the population of the closed conformation. The intermediate conformation is the dominant one (95% in

---

[2]At a given temperature, the energy of a collection of molecules is distributed among the various modes of motion (that is, translational, rotational, vibrational, electronic, ...). All distributions are equally likely; however, the particles in the distributions are indistinguishable. Therefore, some distributions are more probable than others. This principle is known as the equipartition of energy. As the numbers get larger, that is 10^{23}, the most probable distribution for a given amount of macroscopic energy among microscopic energy states is given by the Boltzmann distribution.

theoceptor derived populations and 60% in refinement). These discrepancies are probably linked to the stereochemical preference that seems to exist when the protein is in a closed conformation, but not when the ligand binds to the intermediate conformation, which makes selecting the lowest energy ligand geometry in the binding site a challenging task.

In the *n*-butylbenzene, 60% is assigned to intermediate (60% in the refinement), 40% to open (30% in the refinement), while 10% of closed conformation observed in the refinement is not present in the significant proportions among calculated populations.

In the case of *n*-pentylbenzene, the open conformation seems to be underpredicted by the calculations. This is somewhat surprising, given that the calculation results for *n*-hexylbenzene are in very good agreement with the refinement (70% is assigned to open and 30% to closed, compared to 65% and 35% in the refinement). Likely explanation is that the suggested geometry of the ligand in the open conformation is not the optimal one. Different ligand conformations in theoceptor were tested in this work, however, lower energy complexes could not be found.
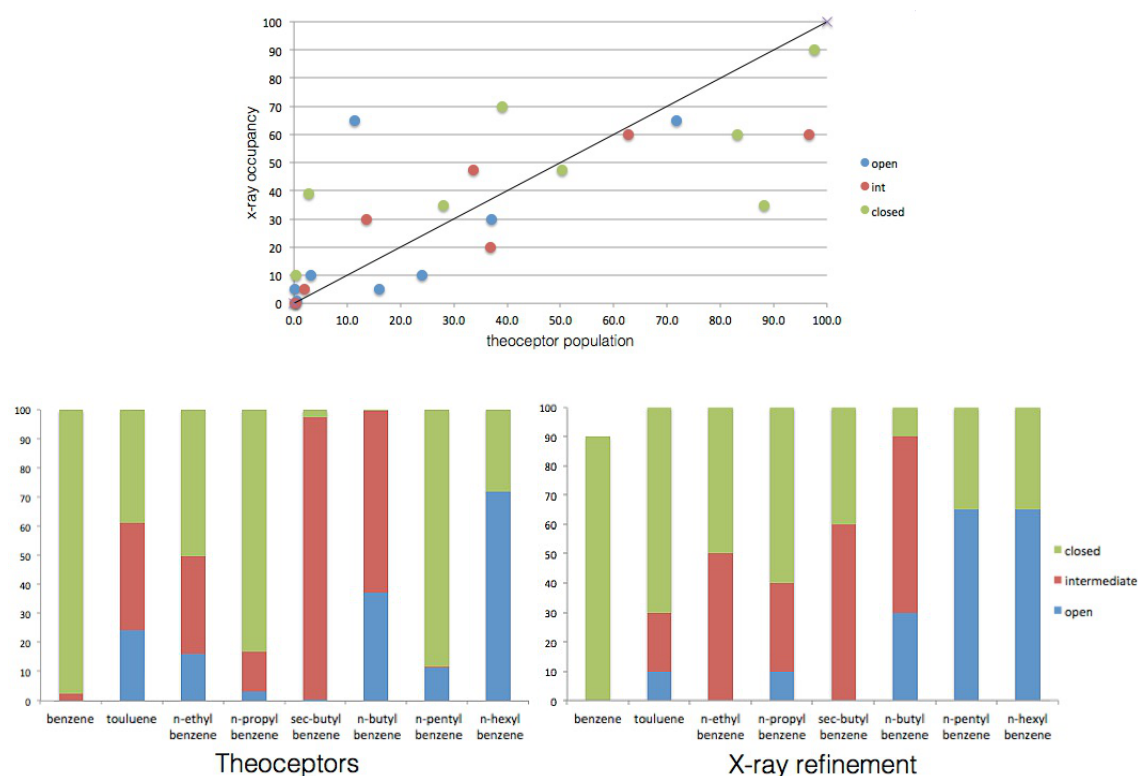

Fig. 6.16 Crystallographic occupancies plotted against theoceptor derived occupancies. Theoceptor derived occupancies plotted as histograms. Crystallographic occupancies plotted as histograms.

6.4.6   Conclusions

The goal of the work presented in this chapter was to investigate whether quantum mechanical calculations are able to correctly identify which ligands would bind preferentially to which protein conformations. The results suggest that the method is capable of detecting the preferred protein conformation in all cases but one (*n*-pentylbenzene). Furthermore, it was possible to predict the relative populations with reasonable accuracies as well: two cases where this was ambiguous were *sec*-butylbenzene and *n*-pentylbenzene. In case of *sec*-butylbenzene this could be linked to stereochemcial preferences and it is encouraging to see that the theoceptor approach can help elucidate stereoselectivity in this case as well. Theoceptor calculations also suggested possible larger ligand conformations that bind to the closed state, something that could not be inferred from crystal structures alone. In two of the cases where the ligand conformation was guessed from the calculations (*n*-butylbenzene and *n*-hexylbenzene), the ligand seems to fit the the density with good accuracy, whereas in the case of *n*-pentylbenzene, the suggested ligand conformation presents a poor fit to the density map. Given the success with modelling *n*-butylbenzene and *n*-hexylbenzene structures, it is not excluded that such conformations exist for *n*-pentylbenzene as well.

Chapter 7

3-Phosphoglycerate dehydrogenase (PHGDH)

7.1 Introduction

3-Phosphoglycerate dehydrogenase (PHGDH) is an enzyme that catalyses the transformation of 3-phosphoglycerate into 3-phosphohydroxypyruvate. This is the first and rate-limiting step in the phosphorylation pathway of serine bionsynthesis, using NAD^+/NADH as a cofactor (Figure 7.1)[209].

Serine supports a number of anabolic processes, including protein, lipid, and nucleic acid synthesis. Overexpression of the gene encoding the PHGDH has been identified in human cancers, particularly in breast cancers and melanomas [210–212]. The identification of a small molecule inhibitor of PHGDH would not only enable thorough preclinical evaluation of PHGDH as a target in cancers, but would also provide a tool with which to study serine metabolism [213].

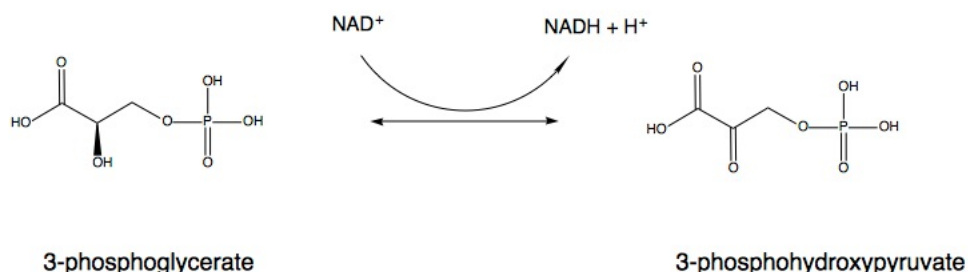Up until recently, no small molecule inhibitors of 3-phosphoglycerate dehydrogenase had been reported [213].

Fig. 7.1 Reaction catalysed by 3-phosphoglycerate dehydrogenase

Work detailed in this chapter describes the efforts in characterising thermodynamic and kinetic signatures of some of the compounds identified in the joint project between AstraZeneca and the Cancer Research Trust. Around 30 inhibitors for this protein target have been tested experimentally using SPR and ITC, however, it was possible to collect only data for a limited number of structures (Table 7.1). These examples are discussed here.

7.2 PHGDH inhibitor selection and potency

Structures of the twelve inhibitors for which the data were collected are shown in Table 7.1, together with the inhibitor potency that was retrieved from the AstraZeneca database. Around thirty inhibitors were initially selected for thermodynamic and kinetic characterisation, however, it was only possible to collect data for the twelve structures shown here (possible reasons will be discussed in a subsequent sections). The twelve inhibitors were divided into three groups, based on common features. It can be noted that all twelve structures are weak inhibitors, with pIC_{50} close to or below the assay threshold, even though the starting set contained inhibitors spanning a wide pIC_{50} range.

Table 7.1 Structures and inhibitory potency of selected PHGDH inhibitors. Potency data presented as pIC_{50} (pIC_{50}=-log_{10}(IC_{50}))

1	**6**	**9**
$pIC_{50} = < 3$	$pIC_{50} = 3.59$	$pIC_{50} = < 3$

R=
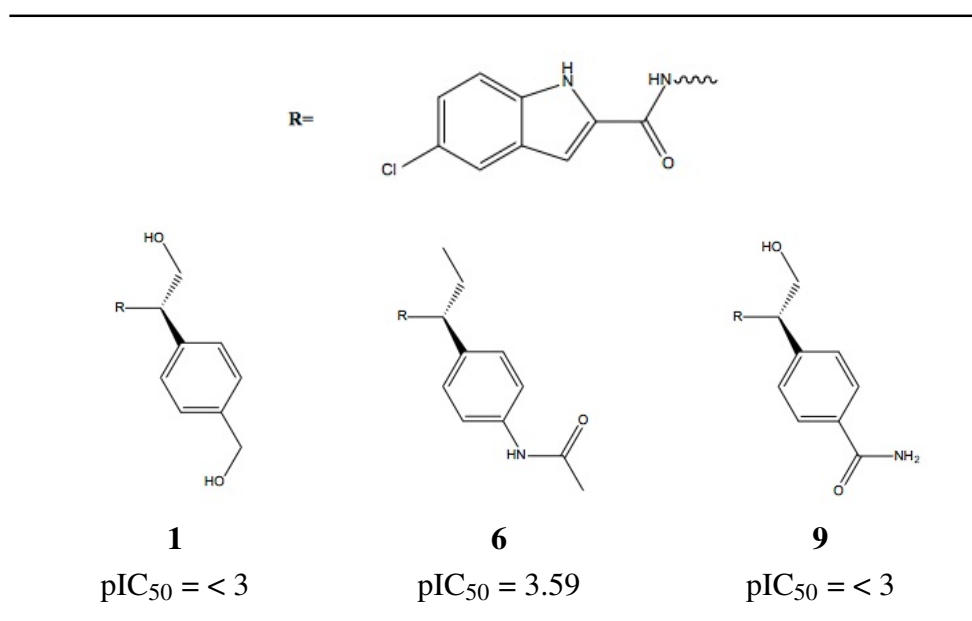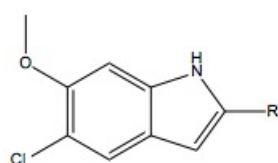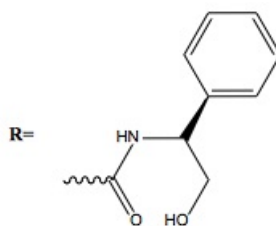
4
pIC$_{50}$ = 3.6

7
pIC$_{50}$ = 3.3

R=
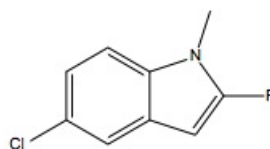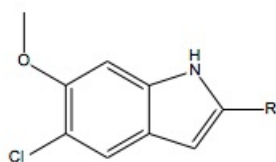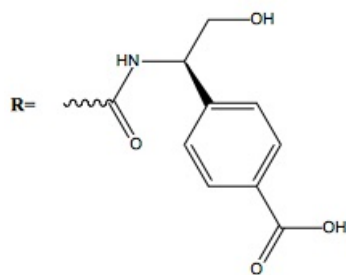
2
pIC$_{50}$ = < 3
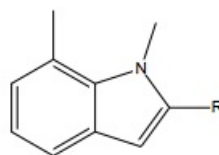
5
pIC$_{50}$ = < 3
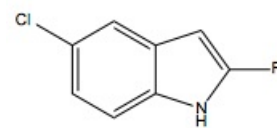
8
pIC$_{50}$ = < 3
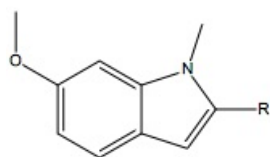
10
pIC$_{50}$ = < 3
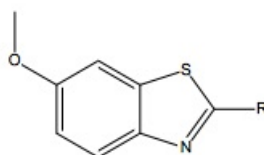
11
pIC$_{50}$ = <3
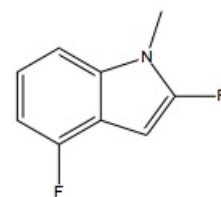
12
pIC$_{50}$ = < 3
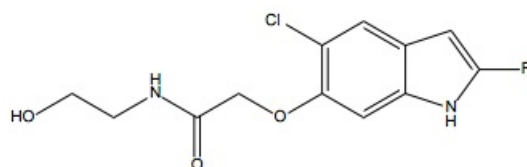
3

$pIC_{50} = < 3$

7.2.1   Structural analysis

The structures of the complexes with six inhibitors (**1-4**, **6**, **12**) were retrieved from the AstraZeneca corporate database. The high resolution of these structures ($\sim 2\text{Å}$) enabled fine details of the interactions to be seen.

Given that there is no difference in the inhibitor potency between these twelve inhibitors, no information on structure-activity relationship (SAR) could be provided. Comparison of the inhibitor binding modes as well as the accompanying thermodynamic and kinetic fingerprints should provide more detailed insights into molecular determinants governing binding of these three groups of inhibitors.

Across all the structures, the most prominent protein-ligand interaction observed is a hydrogen bond between Asp174 and the central amide hydrogen. This interaction is present in all three series. Interactions with water were observed in some of the crystal structures. These water molecules are not conserved and are therefore not considered a part of the binding site. Edge-to-face π stacking between Tyr173 and the indole ring can also be observed in all of the available crystal structures. Members of the series with acid substituent (**2**, **3**, **5**, **8**, **10-12**) form an additional salt bridge with the guanidinium moiety of Arg235 and hydrogen bonds with two water molecules. Because representative crystal structure from each of the three groups were available and the binding mode was well conserved through all groups, the other six complex structures (**5**, **7-11**) were modelled based on their closest analogue. The same binding mode was observed in all the modelled structures-all the inhibitors interact with the hydrophobic environment consisting of Pro175, Ile176, Ile177, Pro207 and Leu215 (Figure 7.2). The binding site is well conserved. The only noticeable difference is orientation of Arg235: in the presence of the COO⁻ group, the guanidinium moiety of Arg235 is pointing towards the binding site, forming a salt bridge with a deprotonated acid group. Arg235 is in the same orientation in **1**, which could be explained by formation of a weak hydrogen bond between its hydroxyl group and Arg235. In other structures, this residue is pointing away from the binding site and into the solvent.
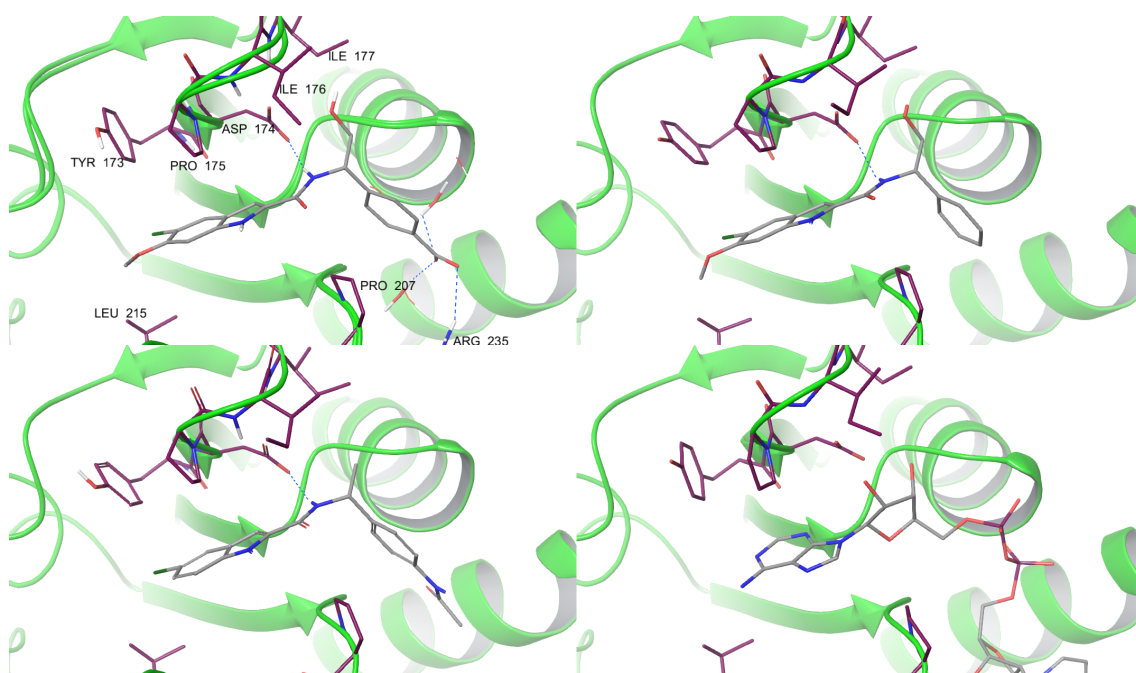
Fig. 7.2 x-ray structures of PHGDH in complex with inhibitors **2**, **4**, **6** and cofactor NAD (PDB: 2G76). Protein backbone is show as green cartoon, inhibitors, cofactor and binding site residues as sticks.

7.3 Docking studies

To verify and support the approach used for modelling six complex structures (**5**, **7-11**), ligands were docked using AutoDockVina. Available crystal structures were processed with Schrödinger's protein preparation wizard [197] to ensure a reasonable starting point for docking: hydrogen atoms were added after deleting any original ones, followed by adjustment of bond orders for amino acid residues and the ligand. The protonation and tautomeric states of Asp, Glu, Arg, Lys and His were adjusted to match a pH of 7.4. Docking studies were performed with AutoDock Vina 1.1.2 [196]. Ligands were extracted from the crystal structures of the complexes and polar ligand hydrogens were added with AutoDock Tools (ADT 1.5.4) [198]. Two types of docking experiment were performed: redocking into the inhibitor bound conformation and docking of ligands **5**, **7-11** into the crystal structure of **2** from which the ligand has been removed.

All the ligands had comparable docking energies (±0.7 kcal/mol) with several incorrect docking poses, therefore this approach can not be used to rank the ligands based on their affinity. However, docking was useful in terms of determining the ligand orientation: in all the docked cases, a hydrogen bond between Asp174 and the ligand's amide hydrogen could be observed, as well as aromatic π stacking between Tyr173 and the ligand's indole ring. AutoDock Vina did not position the acid group of the inhibitor correctly-no interactions with Arg235 were found (Figure 7.3).
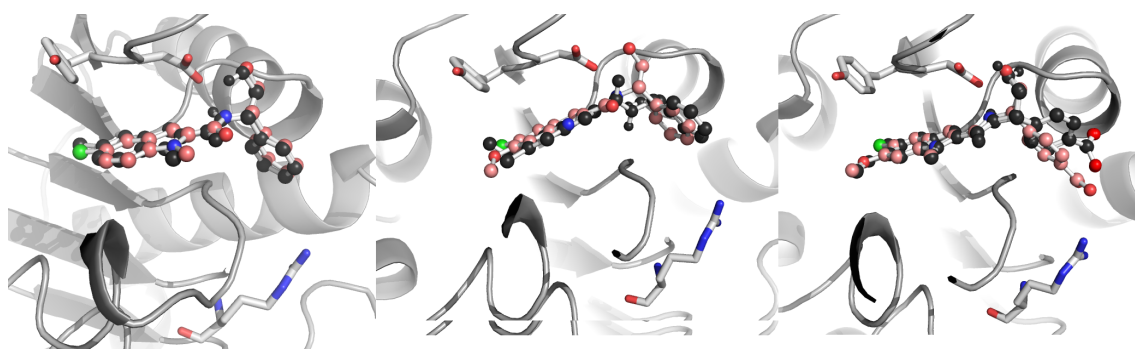
Fig. 7.3 Redocking of the crystal structures of **6**, **4**, and **12**. The ligand from the crystal structure is shown in pink, the docked ligand in black

Docking results were similar for modelled structures: the acid moiety of **5**, **8**, **10** and **11** was placed away from the Arg235 and key interactions with Tyr173 and Asp174 were preserved for all the structures.

7.4 Biophysical properties

To evaluate whether the structural differences translate into different thermodynamic and kinetic effects, the thermodynamic and kinetic profiles were analysed. Changes in thermodynamic parameters were determined by isothermal titration calorimetry (ITC) and binding kinetic data were obtained using surface plasmon resonance (SPR) analysis. Despite the similarity in the pIC_{50} values, some distinction in the measured binding affinity can be observed (Table 7.2).

Observed equilibrium dissociation constant (K_d) values measured by ITC and SPR follow the same trend (Figure 7.4).
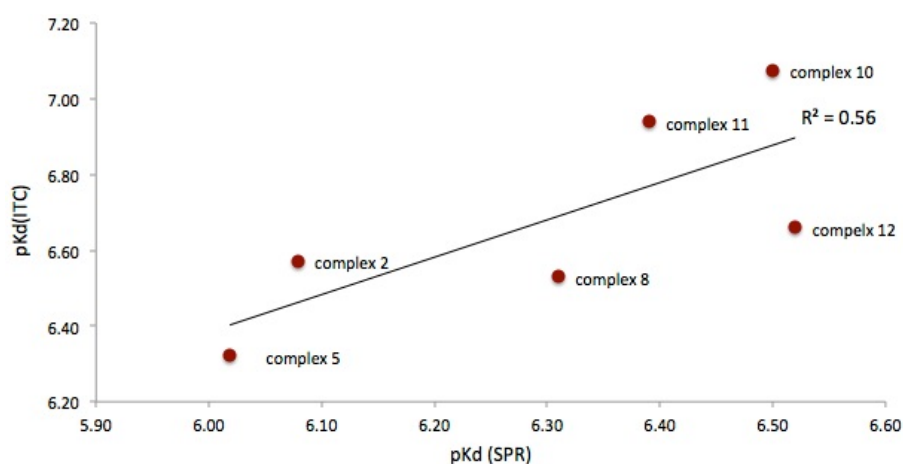
Fig. 7.4 Observed equilibrium dissociation constant (K_d) values measured by ITC and SPR follow the same trend.

Table 7.2 Inhibitory potency and binding affinity of PHGDH inhibitors. pIC_{50} values were retrieved from the AstraZeneca database, pK_d values were measured using SPR and ITC.

	pIC_{50}	pK_d(SPR)	pK_d(ITC)
1	<3	4.54	NC
2	<3	6.07	6.57
3	<3	6.61	NC
4	3.6	5.12	NC
5	<3	6.01	6.32
6	3.59	5.27	NC
7	3.3	5.33	NC
8	<3	6.3	6.50
9	<3	NC	NC
10	<3	6.5	7.02
11	<3	6.3	6.82
12	<3	6.5	6.66

[*] NC=not calculated

7.4.1 Thermodynamic analysis

7.4.1.1 Experimental setup

ITC experiments were carried out using an ITC200 instrument (GE Healthcare) using a protein system that has been passed through a NAP-5 column (GE Healthcare) equilibrated with a) 50 mM potassium phosphate, 150 mM KCl, 1 mM TCEP and 0.005 %T20, 1% DMSO b) 20 mM TRIS, 150 mM NaCl, 1 mM TCEP, 0.005% T20, 1% DMSO c) 50 mM HEPES, 150 mM NaCl, 1 mM TCEP, 0.005% T20, 1% DMSO, all pH 7.5.

Protein concentration was determined by measuring the absorbance at 280 nm using a theoretical molar extinction coefficient. DMSO concentration in the protein solution was adjusted to 1% (v/v). Concentrations of ligand stock solutions in DMSO were determined by weight of a compound. Final ligand concentrations were achieved by diluting 1:100 (v/v) in the experimental buffer resulting in a final DMSO concentration of 1% (v/v). ITC measurements were performed at 25 °C.

In order to check whether the observed binding enthalpy was perturbed by proton transfer upon ligand binding, titrations in all three buffers (a, b and c) were performed. Complete titration of 25 μM protein was carried out in the 200 μL sample cell using 2μL injections of 0.2 mM compound every 180 s. Raw data were collected and the area under each peak was integrated. The experimental data were fitted to a theoretical titration curve (single site binding model) using MicroCal Origin 7 software. Thermodynamic parameters were calculated from the equation:

$$\Delta G = \Delta H - T\Delta S = -RT \ln K_a = RT \ln K_d$$

where ΔG, ΔH and ΔS are the changes in free energy, enthalpy and entropy of binding respectively, T is the absolute temperature and R is the gas constant: 1.986 cal mol^{-1}K^{-1}.

7.4.1.2   Results and discussion

Compound **1** was identified as the weakest PHGDH inhibitor, whereas other compounds exhibited comparable affinities. All the compounds show similar half maximal inhibitory concentrations (IC$_{50}$), so no meaningful comparison with the trends in binding affinity could be achieved. In a typical enzyme inhibition assay, the inhibitor and substrates are serially diluted to achieve concentrations in the assay that span their respective binding constants. The addition of cofactor is necessary to initiate the enzymatic reaction. Although the inhibitor is necessary for enzyme activity, it is not necessary for inhibitor binding. This suggests that all weak binding inhibitors shown here are competing with cofactor NAD for binding in the PHGDH *in vitro* enzymatic assay.

To check if the observed enthalpy is perturbed by any protonation effect, titration experiments were performed in the buffers of different ionisation enthalpies (Appendix 1). The enthalpies were essentially the same in two of the buffer systems (HEPES and TRIS), which have significantly different heats of ionisation. There were some differences in enthalpies measured in phosphate buffer: however, superimposing **5**, **8**, **10**-**12** to a cofactor bound structure (PDB code 2G76) revealed that the phosphate moiety of NAD and inhibitors occupy the same space, suggesting that inhibitors and phosphate are competing for binding (Figure 7.5). Therefore, enthalpies measured in phosphate buffer were excluded from further analysis. Corrected values (geometric means of the values obtained from the experiments performed in HEPES and TRIS buffer) are summarised in table 7.3.

Table 7.3 Thermodynamic data for inhibitors binding to PHGDH[a]

	ΔH (kcal/mol)	-TΔS (kcal/mol)	ΔG (kcal/mol)	K_d (µM)
2	-9.22 ± 0.25	0.29 ± 0.25	-8.92 ± 0.35	0.26 ± 0.00
5	-8.01 ± 0.15	-0.66 ± 0.16	-8.58 ± 0.22	0.47 ± 0.05
8	-8.23 ± 0.10	-0.63 ± 0.12	-8.87 ± 0.15	0.29 ± 0.03
10	-10.09 ± 0.18	0.39 ± 0.25	-9.61 ± 0.31	0.08 ± 0.02
11	-10.48 ± 0.12	1.05 ± 0.15	-9.43 ± 0.19	0.11 ± 0.01
12	-8.94 ± 0.15	-0.11 ±0.14	-9.04 ± 0.18	0.21 ± 0.02

[a] ITC titrations were performed at 298.15 K. Data represent geometric mean ± SE from at least two independent experiments
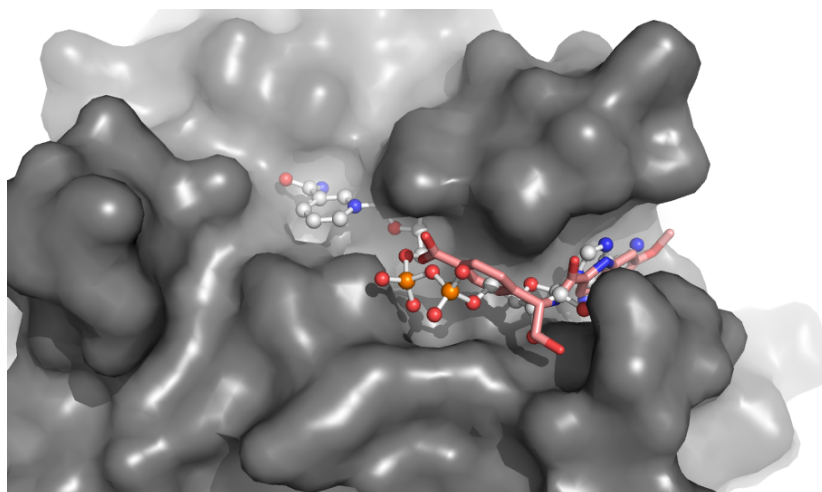
Fig. 7.5 Cofactor bound structure superimposed to **1**. NAD is shown as ball and sticks and inhibitor **1** as pink sticks.

ΔG values of all inhibitors are within ±1 kcal/mol, which translates to a 1 fold change pK_d, which is insufficient for structure-activity relationship analysis. There is a small variation in observed thermodynamic signatures. All tested inhibitors exhibited strongly enthalpically driven binding. Three of the inhibitors tested, **2**, **10** and **11**, have unfavourable entropic component (Figure 7.6).
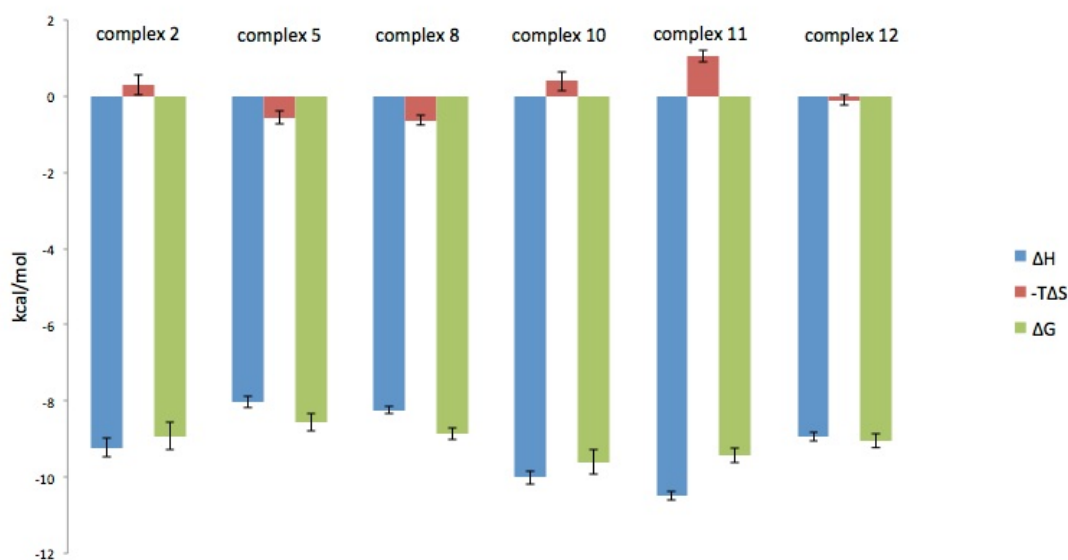
Fig. 7.6 Thermodynamic profile of the binding of **2**, **5**, **8**, **10**-**12** to human PHGDH based on ITC measurements

This entropic penalty is compensated by an increased enthalpic contribution: **2**, **10** and **11** have the highest binding enthalpies in the series. Enthalpy-entropy compensation [52]

resulted in these slightly different thermodynamic signatures giving essentially the same ΔG values (Figure 7.7).
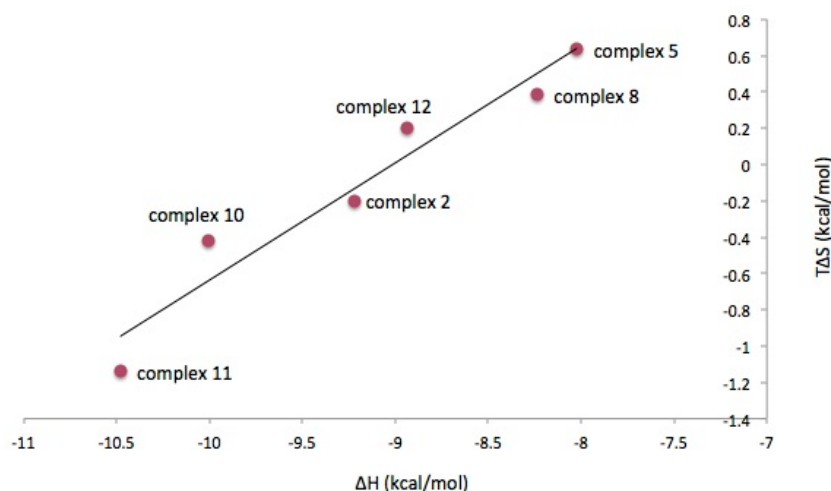
Fig. 7.7 Compensation behaviour between enthalpic (ΔH) and entropic (TΔS) components of free energy of binding.

Although the molecular origin of this small difference in entropic contribution has remained elusive, it is tempting to speculate about the possible causes. The crystal structure of **2** and models of **10** and **11** provide insights in this regard. These structures show the methoxy group projecting into solution (Figure 7.8). The first hydration shell around proteins is usually ordered, showing coherent hydrogen-bond patterns [214]. The methoxy group is pointing towards the solvent environment, which might have caused orientational reordering of the hydration shell water molecules, disrupting the hydrogen bonding network and resulting in the observed entropic penalty. The entropic penalty could also be incurred by restriction of rotational and translational degrees of freedom of the methoxy group upon binding [215].
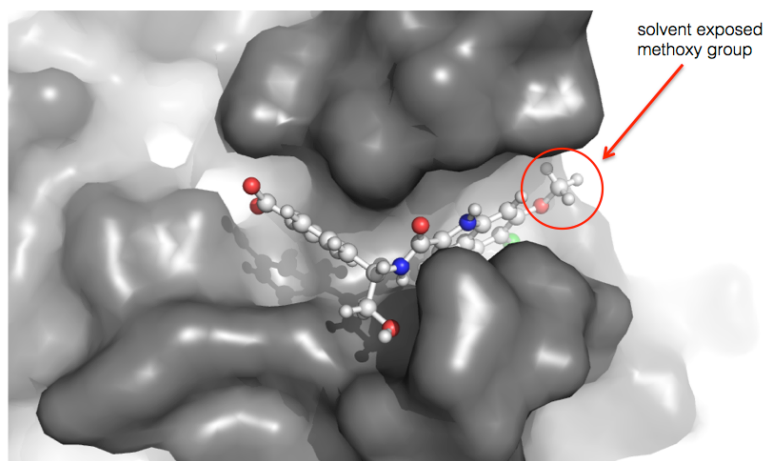
Fig. 7.8 Structure of **2** in complex with PHGDH. The methoxy group is pointing towards the solvent environment, which might have enhanced the water hydrogen bonding network and resulted in an entropic penalty.

Even though thirty inhibitors were analysed using ITC and following the same protocol, data could only be collected for six structures. These six structures share a carboxylic group as a common motif, carrying negative charge at physiological pH.

If the observed enthalpies were caused by the proton exchange, it would be reasonable to expect to see the difference in the observed enthalpies in different buffers, given the ionisation enthalpies of HEPES and TRIS buffers. However, there is no significant difference between these values, so buffer effects are considered to be unimportant. In the case of compound **1** TRIS buffer titration (Figure 7.9, left) and compound **3** HEPES buffer titration (Figure 7.10, right), the integrated binding isotherms do not seem to have sufficient curvature as required to confidently determine thermodynamic parameters.

The critical parameter which determines the shape of the binding isotherm is the unitless constant c [216]:

$$c = \frac{M_{tot}}{K_d}$$

or, more generally, for receptors with several identical noninteracting binding sites,

$$c = n K_a M_{tot}$$

where M_{tot} is the total macromolecule concentration in the cell at the start of the experiment, and n is the number of binding sites per receptor M.

c values of 10-500 [216, 217] or even 20-100 [218] are most often recommended. However, easily modelled curves can also be obtained using a low c-value protocol [219]. Even though c values for this experimental setup were within the recommended range
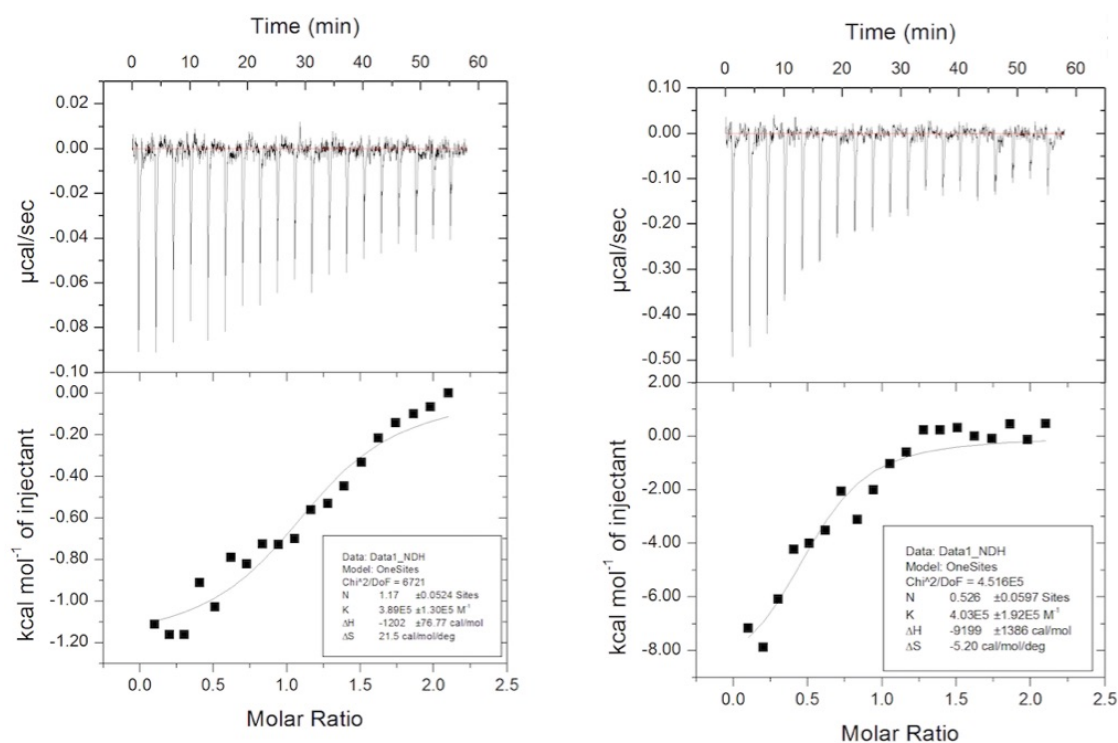
Fig. 7.9 Representative ITC titrations for compound **1** in TRIS and phosphate buffer.

(M_{tot}=25 μM and K_d < 1 μM), experiments at higher c values (M_{tot}=250 μM) were performed, but no meaningful data could be extracted. Between every measurement, cell and syringe were cleaned thoroughly.

Some of the compounds (such as compound **4**, Figure 7.11) showed no binding in the ITC experiment. When this was the case, the experiment was repeated to ensure no error in sample handling occurred, however, binding could still not be observed. Given that (at least in the case of compound **4**) the crystal structure of the protein-ligand complex is available, the fact that the binding was confirmed in several independent SPR experiments, and that the IC_{50} values were available, it could be hypothesised that the binding is not observed due to low protein concentration in the mixture. However, due to limited amounts of sample, no additional attempts at measuring thermodynamic properties of these compounds were made.

In all the experiments, the measured stoichiometry, *n*, was lower than what would be expected for a 1:1 binding mode, and varies throughout the experiments from 0.3 to 1. To obtain accurate values for stoichiometry, binding affinity, and normalised heats or the enthalpy of binding, the concentration of each species must be known. Error in the concentration of macromolecule in the cell will directly affect the stoichiometry while error in the concentration of ligand will affect all three fitted parameters of the binding experiment. Ligand samples were prepared by diluting a stock solution of known
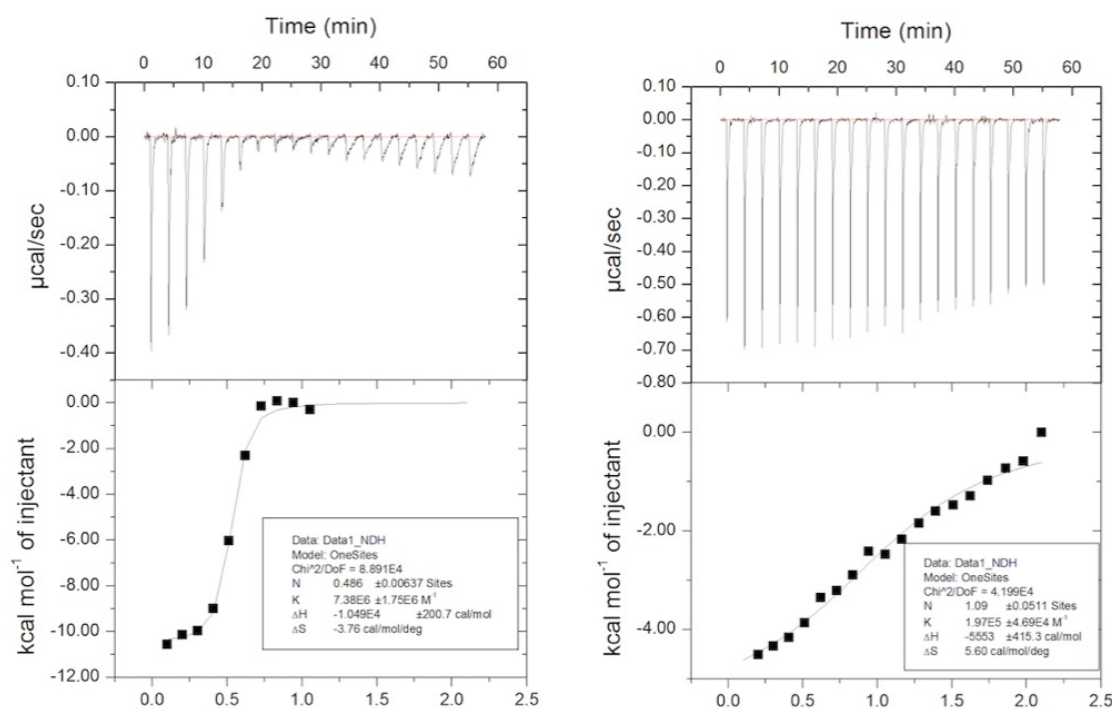
Fig. 7.10 Representative ITC titrations for compound **3** in TRIS and HEPES buffer.

concentration. Some of the ligands had poor solubility, which might have affected n values (although, this would be expected to cause higher n values, rather than lower). A more likely explanation is that low n values are linked to the fraction of active protein in the sample. The experiments were performed with different batches of protein and there were some concerns about the state of the protein in older batches. Another possible explanation is that the proposed 1:1 binding mode is not correct. This is unlikely given that the value of n varies so much across the titrations (if the binding mode was indeed 0.5:1, the value would not be expected to deviate more than 20%). No attempts were made at adjusting the concentrations before fitting the isotherm to the data.
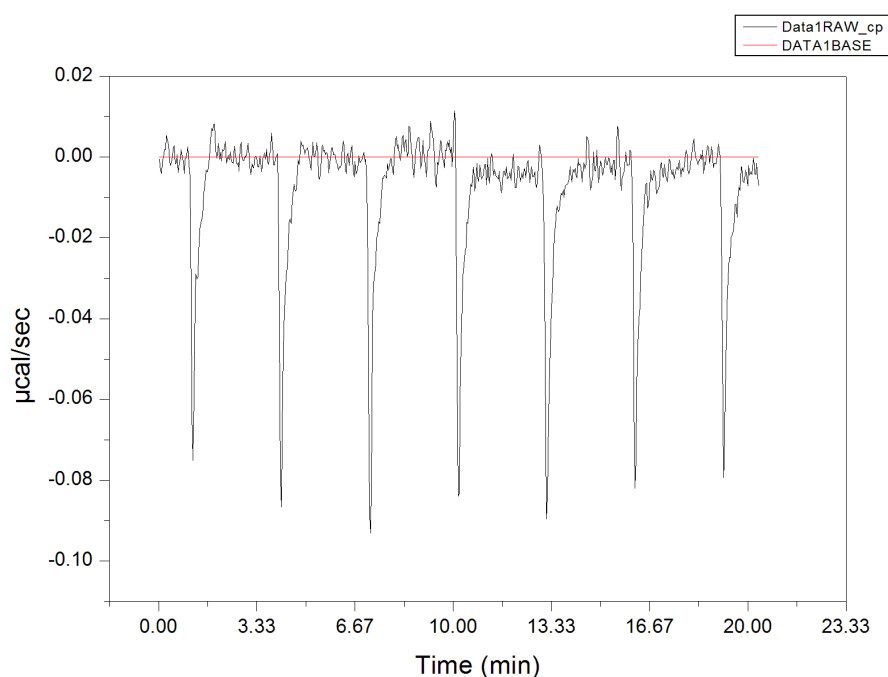
Fig. 7.11 Representative ITC titrations for compound **4** HEPES buffer shows no ligand binding.

7.4.2   Kinetic analysis

7.4.2.1   Experimental setup

Surface plasmon resonance (SPR) experiments were performed using a Biacore T2000 (GE Healthcare) equipped with a CM7 sensor chip. Protein was immobilised using amine-coupling chemistry. All the experiments were carried out using 50 mM potassium phosphate, pH 7.5, 150 mM NaCl, 0.005% T20, 1 mM TCEP, 1% DMSO. The surfaces of flow cells 2 and 3 were first activated with a 7 min injection of a 1:1 ratio of 0.2 M EDC and 0.5 M NHS at a flow rate of 5 μl/min.

Protein at a concentration of 10 and 20 μg/ml in 10 mM HEPES, pH 7.5 was immobilised at a density of 8000 RU on flow cells 2 and 3, and flow cell 1 was left blank to serve as a reference surface. Any remaining activated groups were blocked with a 7 min injection of a 0.1 M TRIS, pH 8.0. To collect kinetic binding data, multi cycle kinetic analysis was carried out at a constant flow rate of 30 μL/min and at a temperature of 25 °C. Compounds were diluted in the running buffer to yield 9 point concentration responses with a top concentration of 20 μM and 3 fold dilutions. Control (NADH 20μM prepared as test compound) and buffer injections were interspersed throughout the run every 25 cycles respectively.

Individual concentrations were injected from lowest to highest, with 90 s association and 300 s dissociation time.

K_d is determined from a Langmuir (1:1) binding isotherm. There is a very rapid interaction between the inhibitors and the protein, and it was impossible to extract k_{on} values confidently from the sensorgram; therefore only k_{off} rate constants were determined, by fitting the following equation:

$$R = R_0 \times e^{(-k_d(t-t_0)+offset)}$$

k_{on} was calculated from the $K_d = k_{off}/k_{on}$ relationship.

7.4.2.2   Results and discussion

Binding kinetic data was obtained using surface plasmon resonance (SPR) and demonstrated binding affinities comparable to those measured by ITC (Figure 7.4 and Table 7.2). Because of a rapid association and dissociation, it is necessary to interpret the individual data with caution. However, it is clear that the strongest binders ($K_d < 1$ µM) have the fastest observed k_{on} rates, while the k_{off} rates are essentially the same for all the inhibitors (Table 7.4).

Table 7.4 Kinetic data for inhibitors binding to PHGDH[a,b]

	pK_d	$\log(k_{on})$ (M^{-1} s^{-1})	$\log(k_{off})$ (s^{-1})
1	4.54	2.97	-1.58
2	6.07	4.65	-1.43
3	6.61	5.18	-1.44
4	5.12	3.51	-1.62
5	6.01	4.71	-1.30
6	5.27	NC	NC
7	5.33	3.82	-1.52
8	6.3	5.1	-1.21
9	NC	NC	NC
10	6.5	5.06	-1.44
11	6.3	5.12	-1.27
12	6.5	5.43	-1.09

[a] SPR experiments were performed at 298.15 K.
[b] Equilibrium dissociation constants (k_{on}/k_{off}). Data represent geometric mean ± SE from at least three independent experiments.

Kinetic analysis shows that, despite essentially the same dissociation rates, association rates vary, and cause slight differences in the overall binding affinity (Figure 7.12).
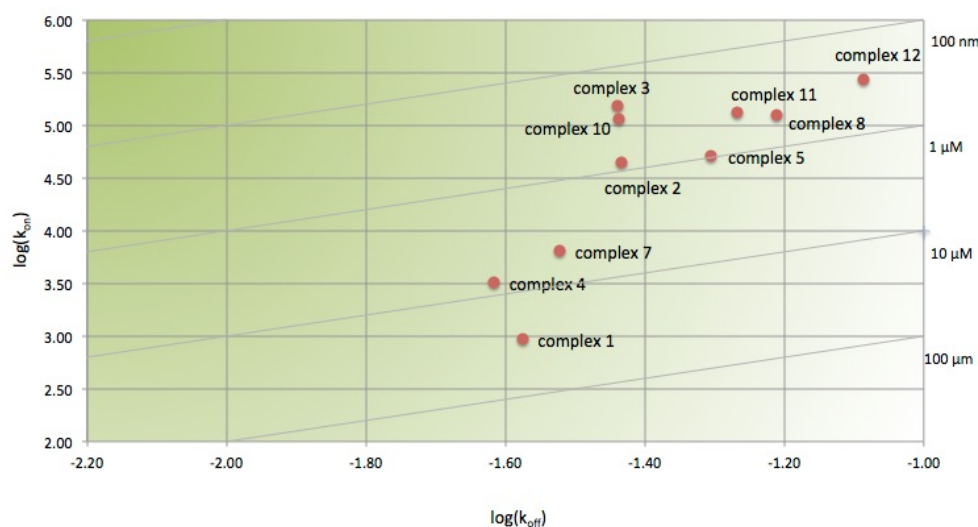
Fig. 7.12 Association rate constant k_{on} plotted against dissociation rate constant k_{off} based on SPR measurements. Diagonal lines in the plot represent equilibrium dissociation constant (K_d) values.

The association event is often expected to be rate-limited by diffusion [220]. If there was no underlying rate-limiting mechanism other than diffusion that is affecting binding, smaller compounds would be expected to bind quicker. A simple approach allowing an estimate of diffusion coefficient from the molecular weight of the solute and the identity of the solvent has been reported [221]. The calculator that implements this calculation [222] has been used to obtain the estimates of the diffusion coefficients of the inhibitors **1-12** (Table 7.5). Estimated diffusion coefficients plotted against association rates show no correlation between the two (R^2=0.26, Figure 7.13), or, in other words, that there are additional mechanisms affecting binding.

It can be noted that inhibitors with an acid moiety (**2**, **3**, **5**, **8**, **10-12**) are the quicker binders, despite the fact that they are not the smallest inhibitors tested. This is linked to the shape and nature of the binding site: a very shallow binding cleft and positive surface potential promotes faster association of negatively charged inhibitors that overrides the diffusion effects (Figure 7.14) [223]. Looking at the electrostatic surface of the protein, it is reasonable to assume faster complex recognition is governed by the charge complementarity and long range electrostatic interactions [224].

Even though SPR experiments were performed for around thirty compounds, it was possible to collect data for only twelve of them. When data could not be collected, it was either because no binding could be observed, observed signal was negative, or because of the poor fit of the curve to the sensorgram. In an attempt to overcome these issues, different immobilisation densities and sensor chips were tried, varying flow rates, as well as different compound concentrations, yielding no improvements.

Table 7.5 Molecular weight and diffusion coefficient[*] of the inhibitors

	$\log(k_{on})$ $(M^{-1}\,s^{-1})$	molecular weight	diffusion coefficient x 10^{-10}
1	2.97	344	8.19
2	4.65	388	7.76
3	5.18	457	7.09
4	3.51	344	8.19
5	4.71	352	8.1
7	3.82	328	8.36
8	5.1	358	8.04
10	5.06	368	7.94
11	5.12	372	7.9
12	5.43	356	8.06

[*] Diffusion coefficient was estimated given the molecular weight and the viscosity of methanol [221, 222]
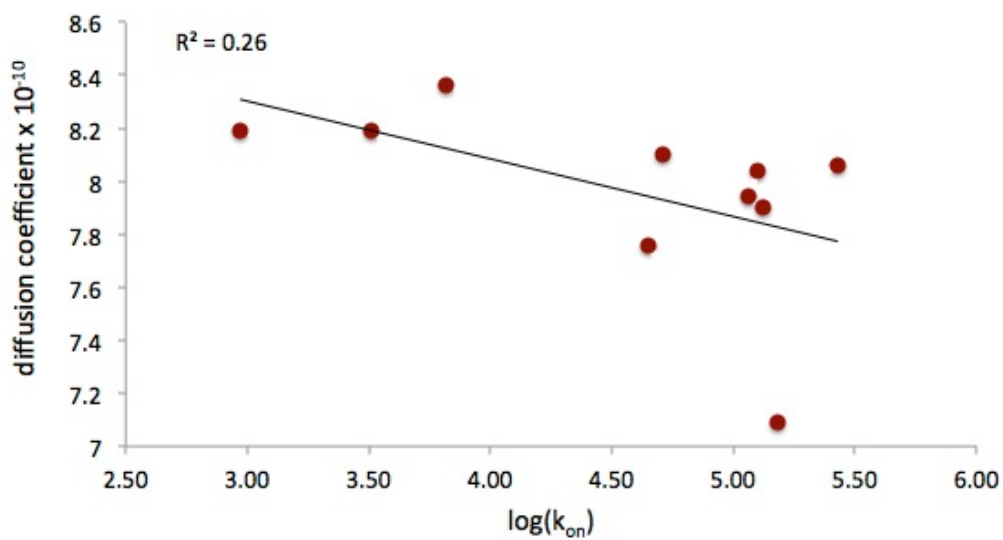
Fig. 7.13 Diffusion coefficients plotted against $\log(k_{on})$ values.
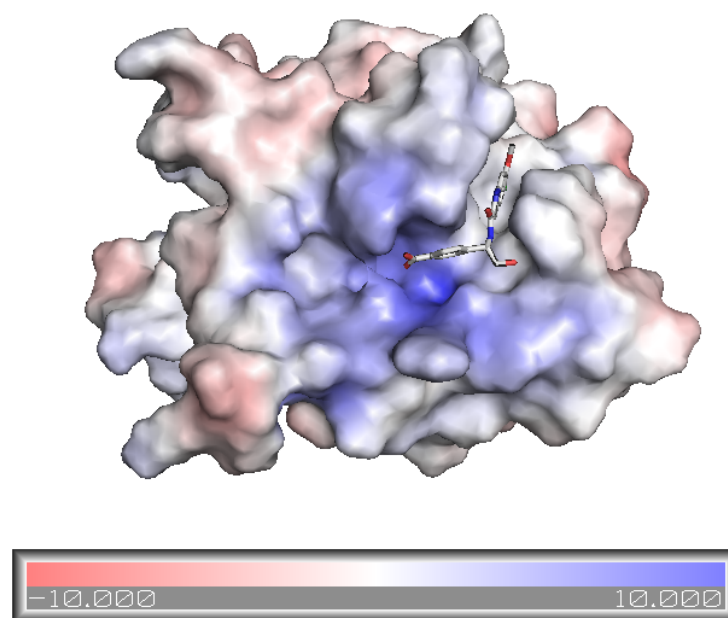
Fig. 7.14 Electrostatic surface of PHGDH. Inhibitor **2** is shown in sticks. Surface is coloured by calculated charge from red (–10 k_bT/e_c) to blue (+10 k_bT/e_c) using a dielectric constant of 78. The surfaces were generated using APBS plugin for PyMol [225].

7.5 Conclusions

Work detailed in this chapter describes the efforts in characterising thermodynamic and kinetic signatures of some of the compounds identified in the joint project to investigate PHGDH between AstraZeneca and the Cancer Research Trust. Although 30 inhibitors for this protein target were tested experimentally using SPR and ITC, the data were collected only for a limited number of structures. Both ITC and SPR are often considered to be "gold standard" methods for the determination of affinities as both measure the direct interaction of a compound with the target protein without any modification. This chapter shows that great care has to be taken during the establishment of conditions for the experiment, as these can dramatically affect the observed results and lead to an erroneous SAR interpretation. Even though binding affinities could be confidently determined over a large number of experiments, thermodynamic and kinetic fingerprints had to be carefully interpreted by combining available crystal structures, molecular modelling and chemical intuition.

Chapter 8

Mycobacterium tuberculosis Enoyl-ACP reductase (InhA)

8.1 Introduction

Tuberculosis is a major global cause of morbidity and mortality, due to the infectious pathogen *Mycobacterium tuberculosis* (Mtb). After dramatic outbreaks of multidrug-resistant tuberculosis in the early 1990s, resistance became recognised as a global problem [226].

Fatty acids are essential for bacterial growth. They cannot be scavenged from the host and must be synthesised *de novo* [227]. Enoyl-ACP reductases participate in fatty acid biosynthesis by utilising NADH to reduce the trans double bond between positions C2 and C3 of a fatty acyl chain linked to the acyl carrier protein (Figure 8.1). The enoyl-ACP reductase from *Mycobacterium tuberculosis*, known as InhA, is a member of an unusual FAS-II system that prefers longer chain fatty acyl substrates for the purpose of synthesising mycolic acids, a major component of mycobacterial cell walls [228]. This protein presents an attractive target for the development of novel drugs against tuberculosis, as it is essential in Mtb as well as sufficiently different from human enzymes [229].

The structures discussed in this chapter were selected from the AstraZeneca corporate database, based on their potency and clogP values, calculated by the in-house tool. The initial set contained 30 inhibitors with clogP values between 0 and 5 and pIC_{50} values up to 9.5.

Work detailed in this chapter describes the efforts in characterising thermodynamic and kinetic signatures of some of these compounds.
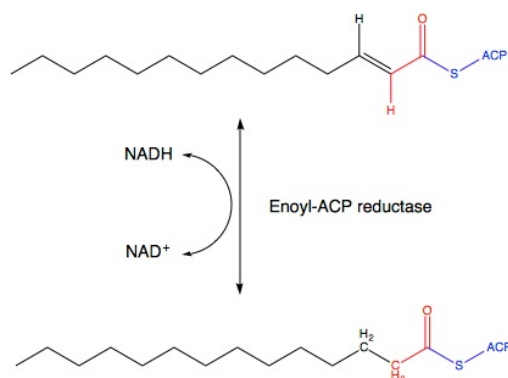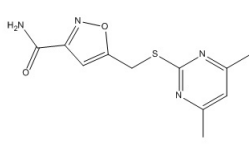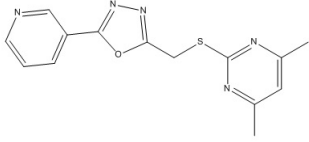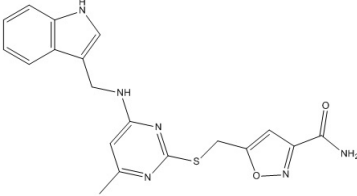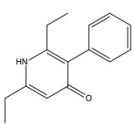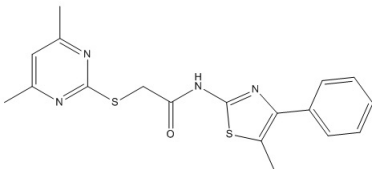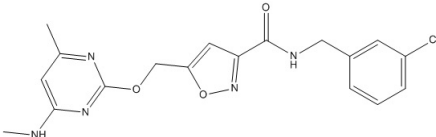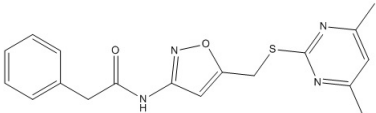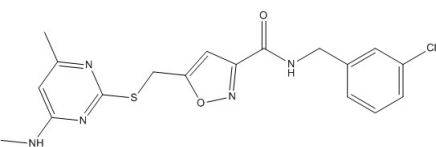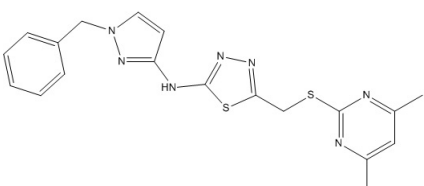
Fig. 8.1 Reaction catalysed by Enoyl-ACP reductase

8.2 InhA inhibitor selection and potency

Even though thirty inhibitors were analysed using ITC following the same protocol, data could only be collected for a limited number of structures. Data for six fragments binding to Mtb Enoyl-ACP reductase was collected using ITC, and eight fragments using SPR. Inhibitory potency measured in a Fluorescence based assay as retrieved from the AstraZeneca database and binding affinity measured by SPR and ITC of all the inhibitors is summarised in table 8.1.

Table 8.1 Inhibitory potency and binding affinity of InhA inhibitors. pIC_{50} values were retrieved from the AstraZeneca database.

		pIC_{50}	pK_d (SPR)	pK_d (ITC)
1		6.44	NC	6.28
2		7.57	NC	5.62
3		7.20	NC	6.29

4		5.54	NC	5.55
5		5.28	4.78	5.56
6		7.40	5.79	5.91
7		5.94	6.43	NC
8		6.73	6.06	NC
9		8.15	6.68	NC
10		7.07	5.96	NC
11		8.48	5.44	NC
12		8.52	6.94	NC

* NC=not calculated

All compounds belong to the di-methyl pyrimidine fragment series. Crystal structures for the complexes **1**, **3**, **4**, **6**, **8**, **11**, **12** were available in the AstraZeneca database. The

binding mode is conserved throughout the series: a H-bond donor-acceptor pair is formed between an isoxazole ring nitrogen (oxadiazole **5**, thiadiazole in **12**) and the backbone amide in Met98, and the amide or amine of the inhibitor and the backbone carbonyl in Met98 (Figure 8.2). Compound **8** has a thiazole ring with nitrogen pointing away from Met98 and does not form this interaction. The pyrimidine group engages with the nicotinamide and ribose groups of the NAD[1] cofactor (Figure 8.2). These specific interactions with the cofactor are in common with known substrate competitive InhA inhibitors [230]. The active site loop of InhA (residues 198-206), a helix which may be disordered or ordered in crystal structures of InhA dependent on interactions with ligand, makes nonpolar contacts with the ligand.
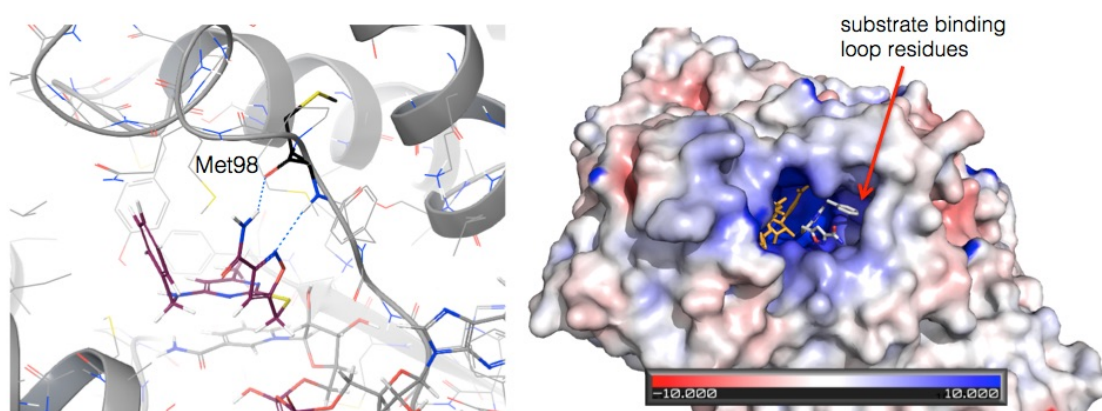
Fig. 8.2 Compound **1** binding mode. All the compounds (apart from **8**) form hydrogen bonds with Met98. Electrostatic surface of Inha. Inhibitor **1** is shown in white sticks, and cofactor in yellow. Surface is coloured by calculated charge from red (-10 k_bT/e_c) to blue ($+10$ k_bT/e_c) using a dielectric constant of 78. The surfaces were generated using APBS plugin for PyMol [225].

8.2.1 Kinetic analysis

Surface plasmon resonance (SPR) experiments were performed using a Biacore T2000 (GE Healthcare) equipped with a CM7 sensor chip. Untagged InhA was immobilised at a density of 3000 RU on flow cells 2 and 3, and flow cell 1 was left blank to serve as a reference surface. Protein was immobilised using amine-coupling chemistry, 50 mM HEPES buffer, pH 7.5, 150 mM NaCl, 0.005% T20 (v/v), 1 mM TCEP was used as immobilization buffer. All the experiments were carried out using 50 mM HEPES buffer, pH 7.5, 150 mM NaCl, 0.005% T20, 1 mM TCEP, 1% DMSO and 100μM NADH

---

[1]resolution is not sufficient to determine whether the cofactor is present in a reduced or oxidized state. NAD is used to denote cofactor when either the actual redox state is not confirmed or the statement refers equally to NAD$^+$ or NADH. However, based on the experimental settings [230], it can be assumed that the cofactor is in the reduced (NADH) form

as running buffer. The surfaces of flow cells 2 and 3 were first activated with a 10 min injection of a 1:1 ratio of 0.2 M EDC and 0.5 M NHS. Remaining activated groups were blocked with an injection of 0.1 M Tris, pH 8.0.

To collect kinetic binding data, multi-cycle kinetic analysis was carried out at a constant flow rate of 30 μL/min and at a temperature of 25 °C. Compounds were diluted in DMSO to yield 8 point concentration responses with a maximum concentration of 2 μM and 2 fold dilutions. Control (compound **7**) and buffer injections were interspersed with injections of compound to allow monitoring of the functionality of the immobilised protein surface. Individual concentrations were injected from lowest to highest, with 60 s association and 60 s dissociation time. K_d is determined from Langmuir (1:1) binding isotherms.

It was not possible to accurately collect kinetic data for this protein system. Many compounds did not show binding at all, binding only at high concentrations, signal below zero or higher than the theoretical binding maximum, or the signal was too weak for the model to be confidently fitted. Different buffer compositions were tested (such as 20 mM potassium phosphate, 150 mM NaCl, 1 mM DTT and 0.00 5%T20, 1% DMSO, 100μM NADH), compound concentrations (500, 250, 125 , 62.5, 31.25, 15.6, 7.8, 3.9 μM) as well as different flow rates, but no improvement was observed. Some compounds have previously been observed to give refractive index problems or to aggregate, so when the signal was considered unreliable these measurements were discarded. Both multi-cycle and single-cycle kinetics approaches were tested, and, while multi-cycle approach provided better results, only K_d values were determined. Kinetic rate constants could not be determined. This could be for several reasons: compounds may interact quickly with the protein (large k_{on} rate constants) and/or dissociate quickly (large k_{off} rate constants), such that the changes are outside the detection range of the instrument. There were also ambiguities about the state of the protein: an older batch had to be used which might have caused poor performance. All the compounds bind in the presence of the NADH, and, although care was taken to minimise the possibility of the cofactor degrading, it can not be excluded that the cofactor stability might have affected the performance.

Fragments **9** and **12** were the strongest binders in the SPR experiment (pK_d values of 6.68 and 6.94). However, compound **11** had pK_d of 5.44, significantly lower than what would be expected given the high pIC_{50} value of 8.52.

8.2.2 Thermodynamic analysis

The ITC experimental setup was the same as it was for PHGDH. Briefly, ITC measurements were performed at 25 °C in a) 50 mM potassium phosphate, 150 mM KCl, 1 mM TCEP and 0.00 5%T20, 1% DMSO, 100μM NADH b) 20 mM TRIS, 150 mM NaCl, 1 mM TCEP, 0.005% T20, 1% DMSO, 100μM NADH c) 50 mM HEPES, 150 mM NaCl, 1 mM

TCEP, 0.005% T20, 1% DMSO 100μM NADH all pH 7.5. The titrations were performed on 25 μM INHA in the 200 μL sample cell using 2 μL injections of 0.2 mM ligand solution every 180 s.

Even thought the titration experiments were performed in three buffers of different ionisation enthalpies, it was possible to collect data only for some of the compounds: for four of the compounds the data was collected in all three buffers, and for two compounds this was achieved only in TRIS buffer, buffer of the highest ionisation enthalpy. Proton transfer could not be confidently determined either: despite the structural similarity between the compounds, the slope n indicating the number of exchanged protons and the direction is negative (-0.44 net deprotonation of groups in the protein and/or the ligand) in compound **3**, and positive (0.66 net protonation of groups in the protein and/or the ligand) in the case of compound **1** (Figure 8.3), while compound **4** shows no such effects. Compound **2** shows the biggest difference in binding enthalpy (-14.66 kcal/mol in phosphate, -6.07 kcal/mol in TRIS and -0.66 in HEPES), and the direction of possible proton exchange can not be determined. Therefore, it is not possible to draw any conclusions about the thermodynamic signatures of these compounds.
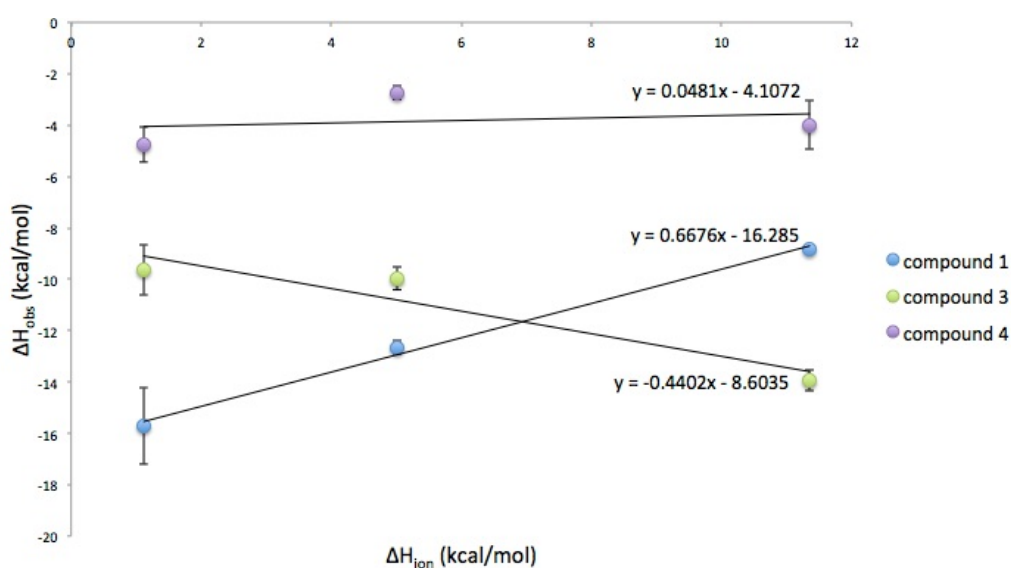
Fig. 8.3 ΔH_{obs} plotted against ΔH_{ion} for compounds **1**, **3** and **4**.

However, K_d values in different buffers were similar, so where possible, the geometric mean across all the K_d values was calculated. The measured data with representative titration curves is shown in Figures 8.4-8.9.

The measured stoichiometry, n varied between 0.5 and 1.5 throughout titrations. Ligand samples were prepared by diluting a stock solution of known concentration. In titration of compound **6**, stoichiometry was determined to be 1.5. Higher n values can be linked to lower than expected ligand concentrations. Given that the clogP of this compound is

2.4, no solubility issues would be expected. A more likely explanation is that the observed discrepancy in n values is linked to the concentration of active protein in the sample. The experiments were performed with different batches of protein and there were some concerns about the state of the protein in older batches. Additional protein binding sites or alternative binding models are also unlikely given that the value of n varies so much across the titrations in both directions, and no additional binding modes were observed in the available crystal structures. No attempts were made at adjusting the concentrations before fitting the isotherm to the data.

Table 8.2 InhA thermodynamic data for compound **1**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.79 ± 0.00	5.61 ± 0.76	-12.65 ± 0.13	-3.48 ± 0.15	0.17 ± 0.02
phosphate	0.60 ± 0.00	0.73 ± 0.17	-15.72 ± 0.75	-7.75 ± 0.76	1.36 ± 0.31
TRIS	1.19 ± 0.00	4.15 ± 0.54	-8.81 ± 0.08	0.17 ± 0.11	0.24± 0.03
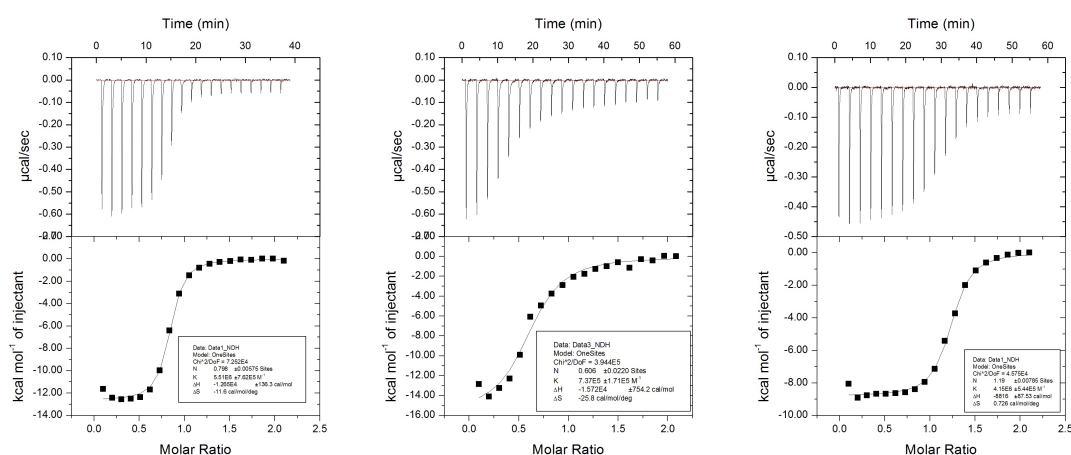
Fig. 8.4 Representative ITC titrations for compound **1** in HEPES, phosphate and TRIS buffer.

In cases where it was not possible to obtain thermodynamic parameters, it was either because no binding could be detected (surprising as crystal structures of some of these complexes were available), or the experiments did not yield a sigmoidal binding isotherm (Figure 8.10).

Table 8.3 InhA thermodynamic data for compound **2**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	1.27 ± 0.08	0.48 ± 0.26	-0.66 ± 0.25	7.05 ± 0.33	2.07 ± 1.14
phosphate	0.59 ± 0.05	0.28 ± 0.09	-14.64 ± 1.83	-7.23 ± 1.84	3.57 ± 1.22
TRIS	0.98 ± 0.02	0.65 ± 0.12	-6.07 ± 0.18	1.83 ± 0.21	1.51± 0.27
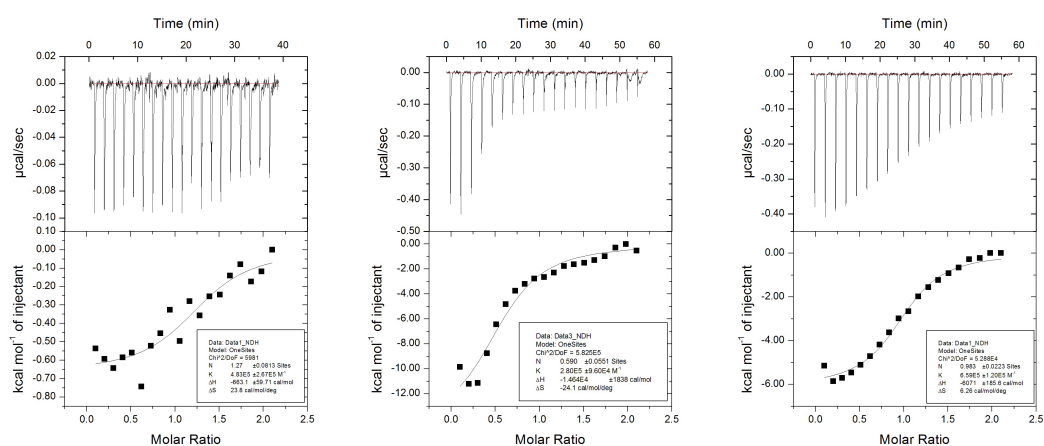
Fig. 8.5 Representative ITC titrations for compound **2** in HEPES, phosphate and TRIS buffer.

Table 8.4 InhA thermodynamic data for compound **3**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.61 ± 0.00	5.23 ± 1.39	-9.96 ± 0.23	-0.83 ± 0.27	0.19 ± 0.05
phosphate	0.75 ± 0.02	0.80 ± 0.23	-9.61 ± 0.50	-1.60 ± 0.53	1.25 ± 0.37
TRIS	0.51 ± 0.00	10.3 ± 2.17	-13.93 ± 0.21	4.35 ± 0.25	0.09 ± 0.02
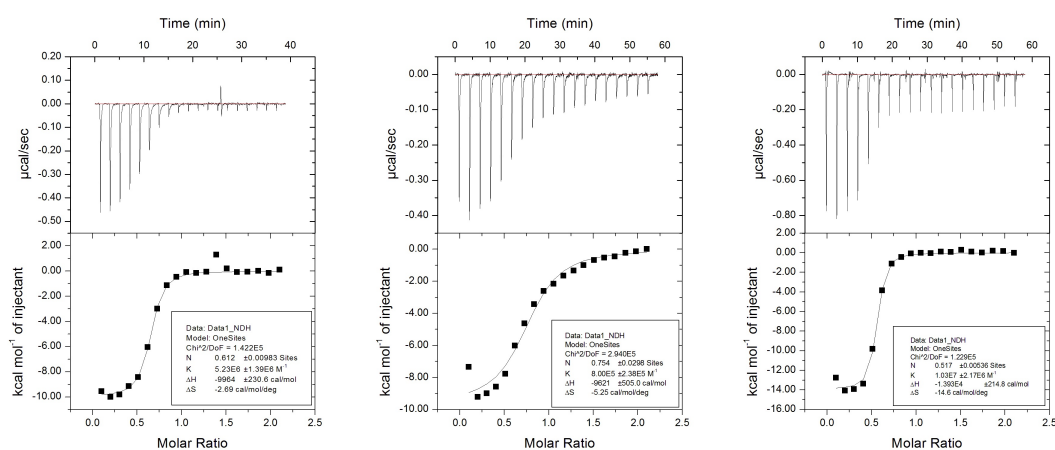
Fig. 8.6 Representative ITC titration for compound **3** in HEPES, phosphate and TRIS buffer.

Table 8.5 InhA thermodynamic data for compound **4**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	1.21 ± 0.04	0.50 ± 0.15	-2.74 ± 0.13	4.99 ± 0.22	2.0 ± 0.06
phosphate	1.09 ± 0.05	0.81 ± 0.37	-4.75 ± 0.35	3.28 ± 0.44	1.22 ± 0.55
TRIS	1.02 ± 0.08	0.18 ± 0.7	-3.99 ± 0.48	3.17 ± 0.53	5.34 ± 2.08
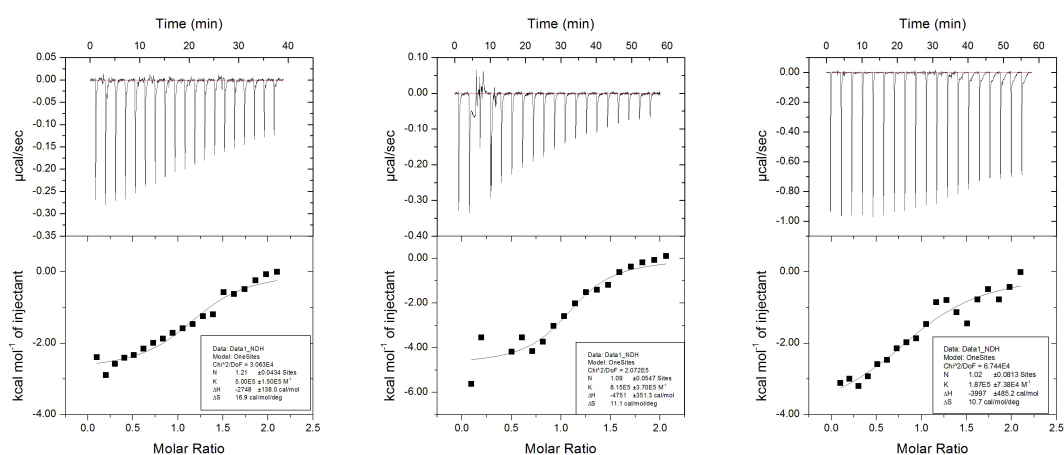
Fig. 8.7 Representative ITC titration for compound **4** in HEPES, phosphate and TRIS buffer.

Table 8.6 InhA thermodynamic data for compound **5**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
TRIS	0.86 ± 0.03	0.35 ± 0.28	-3.17 ± 0.18	4.37 ± 0.23	2.78 ± 0.67
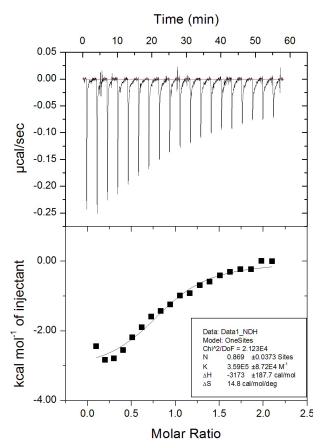
Fig. 8.8 Representative ITC titration for compound **5** in TRIS buffer.

Table 8.7 InhA thermodynamic data for compound **6**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
TRIS	1.49 ± 0.03	0.80 ± 0.24	-2.66 ± 0.09	5.36 ± 0.20	1.23 ± 0.37
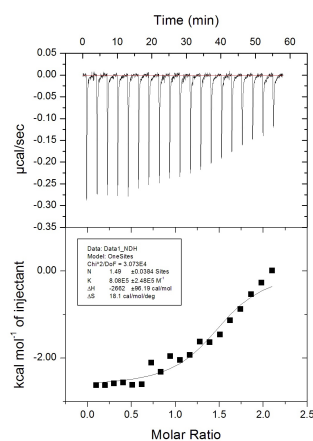
Fig. 8.9 Representative ITC titration for compound **6** in TRIS buffer.
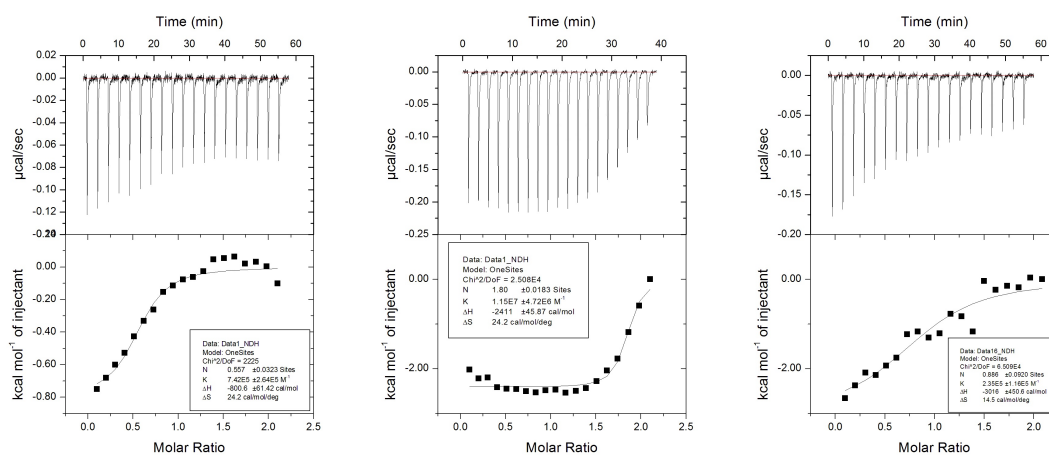
Fig. 8.10 Binding isotherms of some of the titrations that were considered unsuccessful.

8.3 Conclusions

Work detailed in this chapter describes the efforts in characterising thermodynamic and kinetic signatures of some of the compounds identified in AstraZeneca's Mtb Enoyl-ACP reductase project. Although 30 inhibitors for this protein target have been tested experimentally using SPR and ITC, the data were collected only for a limited number of structures. Like the previous chapter (Chapter 7, 3-Phosphoglycerate dehydrogenase (PHGDH)), this chapter highlights some of the difficulties encountered when using the two techniques.

Again, affinity values could confidently be determined across a number of experimental repeats, while thermodynamic and kinetic fingerprints could not be determined. Kinetic rates seem to be outside the range of what can confidently be determined by the instrument and thermodynamic signatures seem to be buffer dependent. However, without additional experiments, these signatures can not be interpreted conclusively.

Summary and future work

Proteins have numerous functions: as enzymes they accelerate vital chemical reactions, they act as regulatory molecules, signaling molecules, selective channels, and many more. To have any of these functions, proteins bind small molecules, nucleic acids or other partners by forming networks of noncovalent interactions between the two binding partners.

In hit selection or ligand optimisation, success is usually measured by evaluating changes in binding affinity. Affinity is usually determined by measuring a binding constant or IC_{50} value.

At equilibrium and under standard conditions, affinity can be understood as Gibbs free free energy, logarithmically related to the dissociation constant. Free energy can be factorised into enthalpy and entropy contributions and these two thermodynamic properties potentially provide further insights into the forces that drive formation of the complex.

Considering binding kinetics of protein–ligand complex formation, the binding constant can be understood as the quotient of the kinetically determined rate constants k_{off} and k_{on}.

Different combinations of enthalpic and entropic contributions or k_{off} and k_{on} can result in the same binding affinity. However, compounds with equal affinities are not equivalent because the interactions they make with the target give rise to different thermodynamic or kinetic signatures. How to exploit and modify the forces contributing to ligand binding is still not clear.

Structure-based drug design relies on knowledge of the three dimensional structure of the biological target to generate high-affinity small molecules that, when bound to the target, elicit the desired functional response. For any small molecule to be considered a likely drug candidate, it must satisfy a number of important properties such as bioavailability, lack of toxicity, lipophilicity, selectivity and stability. However, a small molecule must above all be active, which means that it must bind tightly and selectively to a specific location in the protein target.

Predicting the potency of a compound remains an unsolved problem. Many compounds are synthesised and their potency is tested, most of which turn out to be inactive against the desired target, failing thus at the very first hurdle.

The reliable prediction of affinity is difficult because it is not simply a summation of individual contributions. An interaction involving one part of a molecule perturbs

every other interaction that molecule makes. Furthermore, ligands and targets associate in aqueous media and significant contributions arise from the solvent. Most computational efforts in this area have been structurally "inclusive", employing extensive sampling of orientations and conformations of the full system. However, accuracy is compromised because a poor physical description of the protein and the ligand has to be used because these are computationally cheap and quick and hence amenable for the full system. Although many significant advances in understanding how to rank molecules for their biochemical activity have been made, a robust and general process does not yet exist.

An alternative structurally "exclusive" approach that includes only the parts of the protein that are involved in ligand recognition and neglects the remainder has been introduced in chapter 5. With the example of LDHA inhibitors it has been demonstrated that QM calculations can provide useful insights into interactions between the binding partners, and a deeper understanding of the nature of the interactions can be exploited to generate new ideas: the outcome of modelling different LDHA inhibitors is a new fragment series with high predicted affinity and the potential for fragment growth.

In chapter 6, a similar approach was applied to study if QM can be used to select preferred protein conformations and ligand positions where there is some movement of the protein backbone. The results demonstrate that the method is not only capable of detecting the preferred protein conformation (in all cases but one), but also of predicting the relative populations with reasonable accuracies as well.

It was shown in both chapters 5 and 6 how QM calculations can be used for determining conformational, tautomeric and stereochemical preferences and to help elucidate uncertainties caused by poor crystallographical models or, equally impressive, add an additional level of detail to highly accurate models (as was shown in the theoceptor guided refinement of lysozyme).

Chapters 7 and 8 detailed the efforts in characterising the thermodynamic and kinetic signatures of PHGDH and InhA inhibitors. Although more than 30 inhibitors of both protein systems were tested experimentally using SPR and ITC, it was possible to collect data only for a limited number of structures. Binding affinities could be confidently determined over a large number of experiments, but thermodynamic signatures varied with experimental conditions, while rate constants could be determined only for a limited number of PHGDH inhibitors. In both of these chapters the importance of evaluating the experimental setup and conditions at which the data were measured is discussed in greater detail.

Although the quality of the experimental results does not permit conclusive interpretation, they do provoke appreciation for how much measured parameters can deviate from ideal, error-free values. This is especially important for anyone attempting any compu-

tational modelling, as it highlights the importance of assessing accuracy, reliability and relevance of the experimental data, which is often overlooked by modellers.

Both kinetic and thermodynamic behaviour of a ligand binding to a protein can be viewed as a combination of interactions between the binding partners, paired with conformational effects caused by restricting degrees of freedom of the ligand upon binding and reordering the protein. A lot of work in this area is still retrospective in nature; optimising thermodynamics and kinetics as a part of medicinal chemistry programmes is still not routine, although such efforts are becoming more frequent. The individual contributions of intermolecular interactions like the formation of hydrogen bonds or ionic interactions are not direct experimental observables, which makes correct interpretation of the ITC data very difficult. Due to these complications, raw thermodynamic data as well as thermodynamically derived metrics like enthalpic efficiency should be used with great care. Although appealing, the potential for using thermodynamic or kinetic signatures as end points appears to be limited. It is more likely that the combination of computational approaches, crystallography, and experimental data can contribute to the understanding of the forces that govern protein-ligand binding. Comparison of experimental thermodynamic data with LLE, combined with the QM calculations shown in this work enables the identification of compounds that do not behave as expected. Identifying and analysing those outliers has a great potential of maximising the use of thermodynamic profiling. The ability to manipulate kinetic signatures is likely to improve with advances in the technologies used for studying binding events and the increase of available experimental data. Despite the increase in computer power and the design of specific hardware, accurate predictions of any of the kinetic determinants is still very difficult. Most of the efforts in this area are centred around molecular mechanics and the main reasons for limited success in computing binding kinetics seems to be linked to the known limitations of the method: inability to accurately and adequately sample protein and ligand conformations and reproduce association or dissociation pathways, solvent sampling and force field parametrisation. Still, this area seems to be provoking interest, and it is possible that we will soon begin to rationally design kinetically optimised drugs. QM approaches such as theoceptors could be used to post-process molecular dynamics snapshots, or to evaluate ligand poses from ligand binding/unbinding trajectories. Another application would be to develop and optimise faster methodologies, such as force-field parametrisation of bonded and nonbonded (e.g. partial atomic charges, Lennard-Jones parameters) terms for molecular mechanics, molecular dynamics, crystallographic refinement and docking calculations. It is likely that the use of QM approaches in all phases of *in-silico* structure based and ligand based drug design will experience growth in the coming years. The high accuracy with which QM describes the interactions between receptor and ligand and the

ability to generate novel descriptor classes should attract even more attention to the use of these methods.

Part III

Matched molecular pairs

Chapter 9

Quantifying the effect that chemical environment exerts upon changes in matched molecular pair analysis

9.1   Introduction

Since the beginning of rational drug design, compound designers have been facing the same question: "What to make next?" With the increased availability and applicability of high throughput techniques from around the mid 1990s, the amount of data available and the diversity of compounds for which data exists has increased incredibly. Despite this, an estimated 90% of compounds entering clinical trials fail to reach the market [231]. It is therefore imperative that drug discoverers find ways to use the vast repository of data already available to improve their chances of success in the future.

One of the ways to decide what to make next is to analyse all of the molecules that differ only by a well-defined structural transformation and use the effect of these structural changes in the past to estimate their likely effect if applied again. This is matched molecular pair analysis, MMPA. One of the hypotheses that underpin this approach is that the effect on pharmaceutical properties of a small structural change between molecules is more easily predicted than the absolute value of those properties for each molecule alone [232–235]. MMPA has been successfully applied in a range of ways: suggesting what to make next, predicting the properties of a new compound, identifying cases where a structural change has a minimal effect on key properties (like bioisosteres), or simply to increase our understanding of the links between biology and chemistry.

MMPA has recently been applied to the ADMET databases of Roche, AstraZeneca and Genentech and the output combined to create what could be the world's largest repository of medicinal chemistry knowledge, akin to a textbook. A by-product of this merging is that

a large set of MMPA data are available that can answer fundamental questions about the technique: i) how many pairs are needed in order to be confident that a particular change in structure actually causes an increase (or decrease) in property? ii) for small sets of molecules, how large must the change in property be in order to be reasonably confident that a particular change in structure actually causes an increase (or decrease) in property? iii) is there an upper limit to the number of pairs needed? This chapter details the initial findings from this work that will help others to maximise the value arising from measuring physical and biological properties of compounds.

9.2 Context definition

The analysis used WizePairZ [236], which identifies matched molecular pairs in an unsupervised fashion and encodes the chemical environment (the context) adjacent to any structural changes. The simplest environment is the first atom at the attachment point of the changing structure (context level 1), and the largest environment encompasses up to 4 atoms (level 4) from the attachment point (Figure 9.1). Others have shown that the chemical environment can be used to split groups of pairs into subsets that reveal interesting details about the link between structures and properties [237]. This work aims to investigate whether better chemical definition reduces the number of data required to achieve a given level of accuracy.
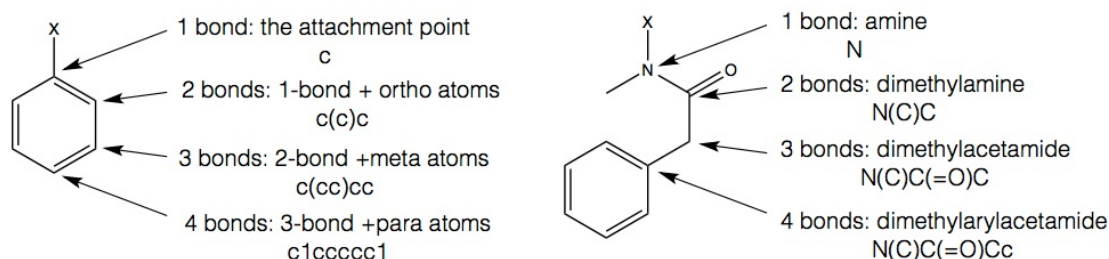
Fig. 9.1 The definition of chemical context for a matched molecular pair

Initial investigations have focused on one endpoint: inhibition of the hERG ion channel, which is linked to a cardiac toxicity risk [238]. hERG is a typical dataset and is used to illustrate the general case. Application of the WizePairZ algorithm found more than 1.1 million pairs of measured values representing more than 700k unique changes in structure.

9.3 MMPs as coin flip experiment

A positive mean change for a set of matched pairs has previously been found for a set that, when expanded, actually has a negative mean change [233]. Hence, a strict approach to selecting pairs was adopted, in which each pair is viewed as a coin flip experiment where the property increasing corresponds to "heads" and the property decreasing corresponds to "tails". This reduces the challenge to detecting biased coins, a well-understood problem [239].

When matched pairs are considered as coin flips, knowing the probability of any outcome due to random chance, the number of actual successes (coin landing heads) can be evaluated for signs of bias. If the coin consistently lands higher or lower than would be expected due to chance, then after a large number of coin flips, there will come a point where the chance probability of an observed outcome is sufficiently small that the observed results are evidence for bias. In the matched pairs analogy, a bias corresponds to a real effect caused by the structural change: a mechanism exists linking the structural change to a property change (Figure 9.2).
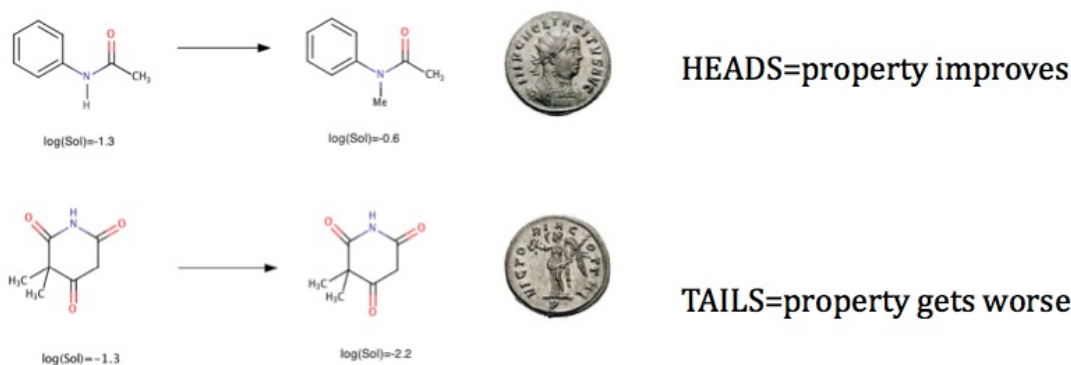
Fig. 9.2 Matched pairs viewed as coins

9.3.1 Detecting biased coins

When is it reasonable to suspect that an observed event is caused by some underlying mechanism? All possible outcomes for flipping a coin up to 60 times were examined (Figure 9.3). For each possible outcome the 95% confidence interval (CI) was calculated. If the CI contained 0.5, which would be a true mean for a fair, unbiased coin, the case was evaluated as 'true', or else, if 0.5 was outside the 95% CI, the case was evaluated as 'false'. The 95% CI defines a range of values that we can be 95% certain contains the population mean: if the interval contains 0.5, which is the mean for a fair coin, there is not evidence to suggest that the coin is biased.
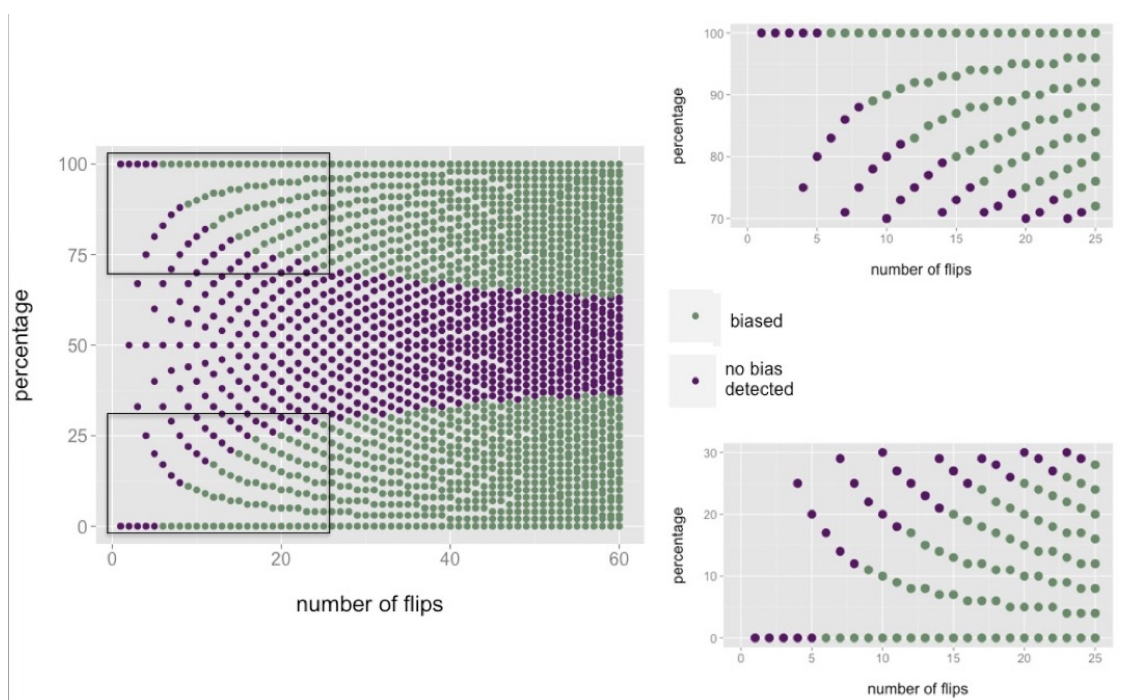
Fig. 9.3 Evaluation of all possible outcomes for flipping a coin up to 60 times. The inset highlights the case of 10 coin flips.

Considering the case of 10 coin flips, heads can appear 0 to 10 times. There are 4 cases when the 0.5 value is outside the 95% CI: 0 or 1 head (equivalent to 10 or 9 tails) or 9 or 10 heads (1 or 0 tails). It is only in these 4 cases that we can be 95% confident 0.5 is outside the CI, which suggests that the coin is biased, or, in other words, that that structural change causes an effect (Table 9.1). If there are 10 matched pairs in a set, only if the measured property changes in the same direction for a least 9 cases it is possible to be 95% confident that this is caused by the structural change.

The first time that a biased coin can be identified with 95% confidence is for 6 coins all landing the same way. This means that in order to make confident assumptions about whether a structural change is indeed changing a property of interest, at least 6 MMPs are required and in all 6 pairs the value of the measured property has to change in the same direction.

Although the coin flip analogy reduces the likelihood of misassigning the direction of change, it analyses only the direction of change in a property and reduces the number of structural transformations that are identified as having a significant effect. Having access to a very large database of matched pairs permits the link between the coin flip approach and the mean changes to be probed. The encoding of the chemical environment permits analysis of its impact.

Table 9.1 All possible outcomes for flipping a coin 10 times, together with 95% CI boundaries.

Number of heads	Fraction of heads	Lower 95% confidence bound	Upper 95% confidence bound	Is the coin biased?
0	0.0	0	0.308	Biased
1	0.1	0.003	0.445	Biased
2	0.2	0.025	0.556	No bias detected
3	0.3	0.067	0.652	No bias detected
4	0.4	0.122	0.738	No bias detected
5	0.5	0.187	0.813	No bias detected
6	0.6	0.262	0.878	No bias detected
7	0.7	0.348	0.933	No bias detected
8	0.8	0.444	0.975	No bias detected
9	0.9	0.555	0.997	Biased
10	1.0	0.692	1	Biased

9.4 Analysis of the hERG dataset

The matched pairs from the hERG dataset were analysed using the coin flip approach. For sets with more than 5 members, the fraction of cases corresponding to increases in inhibition and the 95% confidence interval range for this fraction are computed. If this range does not include 50%, a biased coin has been detected. Those sets causing an increase or decrease in inhibition are styled INC or DEC respectively. The remaining sets were assigned to the class No Effect Determined (NED); this includes structural changes that cause no change in hERG inhibition and those that do change the property but for which there is not yet enough data. Having assigned each set with more than 5 pairs to one of three classes (INC, DEC, NED), the measured difference Δ for each pair of molecules was examined. The variation in the average of Δ, as pairs are added to each set, tracks how sets of matched pairs evolve. If the sequence is $\Delta_1, \Delta_2, \Delta_3 \ldots \Delta_N$, then the mean of Δ_1 and Δ_2 is μ_2, the mean of Δ_1, Δ_2 and Δ_3 is μ_3. The ordering of pairs in this database is arbitrary (in real datasets it is governed by the order in which compounds are made and tested). Hence, in each set the order of Δs was scrambled and μ_1 to μ_N recomputed. The scrambling is repeated 50 times and the average for each μ computed to give $\overline{\mu_2}, \overline{\mu_3} \ldots \overline{\mu_N}$. For each set of pairs, this tracks how the mean value evolves as more pairs are added. These are then averaged over all sets of pairs to give $\overline{\overline{\mu_2}} \; \overline{\overline{\mu_3}} \ldots \overline{\overline{\mu_N}}$. The evolution along this series corresponds to what happens (on average) for sets of pairs in each class (INC, DEC and NED) as more pairs are added. All analysis was performed in R [240–243].

9.4.1 Predicting the transformation outcome

Logistical regression analysis links μ with the likelihood of that set ultimately becoming INC, DEC or NED as more pairs are added. These probabilities are plotted against values of $\overline{\overline{\mu_2}}$ to $\overline{\overline{\mu_{20}}}$ in Figure 9.4. This shows that high values of μ_i are most likely to correspond to INC, low values to DEC and intermediate values to one of the three classes depending upon the exact value. In the following, we focus upon mean changes that are positive but symmetrical arguments concerning negative mean values would hold. There are three distinct zones highlighted:

- the 'Comfort Zone', where the probability of INC is more than 95%

- the 'Twilight Zone' where both INC and NED are equally probable

- the 'Danger Zone' in which there remains a likelihood above 1% that the set of pairs will ultimately be in the DEC class.
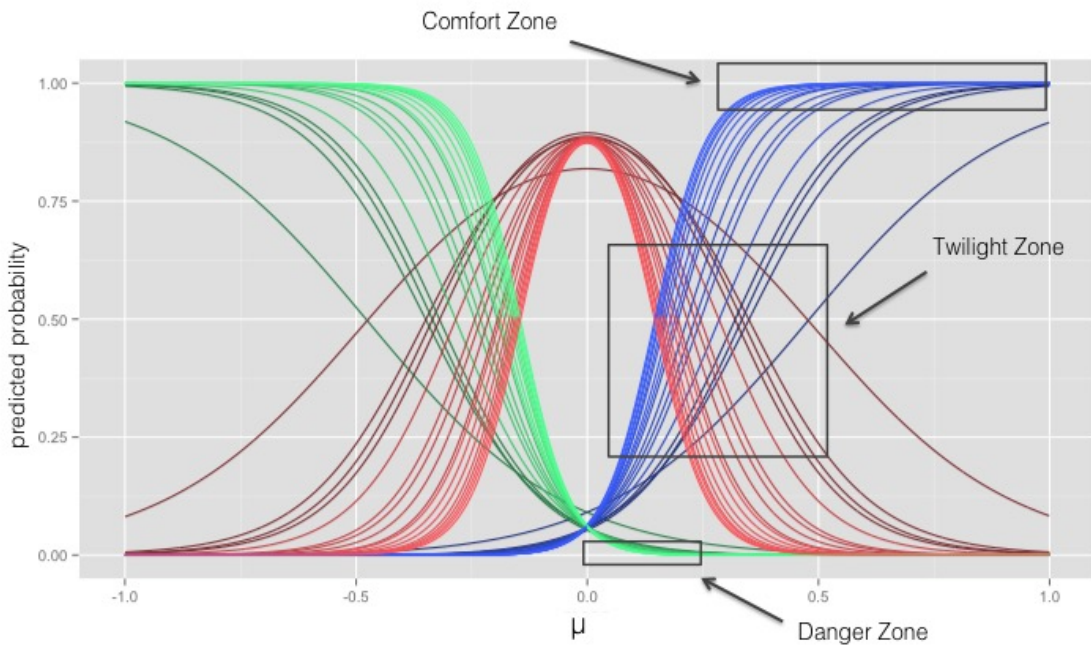

Fig. 9.4 The probability that a set of N pairs with mean value μ will evolve into a set of pairs classified as INC (blue), DEC (green) or NED (red). The colouring starts dark for N=2 and becomes progressively lighter as N increases to 20.

When the number of pairs is low, the comfort zone is only reached for sets of pairs in which the mean change is very high – it is effectively never reached when there are only two pairs in the set. As the number of pairs in the set increases, the comfort zone expands such that sets with lower mean values are increasingly likely to fall into it. However, this converges such that by the time there are 20 pairs, adding more pairs does not provide a marked benefit; a mean change above 0.31 for a set of 20 or more pairs indicates a structural change that is highly likely to increase the property. The twilight zone varies in a similar fashion: a large mean change is required for sets with few pairs before they become most likely to belong to the INC class. For sets with two pairs, the mean change in pIC_{50} has to be above 0.5 before INC is the most likely class for the set to belong to. Again there is convergence as the set size increases to 20 pairs, such that mean values of Δ above 0.15 correspond to INC being the most likely class. Finally, the danger zone extends out to 0.40 when there are two pairs in the set and contracts to 0.08 by the time there are 20 pairs.

9.4.2 The effect of chemical environment

Turning to the importance of chemistry, the sets can be subdivided according to the different levels of description of their chemical environment. When there are low numbers of pairs, chemical environment has a strong influence on the size of the three zones. For instance, with 5 pairs, the comfort zone starts at 0.83 for context level 1, 0.76 for level 2, 0.68 for level 3 and 0.63 for level 4. Meanwhile, the twilight zone is at 0.39, 0.36, 0.32 and 0.31 for context levels 1-4 respectively and the danger zone at 0.25, 0.21, 0.17 and 0.16 (Figure 9.5). Increasing the chemical specificity of small sets of pairs increases the likelihood that they provide correct guidance about the impact of a structural change. A set of 4 pairs with a 4-atom chemical environment specification is as likely to suggest the correct effect as a set of 7-8 pairs with a 1-atom chemical environment specification. This suggests that chemical specificity allows reliable knowledge to be extracted from smaller sets of pairs than would be possible with the 6 pair lower limit required by the strict biased coin detection approach. The power of this reduction is that there are over 12k groups of pairs with 4 in the set but only 2.1 k groups of 7 pairs and 1.6 k groups with 8 pairs. Indeed, there are only 12 k sets with 7 or more pairs in the set. As more pairs are added, however, context becomes less important, suggesting that there is a critical point after which 'statistics overrides chemistry', and more certain outcomes are achieved although they are for less chemically specific sets.

Medicinal chemists often have a value in mind above which changes measured in any given assay might be considered "real". This is typically about two- to three-fold (0.3 – 0.5 log units). Figure 9.6 shows that substantial numbers of pairs would be required before average changes of this size could be considered as providing a useful guide to the effect of
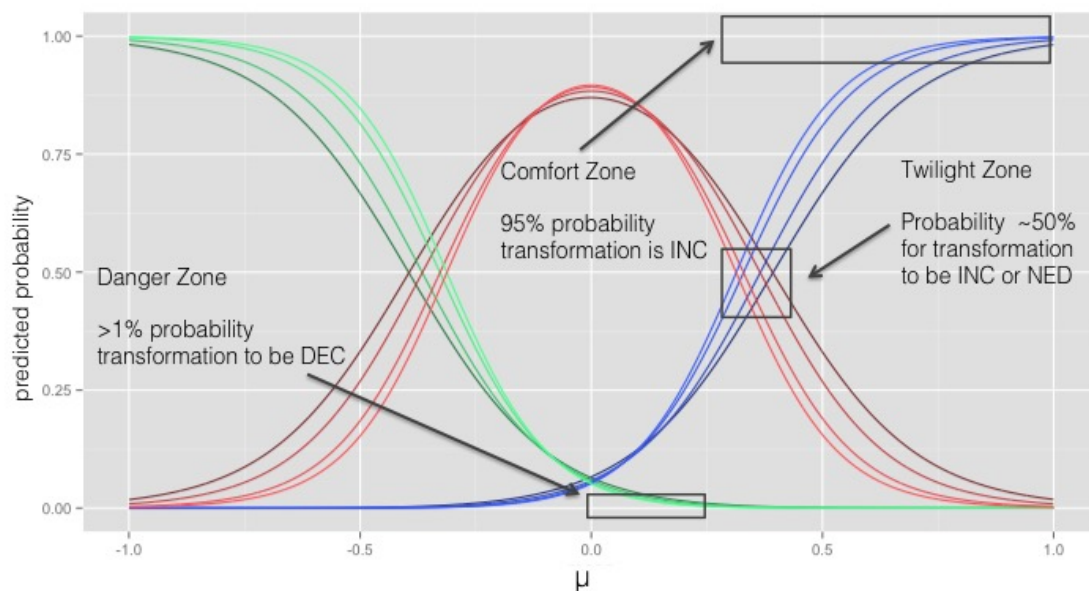
Fig. 9.5 The probability that a set of 5 pairs with mean value μ will evolve into a set of pairs classified as INC (blue), DEC (green) or NED (red). The colouring starts dark for context level 1 and becomes lighter as it increases to 4.

a structural change (9 pairs on average for context level 1 and 6 pairs for context level 4). The requirements for an assay to correctly rank order compounds are different to providing a clear guidance about the effect of changes in structure in a predictive fashion.
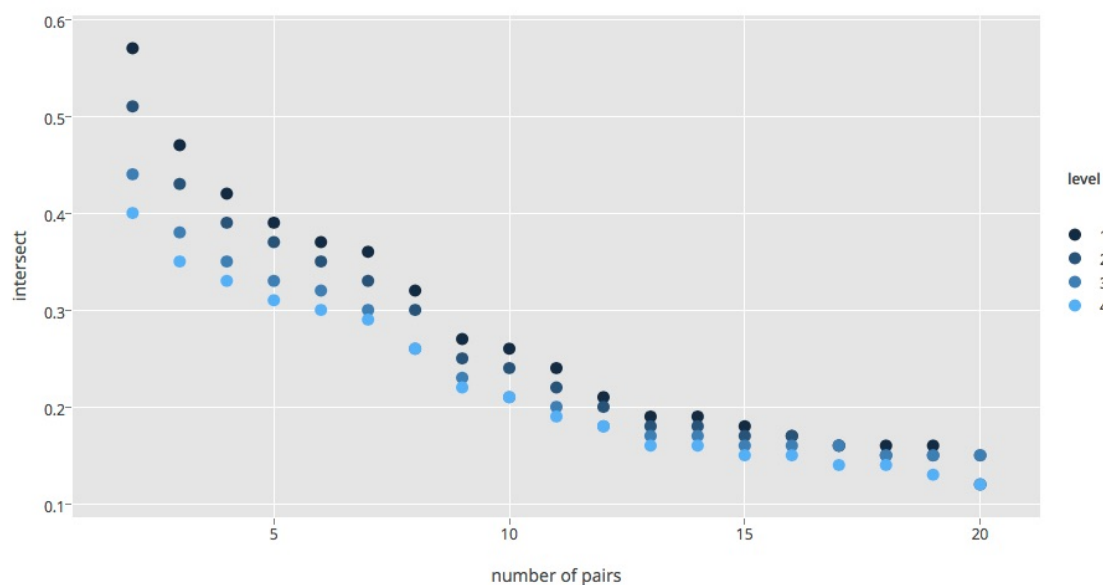
Fig. 9.6 The evolution of the edge of the twilight zone as more pairs are added. When the mean is above the value shown, the set is most likely to become INC.

9.5 ChEMBL hERG IC$_{50}$ case study

In order to bring this analysis to life, the ChEMBL public domain data for hERG IC$_{50}$ values was analysed. Plots similar to those shown in Figures 9.4 and 9.5 were generated from a set comprising 7347 compounds forming 733 matched pairs that could be grouped into sets of size 259, 241, 161 and 72 with context levels 1-4 respectively (Figure 9.7).
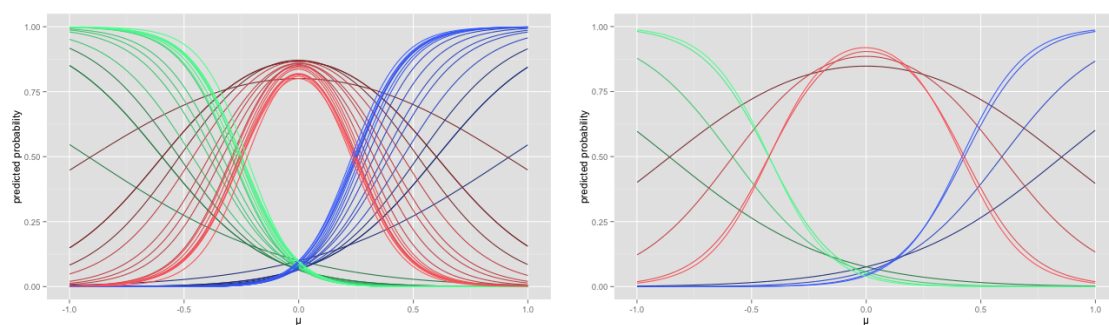
Fig. 9.7 The probability that a set of N pairs (left) and set of 5 pairs (right) with mean value μ will evolve into a set of pairs classified as INC (blue), DEC (green) or NED (red). The colouring starts dark for N=2 or context level 1 and becomes progressively lighter as N increases to 20 or context level 4.

An illustrative set with context level 4 is provided in figure 9.8 wherein exchanging the highlighted fluorine atom with a methyl group increases the hERG binding in all seven

cases, when the set is defined at context level 3 or 4: molecules bearing a *p*-fluorophenyl group.
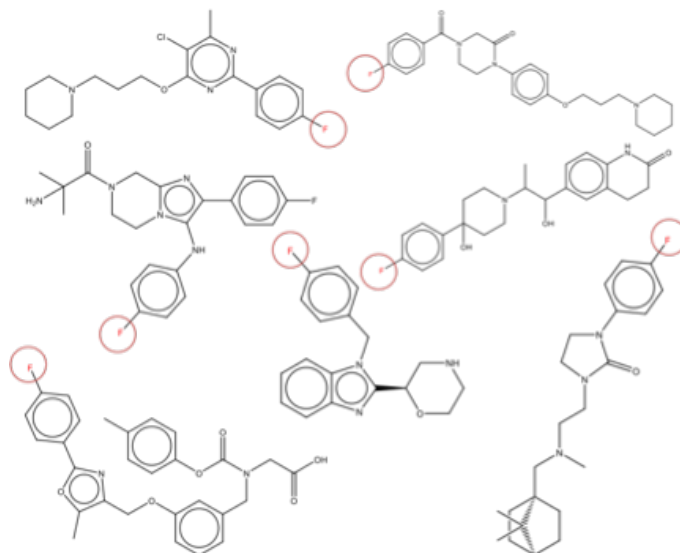
Fig. 9.8 A set of 7 compounds grouped at context level 4 where exchanging the highlighted F for Me increases hERG potency.

When the set is structurally less well-defined, larger sets of pairs are obtained (18 at context level 2 and 51 at level 1) which do not statistically support an underlying bias suggesting that the general effect of exchanging fluorine and methyl is too small to measure easily. Furthermore, the sets of pairs with $\overline{\mu_5}$ in the range 0.45 to 0.55 include 3, 5, 3 and 3 sets of the 108 in the INC group; 9, 5, 4 and 1 of the 516 in the NED group with context levels 1-4 respectively in each case and none of the DEC sets. This distribution is in line with what would be expected given the plot in Figure 9.5.

9.6   Conclusions

In the coin toss analogy, every coin is identical and the outcome is random. For molecules, each is unique and the effect of a structural transformation will usually be dependent on the molecule. The purpose of MMPA is to identify transformations that are highly likely to engender the same effect if applied again.

By using a combination of statistics and chemistry, a progress has been made towards answering the dilemma 'is MMPA more about chemistry or statistics?'- it is both! Having access to a very large dataset enabled monitoring of how the mean change in property evolves as more pairs are added. This led to the conclusion that, when predicting the effect of a structural transformation with only a few examples in the dataset, the uncertainty in the outcome heavily depends on the structural environment in which the transformation is

made. As more and more pairs are added, chemical context becomes less influential, and decisions can be driven by statistics.

References

[1] K. A. Dill, *Journal of Biological Chemistry*, 1997, **272**, 701–704.

[2] G. A. Jeffrey and G. A. Jeffrey, *An introduction to hydrogen bonding*, Oxford university press New York, 1997, vol. 12.

[3] C. Bissantz, B. Kuhn and M. Stahl, *Journal of Medicinal Chemistry*, 2010, **53**, 5061–5084.

[4] E. Barratt, A. Bronowska, J. Vondrášek, J. Černỳ, R. Bingham, S. Phillips and S. W. Homans, *Journal of Molecular Biology*, 2006, **362**, 994–1003.

[5] T. G. Davies, J. Bentley, C. E. Arris, F. T. Boyle, N. J. Curtin, J. A. Endicott, A. E. Gibson, B. T. Golding, R. J. Griffin and I. R. Hardcastle, *Nature Structural & Molecular Biology*, 2002, **9**, 745–749.

[6] K. N. Houk, A. G. Leach, S. P. Kim and X. Zhang, *Angewandte Chemie International Edition*, 2003, **42**, 4872–4897.

[7] R. Perozzo, G. Folkers and L. Scapozza, *Journal of Receptors and Signal Transduction*, 2004, **24**, 1–52.

[8] L. F. Scatena, *Science*, 2001, **292**, 908–912.

[9] J. L. Finney and A. K. Soper, *Chemical Society Reviews*, 1994, **23**, 1.

[10] P. W. Snyder, M. R. Lockett, D. T. Moustakas and G. M. Whitesides, *The European Physical Journal Special Topics*, 2013, **223**, 853–891.

[11] M. Boehringer, H. Fischer, M. Hennig, D. Hunziker, J. Huwyler, B. Kuhn, B. M. Loeffler, T. Luebbers, P. Mattei and R. Narquizian, *Bioorganic & medicinal chemistry letters*, 2010, **20**, 1106–1108.

[12] Y. Mao, M. A. Ratner and M. F. Jarrold, *Journal of the American Chemical Society*, 2000, **122**, 2950–2951.

[13] S. Fischer and C. S. Verma, *Proceedings of the National Academy of Sciences*, 1999, **96**, 9613–9615.

[14] J. M. Chen, S. L. Xu, Z. Wawrzak, G. S. Basarab and D. B. Jordan, *Biochemistry*, 1998, **37**, 17735–17744.

[15] C. A. Hunter and J. K. Sanders, *Journal of the American Chemical Society*, 1990, **112**, 5525–5534.

[16] S. Tsuzuki, K. Honda, T. Uchimaru, M. Mikami and K. Tanabe, *Journal of the American Chemical Society*, 2002, **124**, 104–112.

[17] P. Hobza, H. L. Selzle and E. W. Schlag, *Journal of the American Chemical Society*, 1994, **116**, 3500–3506.

[18] H. Krause, B. Ernstberger and H. Neusser, *Chemical Physics Letters*, 1991, **184**, 411–417.

[19] B. Meyer and T. Peters, *Angewandte Chemie International Edition*, 2003, **42**, 864–890.

[20] J. P. Gallivan and D. A. Dougherty, *Journal of the American Chemical Society*, 2000, **122**, 870–874.

[21] B. W. Berry, M. M. Elvekrog and C. Tommos, *Journal of the American Chemical Society*, 2007, **129**, 5308–5309.

[22] S. Roehrig, A. Straub, J. Pohlmann, T. Lampe, J. Pernerstorfer, K.-H. Schlemmer, P. Reinemer and E. Perzborn, *Journal of Medicinal Chemistry*, 2005, **48**, 5900–5908.

[23] P. Zhou, F. Tian, J. Zou and Z. Shang, *Mini reviews in medicinal chemistry*, 2010, **10**, 309–314.

[24] R. Wilcken, M. O. Zimmermann, A. Lange, A. C. Joerger and F. M. Boeckler, *Journal of Medicinal Chemistry*, 2013, **56**, 1363–1388.

[25] T. Clark, M. Hennemann, J. S. Murray and P. Politzer, *Journal of Molecular Modeling*, 2006, **13**, 291–296.

[26] P. Politzer, J. S. Murray and T. Clark, *Physical Chemistry Chemical Physics*, 2013, **15**, 11178–11189.

[27] M. G. Sarwar, B. Dragisic, L. J. Salsberg, C. Gouliaras and M. S. Taylor, *Journal of the American Chemical Society*, 2010, **132**, 1646–1653.

[28] K. Eskandari and M. Lesani, *Chemistry–A European Journal*, 2015, **21**, 4739–4746.

[29] J. Ren, Y. He, W. Chen, T. Chen, G. Wang, Z. Wang, Z. Xu, X. Luo, W. Zhu and H. Jiang, *Journal of Medicinal Chemistry*, 2014, **57**, 3588–3593.

[30] A. Mittermaier and E. Meneses, in *Protein-Ligand Interactions*, Springer Science Business Media, 2013, pp. 243–266.

[31] M. Pellecchia, I. Bertini, D. Cowburn, C. Dalvit, E. Giralt, W. Jahnke, T. L. James, S. W. Homans, H. Kessler, C. Luchinat, B. Meyer, H. Oschkinat, J. Peng, H. Schwalbe and G. Siegal, *Nature Reviews Drug Discovery*, 2008, **7**, 738–745.

[32] B. Rupp, *Biomolecular crystallography: principles, practice, and application to structural biology*, Garland Science, 2009.

[33] K. M. Merz Jr, D. Ringe and C. H. Reynolds, *Drug design: structure-and ligand-based approaches*, Cambridge University Press, 2010.

[34] A. Davis, S. St-Gallay and G. Kleywegt, *Drug Discovery Today*, 2008, **13**, 831–841.

[35] M. C. Deller and B. Rupp, *Journal of Computer-Aided Molecular Design*, 2015, **29**, 817–36.

[36] C. H. Reynolds, *ACS Medicinal Chemistry Letters*, 2014, **5**, 727–729.

[37] A. L. Lamb, T. J. Kappock and N. R. Silvaggi, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 2015, **1854**, 258–268.

[38] E. Perola and P. S. Charifson, *Journal of Medicinal Chemistry*, 2004, **47**, 2499–2510.

[39] G. Rhodes, *Crystallography made crystal clear: a guide for users of macromolecular models*, Academic press, 2010.

[40] E. Perola and P. S. Charifson, *Journal of Medicinal Chemistry*, 2004, **47**, 2499–2510.

[41] U. Ryde and K. Nilsson, *Journal of Molecular Structure: THEOCHEM*, 2003, **632**, 259–275.

[42] M. Sitzmann, I. E. Weidlich, I. V. Filippov, C. Liao, M. L. Peach, W.-D. Ihlenfeldt, R. G. Karki, Y. V. Borodina, R. E. Cachau and M. C. Nicklaus, *Journal of Chemical Information and Modeling*, 2012, **52**, 739–756.

[43] O. Y. Borbulevych, J. A. Plumley, R. I. Martin, K. M. Merz and L. M. Westerhoff, *Acta crystallographica. Section D, Biological crystallography*, 2014, **70**, 1233–47.

[44] J. Tirado-Rives and W. L. Jorgensen, *Journal of Medicinal Chemistry*, 2006, **49**, 5880–5884.

[45] J. Liebeschuetz, J. Hennemann, T. Olsson and C. R. Groom, *Journal of Computer-Aided Molecular Design*, 2012, **26**, 169–183.

[46] G. Holdgate and W. Ward, *Drug Discov Today*, 2005, **10**, 1543–50.

[47] C. Yung-Chi and W. H. Prusoff, *Biochemical Pharmacology*, 1973, **22**, 3099–3108.

[48] E. Freire, in *Proteomics and Protein-Protein Interactions*, Springer, 2005, pp. 291–307.

[49] J. E. Ladbury, G. Klebe and E. Freire, *Nature Reviews Drug Discovery*, 2010, **9**, 23–27.

[50] Z. Lin, F. Schwartz and E. Eisenstein, *Journal of Biological Chemistry*, 1995, **270**, 1011–4.

[51] V. Srinivas, G. Bhanuprakash Reddy and A. Surolia, *FEBS letters*, 1999, **450**, 181–185.

[52] J. Chodera and D. Mobley, *Annual Review of Biophysics*, 2013, **42**, 121–42.

[53] T. S. Olsson, M. A. Williams, W. R. Pitt and J. E. Ladbury, *Journal of Molecular Biology*, 2008, **384**, 1002–1017.

[54] E. Freire, *Drug Discovery Today*, 2008, **13**, 869–874.

[55] J. C. Martinez, E. S. Cobos, I. Luque, J. Murciano-Calles, J. Ruiz-Sanz and M. Iglesias-Bexiga, *Isothermal titration calorimetry: thermodynamic analysis of the binding thermograms of molecular recognition events by using equilibrium models*, INTECH Open Access Publisher, 2013.

[56] L. MicroCal, *MicroCal, LLC, Northampton, MA*, 2004.

[57] G. A. Holdgate, *Expert Opinion on Drug Discovery*, 2007, **2**, 1103–1114.

[58] H. Fukada and K. Takahashi, *Proteins*, 1998, **33**, 159–66.

[59] R. B. Raffa, G. W. Stagliano and S. D. Spencer, *European Journal of Pharmacology*, 2004, **483**, 323–324.

[60] B. Baum, L. Muley, M. Smolinski, A. Heine, D. Hangauer and G. Klebe, *Journal of Molecular Biology*, 2010, **397**, 1042–1054.

[61] M. M. Lopez and G. I. Makhatadze, in *Differential Scanning Calorimetry*, ed. H. J. Vogel, Springer New York, Totowa, NJ, 2002, pp. 113–119.

[62] M. M. Hann and G. M. Keserü, *Nature Reviews Drug Discovery*, 2012, **11**, year.

[63] M. P. Gleeson, A. Hersey, D. Montanari and J. Overington, *Nature Reviews Drug Discovery*, 2011, **10**, 197–208.

[64] M. Gleeson, A. Hersey, D. Montanari and J. Overington, *Nature Reviews Drug Discovery*, 2011, **10**, 197–208.

[65] P. D. Leeson and B. Springthorpe, *Nature Reviews Drug Discovery*, 2007, **6**, 881–890.

[66] B. Testa, P. Crivori, M. Reist and P.-A. Carrupt, *Perspectives in Drug Discovery and Design*, 2000, **19**, 179–211.

[67] P. W. Kenny, A. Leitão and C. A. Montanari, *Journal of Computer-Aided Molecular Design*, 2014, **28**, 699–710.

[68] A. Hopkins, C. Groom and A. Alex, *Drug Discov Today*, 2004, **9**, 430–1.

[69] T. Ryckmans, M. Edwards, V. Horne, A. Correia, D. Owen, L. Thompson, I. Tran, M. Tutt and T. Young, *Bioorganic & medicinal chemistry letters*, 2009, **19**, 4406–9.

[70] P. W. Kenny, C. A. Montanari and I. M. Prokopczyk, *Journal of Computer-Aided Molecular Design*, 2013, **27**, 389–402.

[71] M. D. Shultz, *Bioorganic & Medicinal Chemistry Letters*, 2013, **23**, 5980–5991.

[72] M. D. Shultz, *Bioorganic & Medicinal Chemistry Letters*, 2013, **23**, 5992–6000.

[73] H.-X. Zhou and M. K. Gilson, *Chemical Reviews*, 2009, **109**, 4092–4107.

[74] P. Ehrlich, 1913.

[75] G. Dahl and T. Akerud, *Drug Discovery Today*, 2013, **18**, 697–707.

[76] M. J. Waring, *Expert Opinion on Drug Discovery*, 2010, **5**, 235–248.

[77] R. A. Copeland, D. L. Pompliano and T. D. Meek, *Nature Reviews Drug Discovery*, 2006, **5**, 730–739.

[78] H. Lu and P. J. Tonge, *Current Opinion in Chemical Biology*, 2010, **14**, 467–474.

[79] S. A. Lipton, *Nature Reviews Drug Discovery*, 2006, **5**, 160–170.

[80] S. Núñez, J. Venhorst and C. G. Kruse, *Drug Discovery Today*, 2012, **17**, 10–22.

[81] I. Dierynck, M. D. Wit, E. Gustin, I. Keuleers, J. Vandersmissen, S. Hallenberger and K. Hertogs, *Journal of Virology*, 2007, **81**, 13845–13851.

[82] N. M. King, M. Prabu-Jeyabalan, E. A. Nalivaika, P. Wigerinck, M.-P. de Bethune and C. A. Schiffer, *Journal of Virology*, 2004, **78**, 12012–12021.

[83] P. J. Tummino and R. A. Copeland, *Biochemistry*, 2008, **47**, 5481–5492.

[84] G. A. Holdgate and A. L. Gill, *Drug Discovery Today*, 2011, **16**, 910–913.

[85] J. McDonnell, *Current Opinion in Chemical Biology*, 2001, **5**, 572–7.

[86] G. De Crescenzo, C. Boucher, Y. Durocher and M. Jolicoeur, *Cellular and Molecular Bioengineering*, 2008, **1**, 204–215.

[87] *SPRPages*, http://www.sprpages.nl, Accessed: 2016-10-04.

[88] P. A. Van Der Merwe, *Surface plasmon resonance*, 2001.

[89] *Guide to SPR Data Analysis on the ProteOn™ XPR36 System*, http://www.bioradiations.com/guide-to-spr-data-analysis-on-the-proteon-xpr36-system/, Accessed: 2016-10-04.

[90] J. J. Maguire, R. E. Kuc and A. P. Davenport, in *Receptor Binding Techniques*, Springer Science Business Media, 2012, pp. 31–77.

[91] D. Guo, E. J. H. van Dorp, T. Mulder-Krieger, J. P. D. van Veldhoven, J. Brussee, A. P. IJzerman and L. H. Heitman, *Journal of Biomolecular Screening*, 2012, **18**, 309–320.

[92] H. Motulsky and L. Mahan, *Molceular Pharmacology*, 1984, **25**, 1–9.

[93] A. Emami-Nemini, T. Roux, M. Leblay, E. Bourrier, L. Lamarque, E. Trinquet and M. J. Lohse, *Nature Protocols*, 2013, **8**, 1307–1320.

[94] W. M. Bujalowski and M. J. Jezewska, in *Methods in Molecular Biology*, Springer Science Business Media, 2012, pp. 105–133.

[95] D. L. Mobley and K. A. Dill, *Structure*, 2009, **17**, 489–498.

[96] M. K. Gilson and H.-X. Zhou, *Annual Review of Biophysics and Biomolecular Structure*, 2007, **36**, 21.

[97] A. K. Bronowska, *Thermodynamics of ligand-protein interactions: Implications for molecular design*, INTECH Open Access Publisher, 2011.

[98] H. Gohlke, M. Hendlich and G. Klebe, *Journal of Molecular Biology*, 2000, **295**, 337–356.

[99] S.-Y. Huang, S. Z. Grinter and X. Zou, *Physical Chemistry Chemical Physics*, 2010, **12**, 12899–12908.

[100] H.-J. Bhm, *Journal of Computer-Aided Molecular Design*, 1992, **6**, 593–606.

[101] M. Eldridge, C. Murray, T. Auton, G. Paolini and R. Mee, *Journal of Computer-Aided Molecular Design*, 1997, **11**, 425–45.

[102] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *Journal of Medicinal Chemistry.*, 2004, **47**, 1739–1749.

[103] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin, *Journal of Molecular Biology*, 1982, **161**, 269–288.

[104] G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *Journal of Molecular Biology*, 1997, **267**, 727–748.

[105] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *Journal of Computational Chemistry*, 2009, **30**, 2785–2791.

[106] I. Muegge and Y. C. Martin, *Journal of Medicinal Chemistry.*, 1999, **42**, 791–804.

[107] J.-Y. Trosset and H. A. Scheraga, *Journal of Computational Chemistry*, 1999, **20**, 412–427.

[108] M. A. C. Neves, M. Totrov and R. Abagyan, *Journal of Computer-Aided Molecular Design*, 2012, **26**, 675–686.

[109] M. Liu and S. Wang, *Journal of Computer-Aided Molecular Design*, 1999, **13**, 435–51.

[110] T. Hart and R. Read, *Proteins*, 1992, **13**, 206–22.

[111] K. P. Clark and Ajay, *Journal of Computational Chemistry*, 1995, **16**, 1210–1226.

[112] J. S. Taylor and R. M. Burnett, *Proteins: Structure, Function and Genetics*, 2000, **41**, 173–191.

[113] S. F. Sousa, P. A. Fernandes and M. J. Ramos, *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**, 15–26.

[114] X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, *Current Computer Aided-Drug Design*, 2011, **7**, 146–157.

[115] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *Journal of computational chemistry*, 1998, **19**, 1639–1662.

[116] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *Journal of Computational Chemistry*, 1983, **4**, 187–217.

[117] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *Journal of the American Chemical Society*, 1995, **117**, 5179–5197.

[118] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger and W. F. van Gunsteren, *Journal of Physical Chemistry A*, 1999, **103**, 3596–3607.

[119] W. L. Jorgensen and J. Tirado-Rives, *Journal of the American Chemical Society*, 1988, **110**, 1657–1666.

[120] F. Jensen, *Introduction to Computational Chemistry*, Wiley, 2007.

[121] H. L. N. de Amorim, R. A. Caceres and P. A. Netz, *Current Drug Targets*, 2008, **9**, 1100–1105.

[122] J. Åqvist, V. B. Luzhkov and B. O. Brandsdal, *Accounts of Chemical Research*, 2002, **35**, 358–365.

[123] H. L. N. de Amorim, R. A. Caceres and P. A. Netz, *Current Drug Targets*, 2008, **9**, 1100–1105.

[124] T. Hansson, J. Marelius and J. Åqvist, *Journal of Computer-Aided Molecular Design*, 1998, **12**, 27–35.

[125] J. Åqvist, C. Medina and J.-E. Samuelsson, *Protein Engineering*, 1994, **7**, 385–391.

[126] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan and W. Wang, *Accounts of Chemical Research*, 2000, **33**, 889–897.

[127] S. Genheden and U. Ryde, *Expert Opinion on Drug Discovery*, 2015, **10**, 449–461.

[128] S. Genheden and U. Ryde, *Journal of Chemical Theory and Computation*, 2011, **7**, 3768–78.

[129] S. Genheden, *Journal of Computer-Aided Molecular Design*, 2011, **25**, 1085–93.

[130] P. Mikulskis, S. Genheden, P. Rydberg, L. Sandberg, L. Olsen and U. Ryde, *Journal of Computer-Aided Molecular Design*, 2012, **26**, 527–41.

[131] B. Kuhn and P. Kollman, *Journal of Medicinal Chemistry*, 2000, **43**, 3786–91.

[132] S. Genheden, I. Nilsson and U. Ryde, *Journal of Chemical Information and Modeling*, 2011, **51**, 947–958.

[133] V. Helms and R. C. Wade, *Journal of the American Chemical Society*, 1998, **120**, 2710–2713.

[134] M. Gilson, J. Given, B. Bush and J. McCammon, *Biophysical Journal*, 1997, **72**, 1047–1069.

[135] G. Archontis, K. A. Watson, Q. Xie, G. Andreou, E. D. Chrysina, S. E. Zographos, N. G. Oikonomakos and M. Karplus, *Proteins*, 2005, **61**, 984–998.

[136] C. Chipot, X. Rozanska and S. B. Dixit, *Journal of Computer-Aided Molecular Design*, 2005, **19**, 765–770.

[137] S. Genheden, I. Nilsson and U. Ryde, *Journal of Chemical Information and Modeling*, 2011, **51**, 947–58.

[138] H. Fujitani, Y. Tanida, M. Ito, G. Jayachandran, C. D. Snow, M. R. Shirts, E. J. Sorin and V. S. Pande, *The Journal of Chemical Physics*, 2005, **123**, 084108.

[139] S. Boresch, F. Tettinger, M. Leitgeb and M. Karplus, *The Journal of Physical Chemistry B*, 2003, **107**, 9535–9551.

[140] S. Donnini and A. H. Juffer, *Journal of Computational Chemistry*, 2003, **25**, 393–411.

[141] J. W. Ponder and D. A. Case, in *Protein Simulations*, Elsevier BV, 2003, pp. 27–85.

[142] S. Izrailev, S. Stepaniants, B. Isralewitz, D. Kosztin, H. Lu, F. Molnar, W. Wriggers and K. Schulten, in *Computational Molecular Dynamics: Challenges Methods, Ideas*, Springer Science Business Media, 1999, pp. 39–65.

[143] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1992, **13**, 1011–1021.

[144] A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences*, 2002, **99**, 12562–12566.

[145] H.-J. Woo and B. Roux, *Proceedings of the National Academy of Sciences*, 2005, **102**, 6825–6830.

[146] J. C. Gumbart, B. Roux and C. Chipot, *Journal of Chemical Theory and Computation*, 2013, **9**, 794–802.

[147] V. Limongelli, M. Bonomi and M. Parrinello, *Proceedings of the National Academy of Sciences*, 2013, **110**, 6358–6363.

[148] N. Ferruz and G. D. Fabritiis, *Molecular Informatics*, 2016, **35**, 216–226.

[149] K. M. ElSawy, R. Twarock, D. P. Lane, C. S. Verma and L. S. D. Caves, *Journal of Chemical Theory and Computation*, 2012, **8**, 314–321.

[150] L. W. Votapka and R. E. Amaro, *PLOS Computational Biology*, 2015, **11**, e1004381.

[151] V. Limongelli, M. Bonomi, L. Marinelli, F. L. Gervasio, A. Cavalli, E. Novellino and M. Parrinello, *Proceedings of the National Academy of Sciences*, 2010, **107**, 5411–5416.

[152] F. Pietrucci, F. Marinelli, P. Carloni and A. Laio, *Journal of the American Chemical Society*, 2009, **131**, 11811–11818.

[153] D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Lerardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. M. Deneroff, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S. C. Wang, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson and K. J. Bowers, *Communications of the ACM*, 2008, **51**, 91.

[154] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan and W. Wriggers, *Science*, 2010, **330**, 341–346.

[155] A. R. Leach, *Molecular modelling : principles and applications*, Prentice Hall, Harlow, New York, 2001.

[156] A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648–5652.

[157] R. Bauernschmitt and R. Ahlrichs, *Chemical Physics Letters*, 1996, **256**, 454–464.

[158] C. Lee, W. Yang and R. G. Parr, *Physical Review B*, 1988, **37**, 785.

[159] Y. Zhao and D. G. Truhlar, *Theoretical Chemistry Accounts*, 2008, **120**, 215–241.

[160] M. J. Dewar and W. Thiel, *Journal of the American Chemical Society*, 1977, **99**, 4899–4907.

[161] M. J. Dewar, E. G. Zoebisch, E. F. Healy and J. J. Stewart, *Journal of the American Chemical Society*, 1985, **107**, 3902–3909.

[162] M. J. Dewar, E. F. Healy, A. J. Holder and Y.-C. Yuan, *Journal of Computational Chemistry*, 1990, **11**, 541–542.

[163] J. J. Stewart, *Journal of Molecular modeling*, 2007, **13**, 1173–1213.

[164] R. Boča, *International Journal of Quantum Chemistry*, 1987, **31**, 941–950.

[165] S. R. Pruitt, D. G. Fedorov, K. Kitaura and M. S. Gordon, *Journal of Chemical Theory and Computation*, 2009, **6**, 1–5.

[166] M. N. Ucisik, Z. Zheng, J. C. Faver and K. M. Merz, *Journal of Chemical Theory and Computation*, 2014, **10**, 1314–1325.

[167] U. Ryde and P. Söderhjelm, *Chemical Reviews*, 2016, **116**, 5520–5566.

[168] M. Peräkylä and T. A. Pakkanen, *Proteins: Structure, Function, and Bioinformatics*, 1994, **20**, 367–372.

[169] M. Peräkylä and T. A. Pakkanen, *Proteins: Structure, Function, and Bioinformatics*, 1995, **21**, 22–29.

[170] E. Nikitina, V. Sulimov, V. Zayets and N. Zaitseva, *International Journal of Quantum Chemistry*, 2004, **97**, 747–763.

[171] E. Nikitina, V. Sulimov, F. Grigoriev, O. Kondakova and S. Luschekina, *International Journal of Quantum Chemistry*, 2006, **106**, 1943–1963.

[172] K. Raha and K. M. Merz, *Journal of Medicinal Chemistry*, 2005, **48**, 4558–4575.

[173] R. Villar, M. Gil, J. Garcia and V. Martínez-Merino, *Journal of Computational Chemistry*, 2005, **26**, 1347–1358.

[174] A. G. Leach, L.-L. Olsson and D. J. Warner, *Medicinal Chemical Communications*, 2013, **4**, 180–186.

[175] K. Roos, J. Viklund, J. Meuller, K. Kaspersson and M. Svensson, *Journal of Chemical Information and Modeling*, 2014, **54**, 818–825.

[176] K. Roos, A. Hogner, D. Ogg, M. J. Packer, E. Hansson, K. L. Granberg, E. Evertsson and A. Nordqvist, *Journal of Computer-Aided Molecular Design*, 2015, **29**, 1109–1122.

[177] M. P. Mazanetz, O. Ichihara, R. J. Law and M. Whittaker, *Journal of Cheminformatics*, 2011, **3**, 1.

[178] T. Otsuka, N. Okimoto and M. Taiji, *Journal of Computational Chemistry*, 2015, **36**, 2209–2218.

[179] Y. Nishimoto and D. G. Fedorov, *Physical Chemistry Chemical Physics*, 2016.

[180] A. Heifetz, E. I. Chudyk, L. Gleave, M. Aldeghi, V. Cherezov, D. G. Fedorov, P. C. Biggin and M. J. Bodkin, *Journal of Chemical Information and Modeling*, 2015, **56**, 159–172.

[181] A. Heifetz, E. I. Chudyk, L. Gleave, M. Aldeghi, V. Cherezov, D. G. Fedorov, P. C. Biggin and M. J. Bodkin, *Journal of Chemical Information and Modeling*, 2015, **56**, 159–172.

[182] O. Warburg, *Science*, 1956, **123**, 309–314.

[183] F. Hirschhaeuser, U. G. Sattler and W. Mueller-Klieser, *Cancer Research*, 2011, **71**, 6921–6925.

[184] D. Balinsky, C. E. Platz and J. W. Lewis, *Cancer Research*, 1983, **43**, 5895–5901.

[185] R. D. Goldman, N. O. Kaplan and T. C. Hall, *Cancer Research*, 1964, **24**, 389–399.

[186] M. G. Vander Heiden, *Nature Reviews Drug discovery*, 2011, **10**, 671–684.

[187] D. A. Tennant, R. V. Durán and E. Gottlieb, *Nature Reviews Cancer*, 2010, **10**, 267–277.

[188] P. S. Dragovich, K. W. Bair, T. Baumeister, Y.-C. Ho, B. M. Liederer, X. Liu, Y. Liu, T. O'Brien, J. Oeh and D. Sampath, *Bioorganic & medicinal chemistry letters*, 2013, **23**, 4875–4885.

[189] P. S. Dragovich, B. P. Fauber, J. Boggs, J. Chen, L. B. Corson, C. Z. Ding, C. Eigenbrot, H. Ge, A. M. Giannetti, T. Hunsaker, S. Labadie, C. Li, Y. Liu, Y. Liu, S. Ma, S. Malek, D. Peterson, K. E. Pitts, H. E. Purkey, K. Robarge, L. Salphati, S. Sideris, M. Ultsch, E. Vanderporten, J. Wang, B. Wei, Q. Xu, I. Yen, Q. Yue, H. Zhang, X. Zhang and A. Zhou, *Bioorganic and Medicinal Chemistry Letters*, 2014, **24**, 3764–3771.

[190] P. S. Dragovich, B. P. Fauber, J. Boggs, J. Chen, L. B. Corson, C. Z. Ding, C. Eigenbrot, H. Ge, A. M. Giannetti, T. Hunsaker, S. Labadie, C. Li, Y. Liu, Y. Liu, S. Ma, S. Malek, D. Peterson, K. E. Pitts, H. E. Purkey, K. Robarge, L. Salphati, S. Sideris, M. Ultsch, E. Vanderporten, J. Wang, B. Wei, Q. Xu, I. Yen, Q. Yue, H. Zhang, X. Zhang and A. Zhou, *Bioorganic and Medicinal Chemistry Letters*, 2014, **24**, 3764–3771.

[191] P. S. Dragovich, B. P. Fauber, J. Boggs, J. Chen, L. B. Corson, C. Z. Ding, C. Eigenbrot, H. Ge, A. M. Giannetti, T. Hunsaker, S. Labadie, C. Li, Y. Liu, Y. Liu, S. Ma, S. Malek, D. Peterson, K. E. Pitts, H. E. Purkey, K. Robarge, L. Salphati, S. Sideris, M. Ultsch, E. Vanderporten, J. Wang, B. Wei, Q. Xu, I. Yen, Q. Yue, H. Zhang, X. Zhang and A. Zhou, *Bioorganic and Medicinal Chemistry Letters*, 2014, **24**, 5683–5687.

[192] P. S. Dragovich, B. P. Fauber, L. B. Corson, C. Z. Ding, C. Eigenbrot, H. Ge, A. M. Giannetti, T. Hunsaker, S. Labadie, Y. Liu, S. Malek, B. Pan, D. Peterson, K. Pitts, H. E. Purkey, S. Sideris, M. Ultsch, E. VanderPorten, B. Wei, Q. Xu, I. Yen, Q. Yue, H. Zhang and X. Zhang, *Bioorganic and Medicinal Chemistry Letters*, 2013, **23**, 3186–3194.

[193] A. Kohlmann, S. G. Zech, F. Li, T. Zhou, R. M. Squillace, L. Commodore, M. T. Greenfield, X. Lu, D. P. Miller and W.-S. Huang, *Journal of Medicinal Chemistry*, 2013, **56**, 1023–1040.

[194] R. A. Ward, C. Brassington, A. L. Breeze, A. Caputo, S. Critchlow, G. Davies, L. Goodwin, G. Hassall, R. Greenwood, G. A. Holdgate, M. Mrosek, R. A. Norman, S. Pearson, J. Tart, J. A. Tucker, M. Vogtherr, D. Whittaker, J. Wingfield, J. Winter and K. Hudson, *Journal of Medicinal Chemistry*, 2012, **55**, 3285–3306.

[195] ChemAxon, *Marvin 15.3.2*, 2015, http://www.chemaxon.com.

[196] O. Trott and A. J. Olson, *Journal of Computational Chemistry*, 2010, **31**, 455–461.

[197] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman, *Journal of Computer-Aided Molecular Design*, 2013, **27**, 221–234.

[198] R. Huey and G. Morris, *La Jolla, CA, USA: The Scripps Research Institute*, 2003.

[199] J. Gasteiger, C. Rudolph and J. Sadowski, *Tetrahedron Computer Methodology*, 1990, **3**, 537–547.

[200] L. Li, C. Li, Z. Zhang and E. Alexov, *Journal of Chemical Theory and Computation*, 2013, **9**, 2126–2136.

[201] F. Buckley and A. A. Maryott, *Tables of dielectric dispersion data for pure liquids and dilute solutions*, US Dept. of Commerce, National Bureau of Standards, 1958, vol. 589.

[202] C. Gao, M.-S. Park and H. A. Stern, *Biophysical Journal*, 2010, **98**, 901–910.

[203] (*a*) A. Deschênes and E. Sourial, *Chemical Computing Group*, 2007; (*b*) P. Labute, C. Williams, M. Feher, E. Sourial and J. M. Schmidt, *Journal of Medicinal Chemistry*, 2001, **44**, 1483–1490.

[204] M. Merski, M. Fischer, T. E. Balius, O. Eidam and B. K. Shoichet, *Proceedings of the National Academy of Sciences*, 2015, **112**, 5039–5044.

[205] N. M. Lim, L. Wang, R. Abel and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2016, **12**, 4620–4631.

[206] P. Adams, P. Afonine, G. Bunkóczi, V. Chen, I. Davis, N. Echols, J. Headd, L. Hung, G. Kapral, R. Grosse-Kunstleve, A. McCoy, N. Moriarty, R. Oeffner, R. Read, D. Richardson, J. Richardson, T. Terwilliger and P. Zwart, *Acta Crystallographica Section D*, 2010, **66**, 213–21.

[207] N. Moriarty, R. Grosse-Kunstleve and P. Adams, *Acta Crystallographica Section D*, 2009, **65**, 1074–80.

[208] *PHENIX Python-based Hierarchical ENvironment for Integrated Xtallography*, https://www.phenix-online.org/documentation/faqs/index.html, Accessed: 2016-10-19.

[209] R. Edgar, M. Domrachev and A. E. Lash, *Nucleic Acids Research*, 2002, **30**, 207–210.

[210] R. Possemato, K. M. Marks, Y. D. Shaul, M. E. Pacold, D. Kim, K. Birsoy, S. Sethumadhavan, H.-K. Woo, H. G. Jang and A. K. Jha, *Nature*, 2011, **476**, 346–350.

[211] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson and M. Urashima, *Nature*, 2010, **463**, 899–905.

[212] J. W. Locasale, A. R. Grassian, T. Melman, C. A. Lyssiotis, K. R. Mattaini, A. J. Bass, G. Heffron, C. M. Metallo, T. Muranen and H. Sharfi, *Nature Genetics*, 2011, **43**, 869–874.

[213] E. Mullarky, N. C. Lucki, R. B. Zavareh, J. L. Anglin, A. P. Gomes, B. N. Nicolay, J. C. Wong, S. Christen, H. Takahashi and P. K. Singh, *Proceedings of the National Academy of Sciences*, 2016, **113**, 1778–1783.

[214] T. Yokomizo, M. Nakasako, T. Yamazaki, H. Shindo and J. Higo, *Chemical Physics Letters*, 2005, **401**, 332–336.

[215] A. C. Chia-en, W. Chen and M. K. Gilson, *Proceedings of the National Academy of Sciences*, 2007, **104**, 1534–1539.

[216] S. Milev, *General Electric*, 2013, **9**, year.

[217] L. MicroCal, *MAU130030 (Rev. A). ed. Northampton: MicroCal LLC*, 2003.

[218] D. Myszka, Y. Abdiche, F. Arisaka, O. Byron, E. Eisenstein, P. Hensley, J. Thomson, C. Lombardo, F. Schwarz and W. Stafford, *Journal of Biomolecular Techniques: JBT*, 2003, **14**, 247.

[219] W. B. Turnbull and A. H. Daranas, *Journal of the American Chemical Society*, 2003, **125**, 14859–14866.

[220] H.-X. Zhou and P. A. Bates, *Current Opinion in Structural Biology*, 2013, **23**, 887–893.

[221] R. Evans, Z. Deng, A. K. Rogerson, A. S. McLachlan, J. J. Richards, M. Nilsson and G. A. Morris, *Angewandte Chemie International Edition*, 2013, **52**, 3199–3202.

[222] *Excel spreadsheet relating diffusion coefficient and molecular weight | Manchester NMR Methodology Group*, http://nmr.chemistry.manchester.ac.uk/?q=node/290.

[223] J. Batra, A. Szabó, T. R. Caulfield, A. S. Soares, M. Sahin-Tóth and E. S. Radisky, *Journal of Biological Chemistry*, 2013, **288**, 9848–9859.

[224] J. A. Caravella, J. D. Carbeck, D. C. Duffy, G. M. Whitesides and B. Tidor, *Journal of the American Chemical Society*, 1999, **121**, 4340–4347.

[225] N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, *Proceedings of the National Academy of Sciences*, 2001, **98**, 10037–10041.

[226] *World Health Organization global tuberculosis control*, 2011.

[227] D. J. Payne, P. V. Warren, D. J. Holmes, Y. Ji and J. T. Lonsdale, *Drug Discovery Today*, 2001, **6**, 537–544.

[228] D. A. Rozwarski, C. Vilchèze, M. Sugantino, R. Bittman and J. C. Sacchettini, *Journal of Biological Chemistry*, 1999, **274**, 15582–15589.

[229] S. R. Luckner, N. Liu, C. W. Am Ende, P. J. Tonge and C. Kisker, *Journal of Biological Chemistry*, 2010, **285**, 14330–14337.

[230] P. S. Shirude, P. Madhavapeddi, M. Naik, K. Murugan, V. Shinde, R. Nandishaiah, J. Bhat, A. Kumar, S. Hameed and G. Holdgate, *Journal of Medicinal Chemistry*, 2013, **56**, 8533–8542.

[231] P. Ward, *GSK Reveals New Collaboration on Genomics and Big Data*, http://www.appliedclinicaltrialsonline.com/gsk-reveals-new-collaboration-genomics-and-big-data, Accessed: 2016-14-11.

[232] A. G. Dossetter, E. J. Griffen and A. G. Leach, *Drug Discovery Today*, 2013, **18**, 724–731.

[233] E. Griffen, A. G. Leach, G. R. Robb and D. J. Warner, *Journal of Medicinal Chemistry*, 2011, **54**, 7739–7750.

[234] A. G. Leach, H. D. Jones, D. A. Cosgrove, P. W. Kenny, L. Ruston, P. MacFaul, J. M. Wood, N. Colclough and B. Law, *Journal of Medicinal Chemistry*, 2006, **49**, 6672–6682.

[235] P. W. K. Jens Sadowski, *Chemoinformatics in Drug Discovery*, 2006, **23**, year.

[236] D. J. Warner, E. J. Griffen and S. A. St-Gallay, *Journal of Chemical Information and Modeling*, 2010, **50**, 1350–1357.

[237] G. Papadatos, M. Alkarouri, V. J. Gillet, P. Willett, V. Kadirkamanathan, C. N. Luscombe, G. Bravi, N. J. Richmond, S. D. Pickett and J. a. Hussain, *Journal of Chemical Information and Modeling*, 2010, **50**, 1872–1886.

[238] M. C. Sanguinetti and M. Tristani-Firouzi, *Nature*, 2006, **440**, 463–469.

[239] H. Shannon and S. Walter, *Canadian Medical Association Journal*, 1995, **152**, 27.

[240] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[241] H. Wickham, *Journal of Statistical Software*, 2007, **21**, 1–20.

[242] H. Wickham, *ggplot2: elegant graphics for data analysis*, Springer Science & Business Media, 2009.

[243] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*, Springer Science & Business Media, 2013.

Appendix A

A.1 Representative ITC titrations for PHGDH

Table A.1 PHGDH raw ITC data for compound **2**

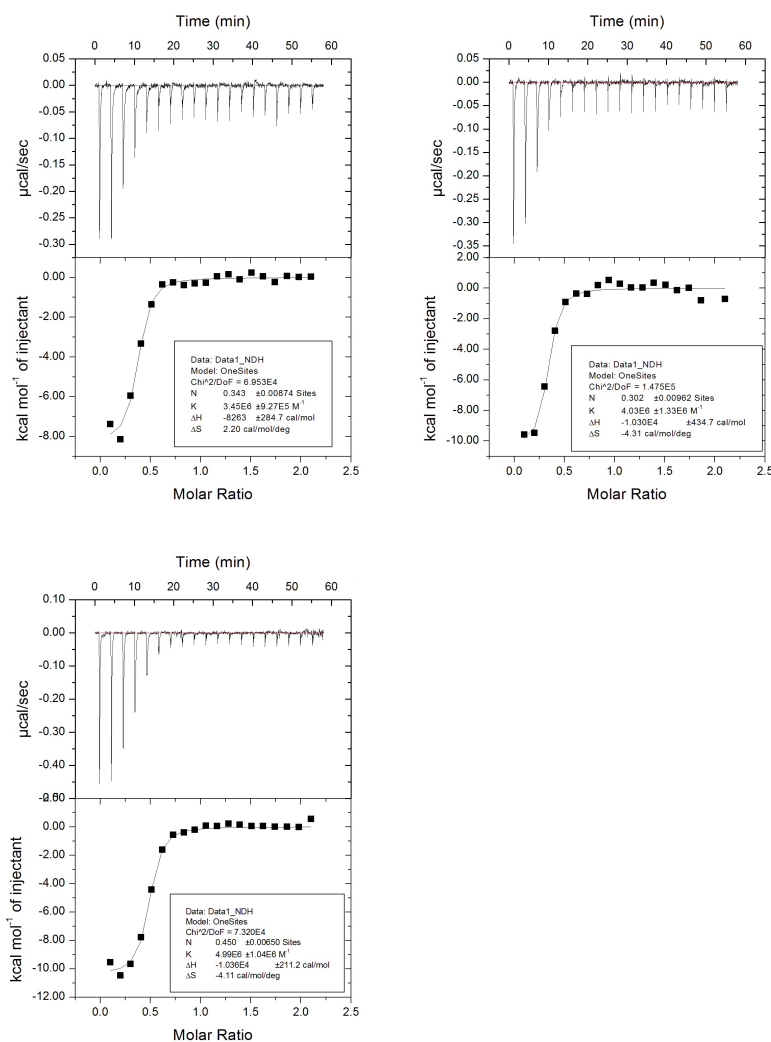| Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.34 ± 0.00	3.45 ± 0.92	-8.26 ± 0.28	0.62 ± 0.32	0.28 ± 0.07
HEPES	0.30 ± 0.00	4.03 ± 1.33	-10.30 ± 0.43	-1.32 ± 0.47	0.24 ± 0.08
phosphate	0.45 ± 0.00	4.99 ± 1.04	-10.36 ± 0.21	-1.25 ± 1.24	0.2 ± 0.41

Fig. A.1 Representative ITC titration for compound **2** in HEPES (upper row) and phosphate buffer (bottom row).

Table A.2 PHGDH raw ITC data for compound **5**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.41 ± 0.01	1.33 ± 0.36	-9.24 ± 0.47	0.91 ± 0.50	0.75 ± 0.20
HEPES	0.50 ± 0.00	3.33 ± 0.51	-7.98 ± 0.16	0.88 ± 0.18	0.30 ± 0.04
phosphate	0.5 ± 0.00	2.36 ± 0.30	-12.84 ± 0.21	-4.18 ± 0.22	0.42 ± 0.05
TRIS	0.69 ± 0.00	0.08± 0.16	-7.50 ± 0.07	1.08 ± 0.09	0.48 ± 0.03

Fig. A.2 Representative ITC titrations for compound **5** in HEPES (upper row), phosphate and TRIS buffer (bottom row).

Table A.3 PHGDH raw ITC data for compound **8**

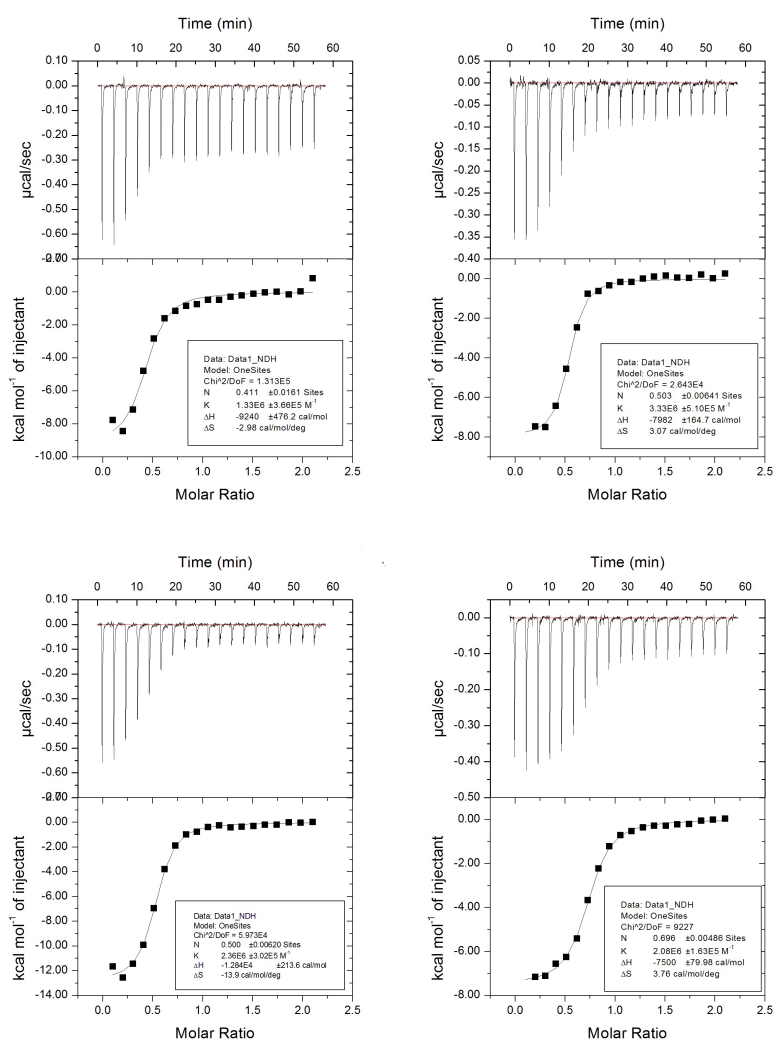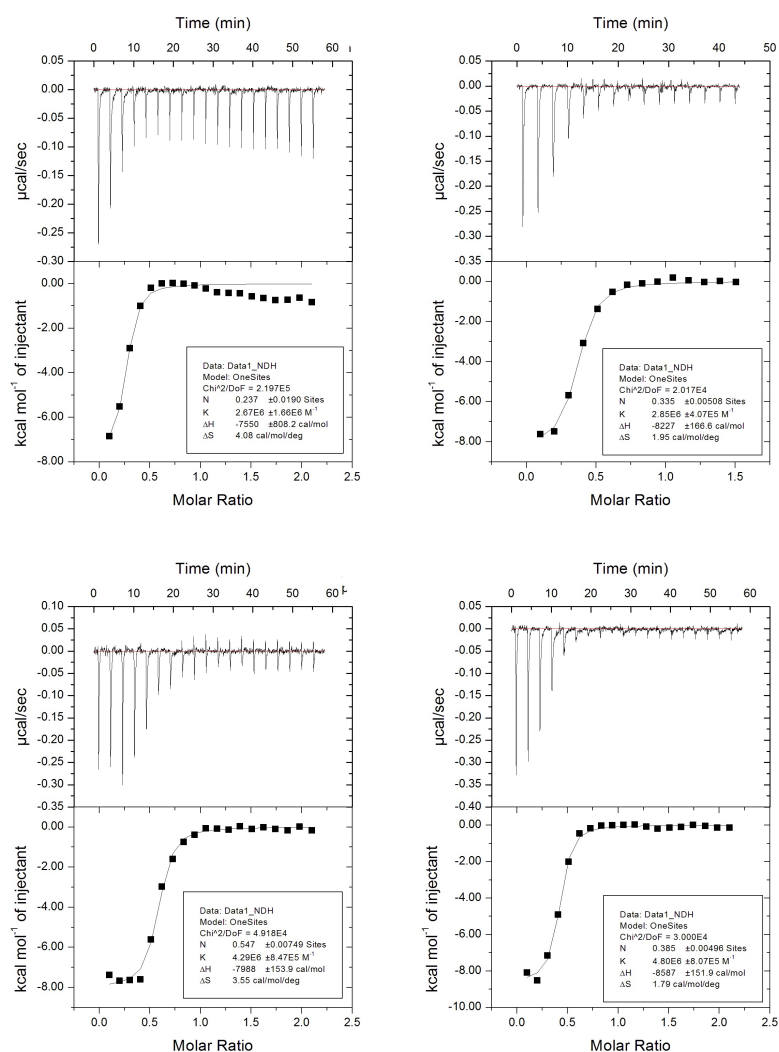| Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.59 ± 0.00	1.99 ± 0.21	-8.79 ± 0.10	-0.23 ± 0.12	0.50 ± 0.05
HEPES	0.33 ± 0.00	2.85 ± 0.40	-8.22 ± 0.16	0.55 ± 0.18	0.35 ± 0.05
phosphate	0.54 ± 0.00	4.29 ± 0.84	-7.98 ± 0.15	1.03 ± 0.19	0.04 ± 0.05
TRIS	0.38 ± 0.00	4.80± 0.80	-8.58 ± 0.15	0.49 ± 0.18	0.20 ± 0.03

Fig. A.3 Representative ITC titrations for compound **8** in HEPES (upper row), phosphate and TRIS buffer (bottom row).

Table A.4 PHGDH raw ITC data for compound **10**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.31 ± 0.00	6.88 ± 1.63	-10.60 ± 0.25	-1.30 ± 0.29	0.14 ± 0.03
HEPES	0.33 ± 0.00	9.12 ± 4.49	-10.20 ± 0.40	-0.74 ± 0.49	0.10 ± 0.05
phosphate	0.35 ± 0.00	22.1 ± 6.15	-13.44 ± 0.16	-3.45 ± 0.23	0.04 ± 0.01
TRIS	0.55 ± 0.00	17.6 ± 9.77	-9.64 ± 0.27	0.20 ± 0.42	0.05± 0.03

Fig. A.4 Representative ITC titrations for compound **10** in in HEPES (upper row), phosphate and TRIS buffer (bottom row).

Table A.5 PHGDH raw ITC data for compound **11**

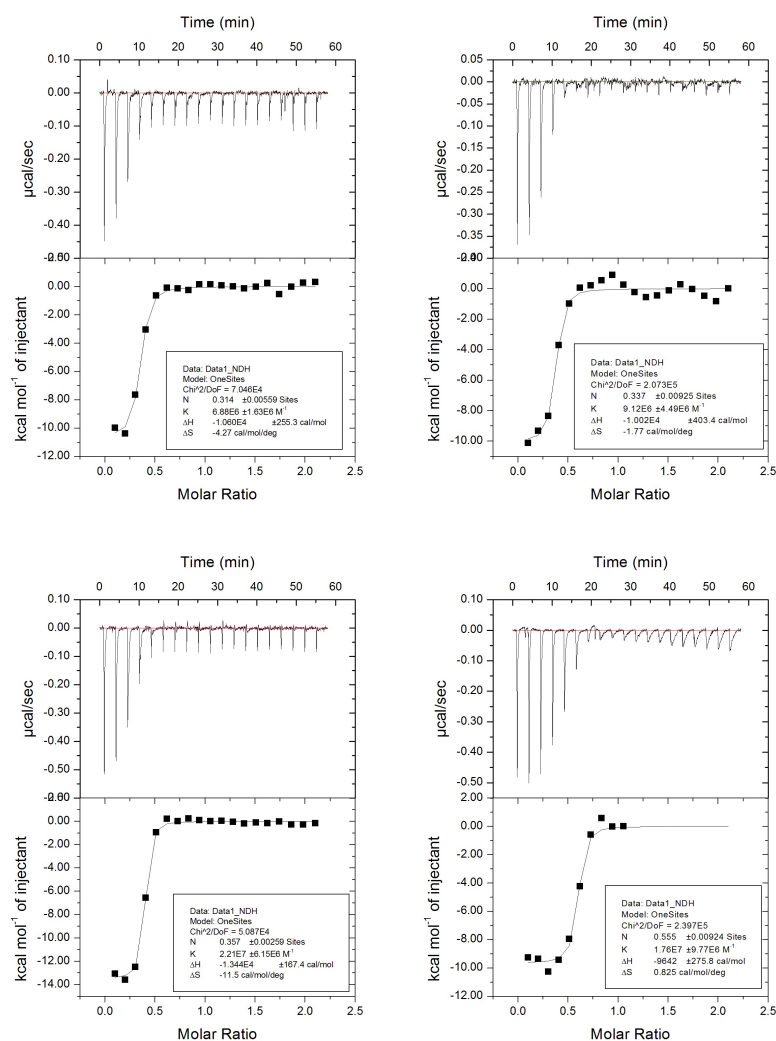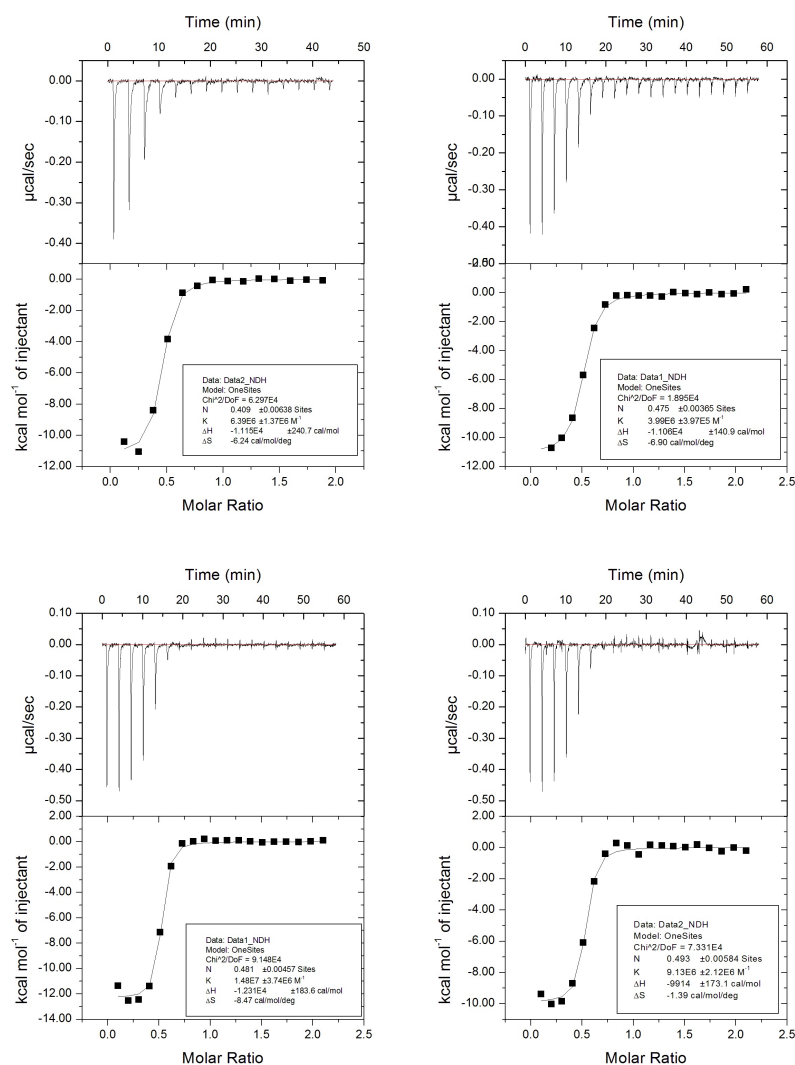| Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.40 ± 0.00	6.39 ± 1.37	-11.5 ± 0.24	-1.90 ± 0.27	0.15 ± 0.03
HEPES	0.50 ± 0.00	3.99 ± 0.39	-11.06 ± 0.14	-2.09 ± 0.15	0.25 ± 0.02
phosphate	0.48 ± 0.00	14.8 ± 3.74	-12.31 ± 0.18	-2.56 ± 0.23	0.06 ± 0.01
TRIS	0.49 ± 0.00	9.13 ± 2.12	-9.91 ± 0.27	-0.45 ± 0.21	0.10 ± 0.02

Fig. A.5 Representative ITC titrations for compound **11** in HEPES (upper row), phosphate and TRIS buffer (bottom row).

Table A.6 PHGDH raw ITC data for compound **12**

Buffer	N	K_a (μM^{-1})	ΔH (kcal/mol)	$T\Delta S$ (kcal/mol)	K_d (μM)
HEPES	0.35 ± 0.00	3.06 ± 0.51	-8.58 ± 0.19	0.23 ± 0.21	0.32 ± 0.05
HEPES	0.42 ± 0.00	5.63 ± 0.78	-10.17 ± 0.17	-0.99 ± 0.18	0.17 ± 0.02
phosphate	0.66 ± 0.00	2.82 ± 0.49	-10.02 ± 0.49	-1.25 ± 0.50	0.35 ± 0.04
TRIS	0.61 ± 0.00	5.01 ± 1.07	-8.55 ± 0.16	0.55 ± 0.20	0.19 ± 0.04
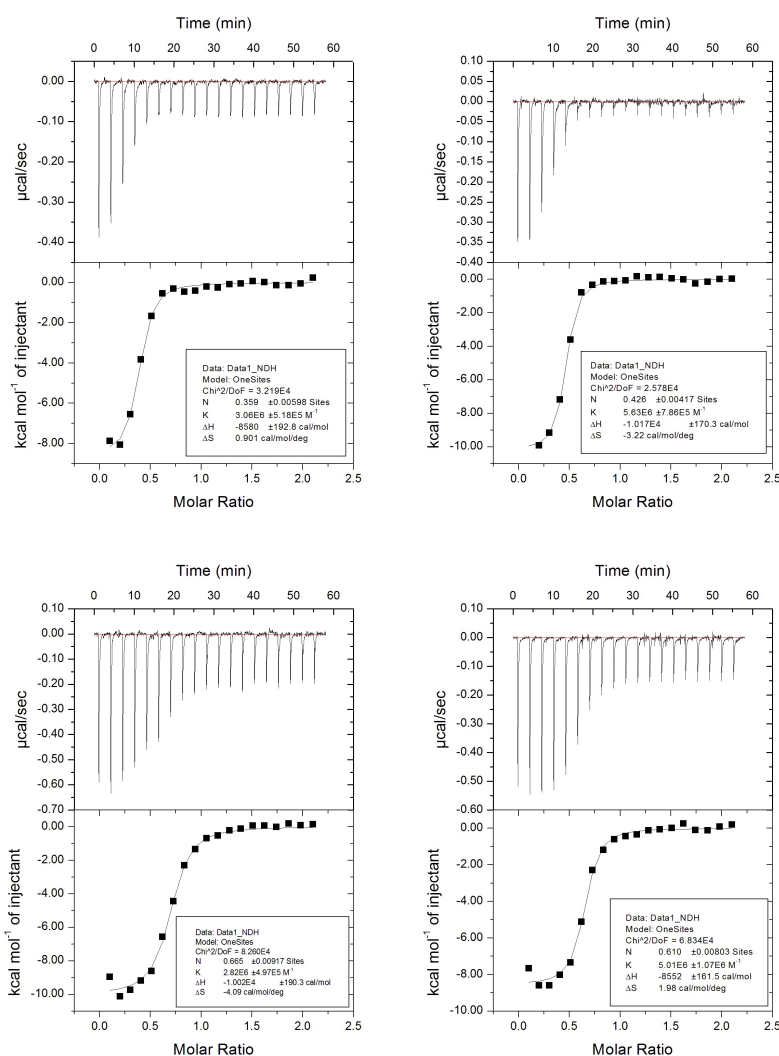
Fig. A.6 Representative ITC titration for compound **12** in HEPES (upper row), phosphate and TRIS buffer (bottom row).

The observed enthalpies were plotted against the buffer ionisation enthalpies (Figure A.7). The enthalpies were essentially the same in two of the buffer systems (HEPES and TRIS), which have significantly different heats of ionisation. There were some differences in enthalpies measured in phosphate buffer: however, superimposing **5**, **8**, **10-12** to a cofactor bound structure (PDB code 2G76) revealed that the phosphate moiety of NAD and inhibitors occupy the same space, suggesting that inhibitors and phosphate are competing for binding (Figure 7.5).
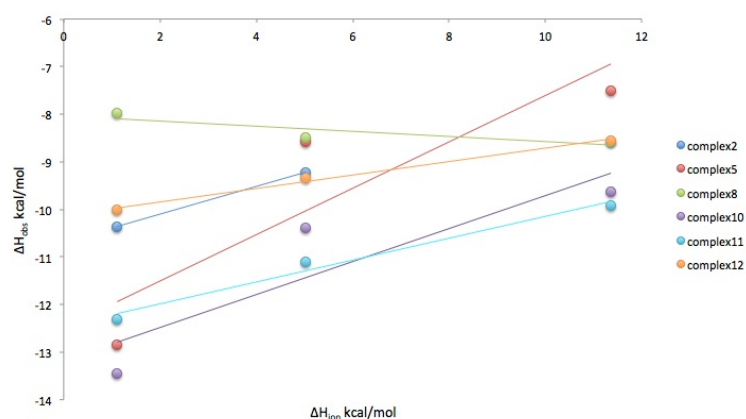

Fig. A.7 ΔH_{obs} plotted against ΔH_{ion}. The data measure in HEPS buffer are geometric means ± SE.

Values measured in phosphate buffer were excluded from further analysis (Figure A.8). The observed enthalpies from the titration experiments were then plotted against buffer ionisation enthalpies of the two buffers.
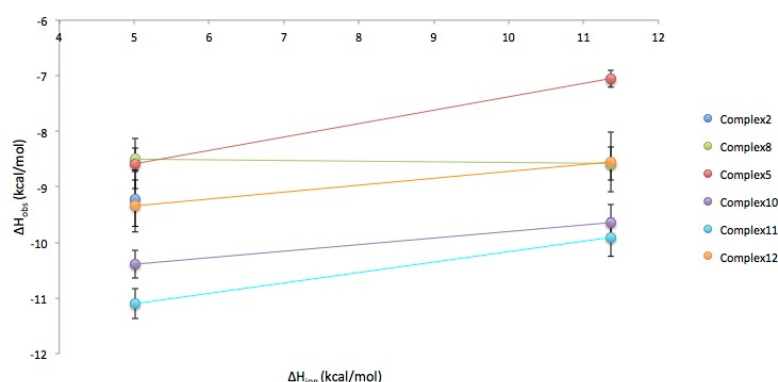

Fig. A.8 The observed enthalpies were essentially the same in HEPES and TRIS buffer. The slope of the line does not indicate proton release or uptake.

The slope of the line (which gives the number of protons released by the buffers) does not indicate proton release or uptake. There is some ambiguity when looking at compound

5. Data suggest other members of the series do not undergo proton exchange, hence, it is reasonable to infer that compound **5** behaves the same.

HEPES and TRIS buffer were therefore treated as an equivalents, and geometric mean from these independent experiments was taken to give values reported in the main text (Table 7.3).

A.2   Representative SPR sensorgrams for PHGDH

Table A.16 PHGDH raw data for compound **4**

measurement number	K_d (µM)	$k_{off}(s^{-1}) \times 10^{-4}$
1	8.43 ± 1.2	0.0225 ± 1.2
2	6.71 ± 0.48	0.0305 ± 1.8
3	NC	0.0206 ± 0.9
4	NC	0.0248 ± 1.6
5	NC	0.0238 ± 0.6

* NC=not calculated

Table A.17 PHGDH raw data for compound **6**

measurement number	K_d (µM)	$k_{off}(s^{-1}) \times 10^{-4}$
1	3.74 ± 0.95	NC
2	3.84 ± 2.5	NC
3	5.77 ± 5.8	NC
4	2.99 ± 2.8	NC
5	0.93 ± 0.59	NC

* NC=not calculated

Fig. A.9 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **1** (dissociation part of the curve) K_d=25.31 $\pm$ 11 μM, k_{off}=0.0533 $\pm$ 2.1 $\times$ 10^{-4} s^{-1}.

Table A.7 PHGDH raw data for compound **1**

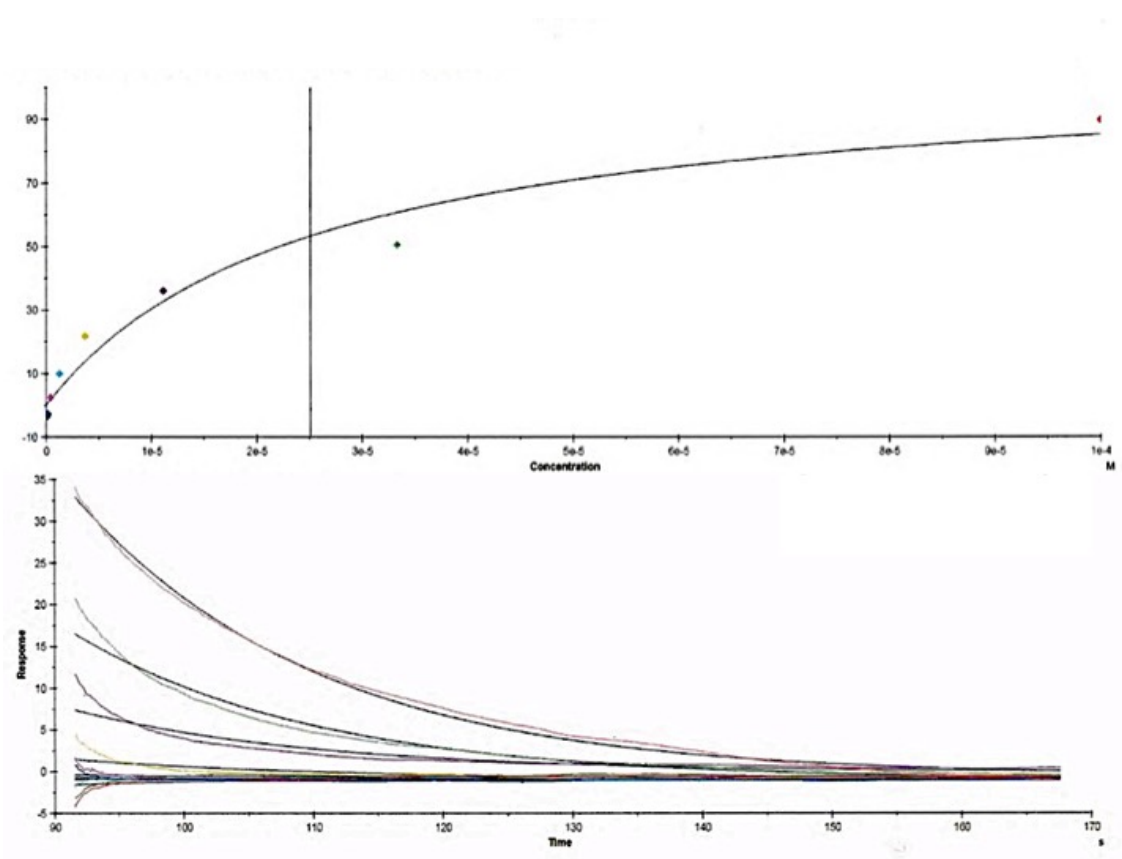| measurement number	K_d (μM)	$k_{off}(s^{-1}) \times 10^{-4}$
1	28.87 ± 12	0.0229 ± 1.0
2	23.02 ± 8.6	0.0234 ± 0.9
3	25.31 ± 11	0.0533 ± 2.1
4	25.09 ± 9.7	0.0264 ± 1.3
5	44.2 ± 18	0.0178 ± 0.9

Fig. A.10 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **2** (dissociation part of the curve) $K_d=0.767 \pm 0.29$ μM, $k_{off}=0.0476 \pm 1.8 \times 10^{-4}$ s^{-1}.

Table A.8 PHGDH raw data for **2**

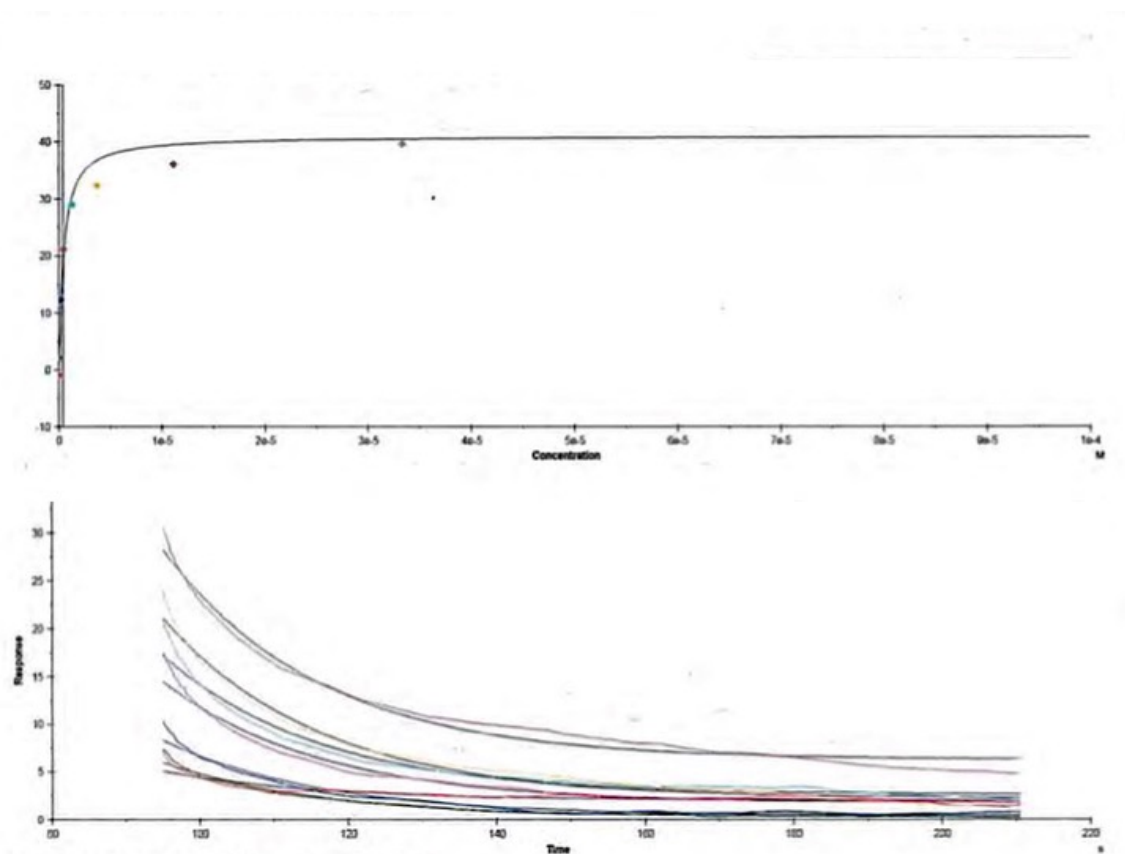| measurement number	K_d (μM)	$k_{off}(s^{-1}) \times 10^{-4}$
1	1.44 ± 0.99	0.0289 ± 1.8
2	0.84 ± 0.53	0.0302 ± 1.5
3	0.83 ± 0.32	0.0336 ± 1.0
4	1.05 ± 0.36	0.0344 ± 1.1
5	0.76 ± 0.29	0.0476 ± 1.8
6	0.41 ± 0.18	0.0521 ± 1.5

Fig. A.11 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **3** (dissociation part of the curve) K_d=0.240 $\pm$ 0.009 μM, k_{off}=0.0478 $\pm$ 1.3 $\times$ 10^{-4} s^{-1}.

Table A.9 PHGDH raw data for compound **3**

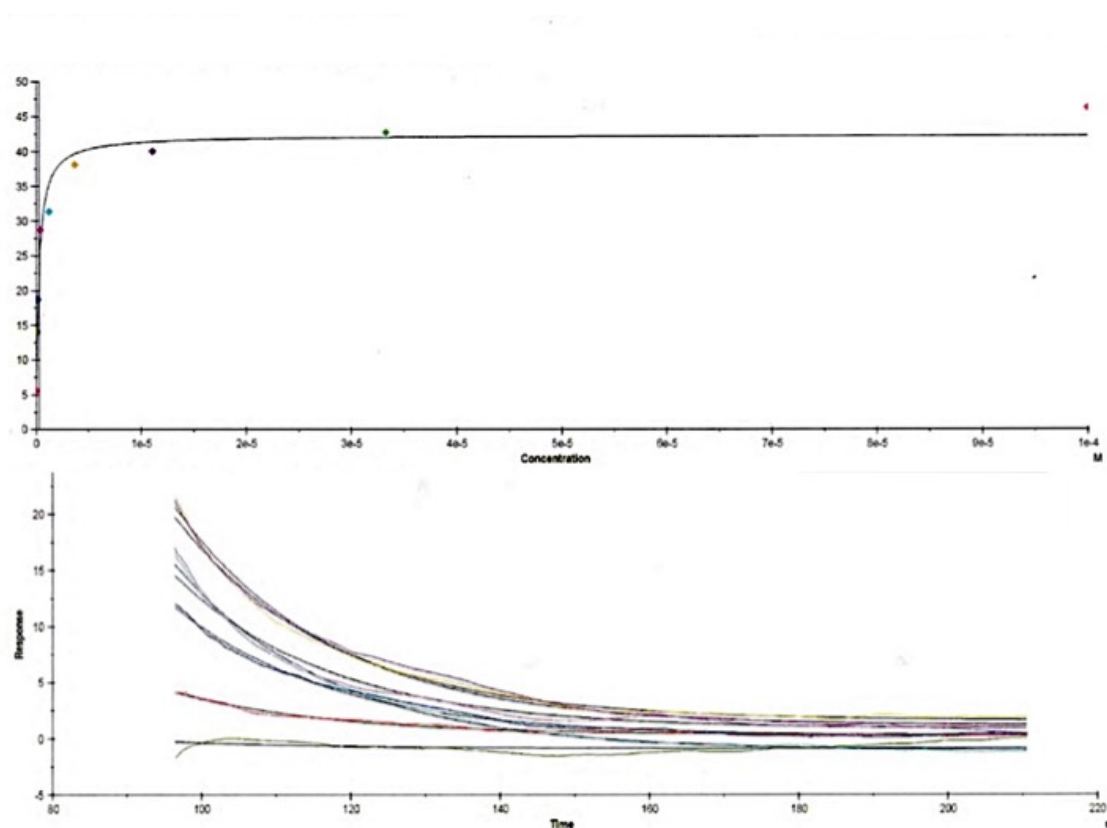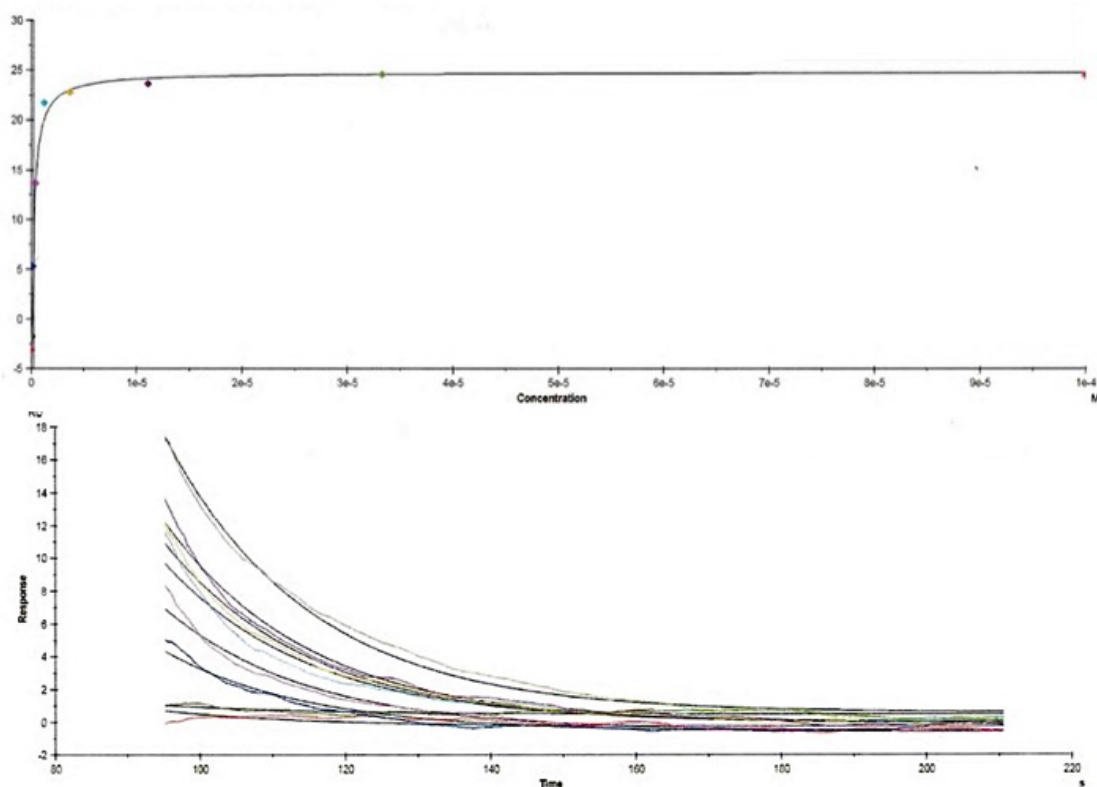| measurement number	K_d (μM)	$k_{off}(s^{-1}) \times 10^{-4}$
1	0.15 $\pm$ 0.07	0.0367 $\pm$ 1.4
2	0.62 $\pm$ 0.12	0.0373 $\pm$ 1.2
3	0.34 $\pm$ 0.09	0.0375 $\pm$ 0.7
4	0.30 $\pm$ 0.11	0.0478 $\pm$ 1.3
5	0.29 $\pm$ 0.04	0.0325 $\pm$ 0.9
6	0.89 $\pm$ 0.68	0.0361 $\pm$ 1.2
7	0.30 $\pm$ 0.06	0.0290 $\pm$ 1.6
8	0.23 $\pm$ 0.08	NC
9	0.01 $\pm$ 0.01	NC

[*] NC=not calculated

Fig. A.12 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **5** (dissociation part of the curve) $K_d = 0.223 \pm 0.03$ μM, $k_{off} = 0.0502 \pm 1.4 \times 10^{-4}$ s^{-1}.

Table A.10 PHGDH raw data for compound **5**

measurement number	K_d (μM)	$k_{off}(s^{-1}) \times 10^{-4}$
1	2.73 ± 2.0	0.0320 ± 1.4
2	2.09 ± 1.5	0.0502 ± 1.2
3	0.40 ± 0.13	0.0604 ± 2.6
4	0.22 ± 0.03	0.0626 ± 2.0
5	0.97 ± 0.19	0.0496 ± 0.9
6	0.62 ± 0.19	NC
7	0.95 ± 0.06	NC

[*] NC=not calculated

Fig. A.13 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **7** (dissociation part of the curve) K_d=7.660 $\pm$ 1.0 μM, k_{off}=0.0426 $\pm$ 2.3 $\times$ 10^{-4} s^{-1}.

Table A.11 PHGDH raw data for compound **7**

measurement number	K_d (μM)	k_{off}(s^{-1})$\times$ 10^{-4}
1	4.44 ± 1.8	0.0335 ± 3.1
2	4.67 ± 3.2	0.0201 ± 2.4
3	5.02 ± 7.2	0.0426 ± 2.3
4	4.29 ± 1.4	0.0287 ± 2.6
5	7.66 ± 1.0	0.0301 ± 1.3
6	5.45 ± 4.4	NC
7	4.39 ± 3.4	NC
8	5.82 ± 4.1	NC
9	3.14 ± 2.4	NC
10	4.59 ± 1.0	NC
11	2.78 ± 0.8	NC

* NC=not calculated
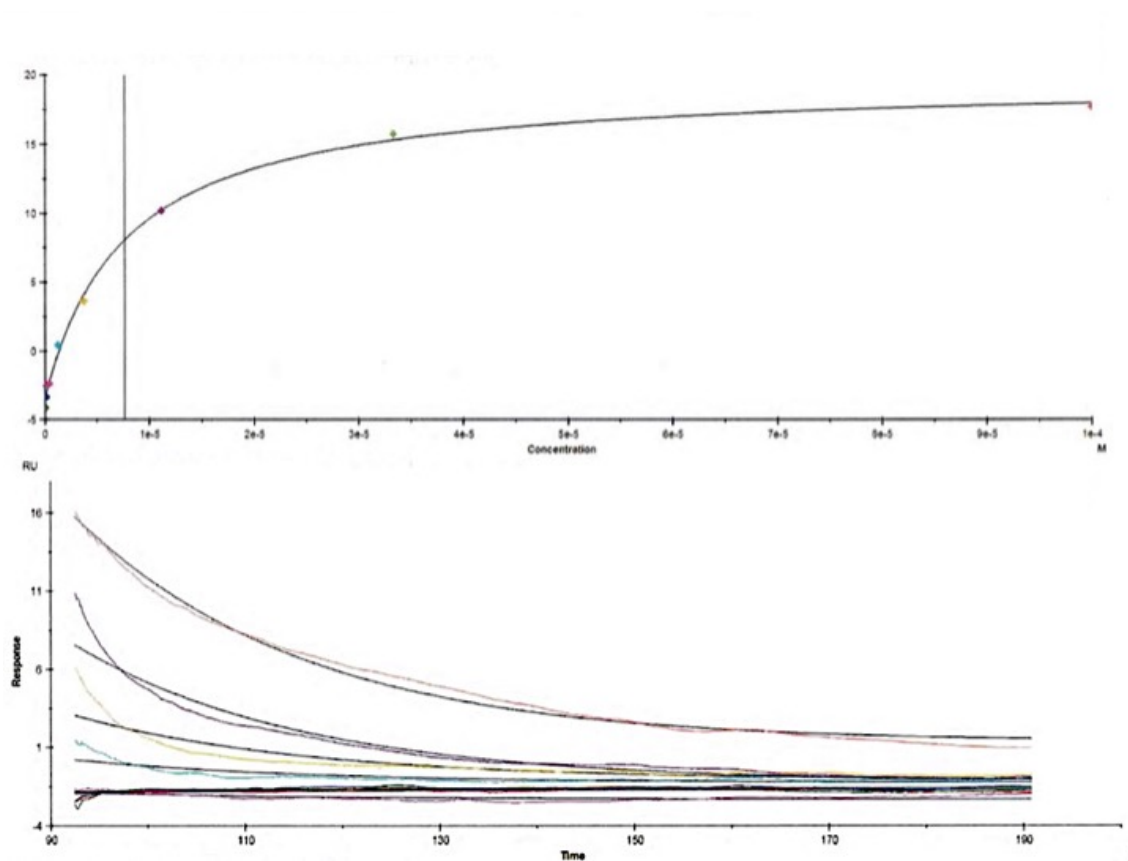
Fig. A.14 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **8** (dissociation part of the curve) $K_d = 0.7811 \pm 0.033$ μM, $k_{off} = 0.0993 \pm 2.4 \times 10^{-4}$ s^{-1}.

Table A.12 PHGDH raw data for compound **8**

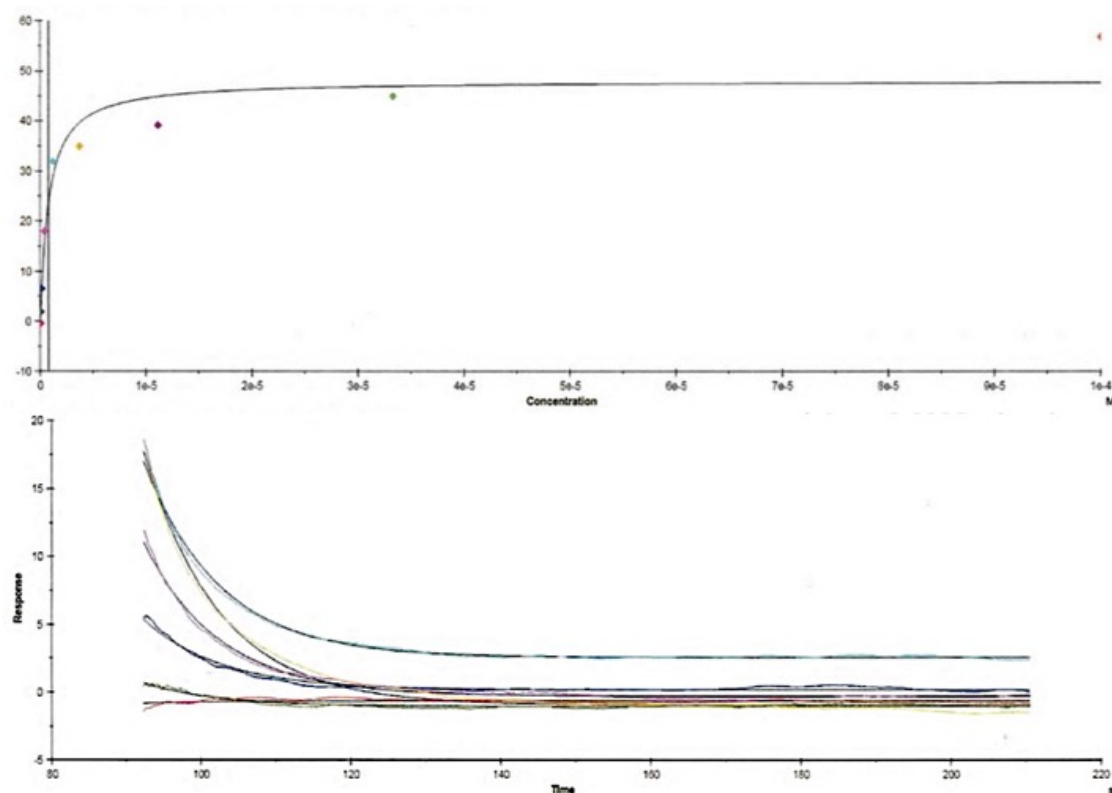| measurement number	K_d (μM)	k_{off}(s^{-1})$\times 10^{-4}$
1	0.860 ± 0.28	0.4574 ± 3.6
2	0.953 ± 0.44	0.0480 ± 3.0
3	0.505 ± 0.23	0.0332 ± 1.9
4	0.607 ± 0.23	0.0324 ± 2.2
5	0.781 ± 0.33	0.0629 ± 2.5
6	0.643 ± 0.26	0.0993 ± 2.4
7	0.677 ± 0.14	0.0472 ± 1.7
8	0.124 ± 0.13	0.0420 ± 2.2
9	0.155 ± 0.17	0.0439 ± 3.5

Fig. A.15 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **10** (dissociation part of the curve) K_d=0.678 $\pm$ 0.49 μM, k_{off}=0.0405 $\pm$ 8.3 $\times$ 10^{-5} s^{-1}.

Table A.13 PHGDH raw data for compound **10**

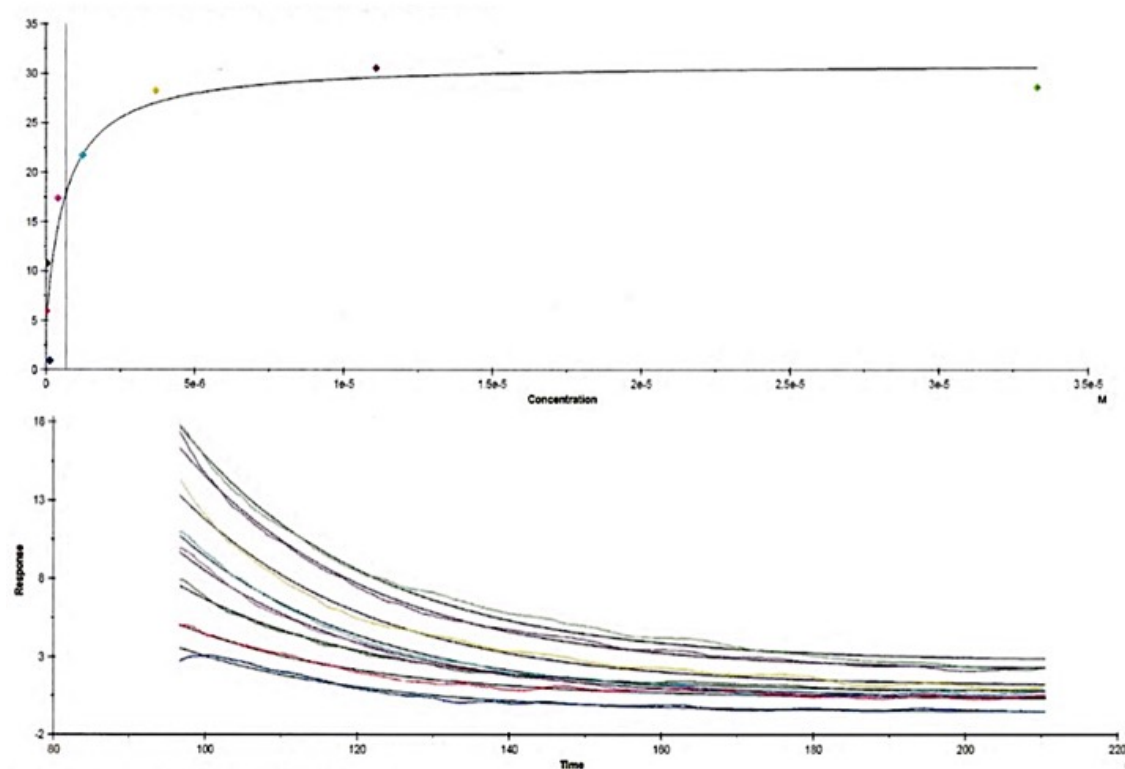| measurement number	K_d (μM)	$k_{off}(s^{-1})\times 10^{-4}$
1	0.435 ± 0.19	0.0378 ± 0.4
2	0.285 ± 0.13	0.0448 ± 0.8
3	0.678 ± 0.49	0.0405 ± 0.8
4	0.162 ± 0.09	0.0389 ± 2.2
5	0.352 ± 0.11	0.0243 ± 1.6
6	0.207 ± 0.04	NC

[*] NC=not calculated

Fig. A.16 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **11** (dissociation part of the curve) K_d=0.398 $\pm$ 0.39 µM, k_{off}=0.0416 $\pm$ 1.0 $\times$ 10^{-4} s^{-1}.

Table A.14 PHGDH raw data for compound **11**

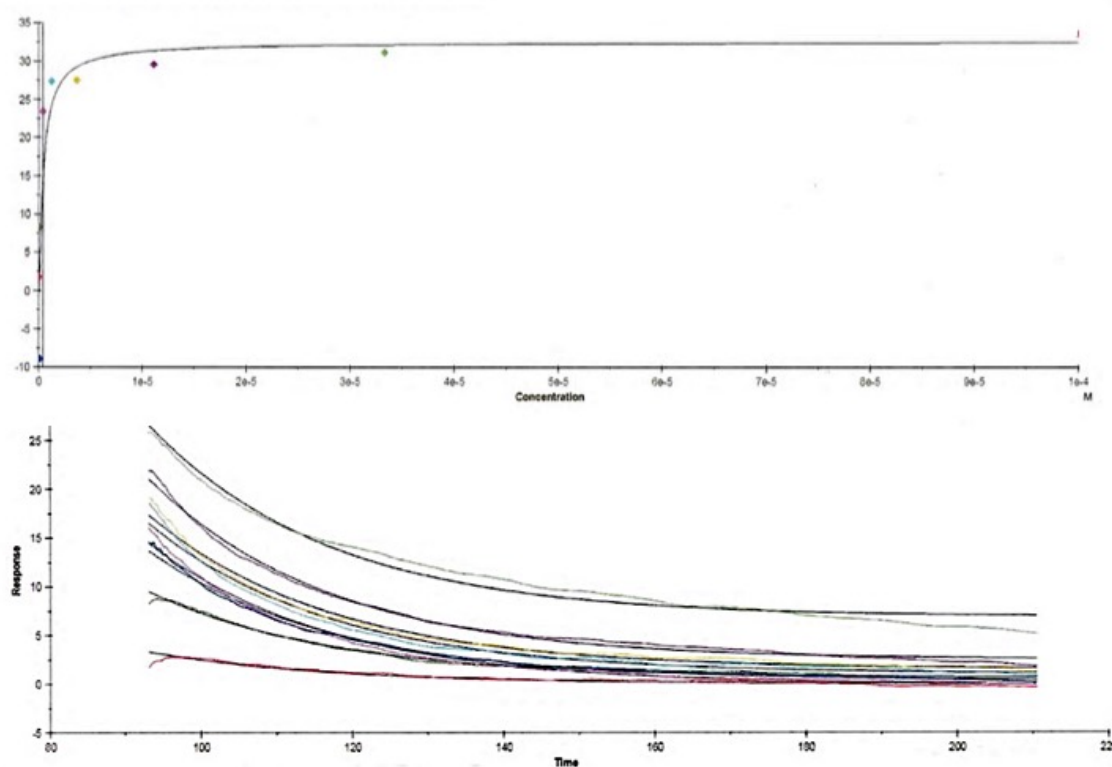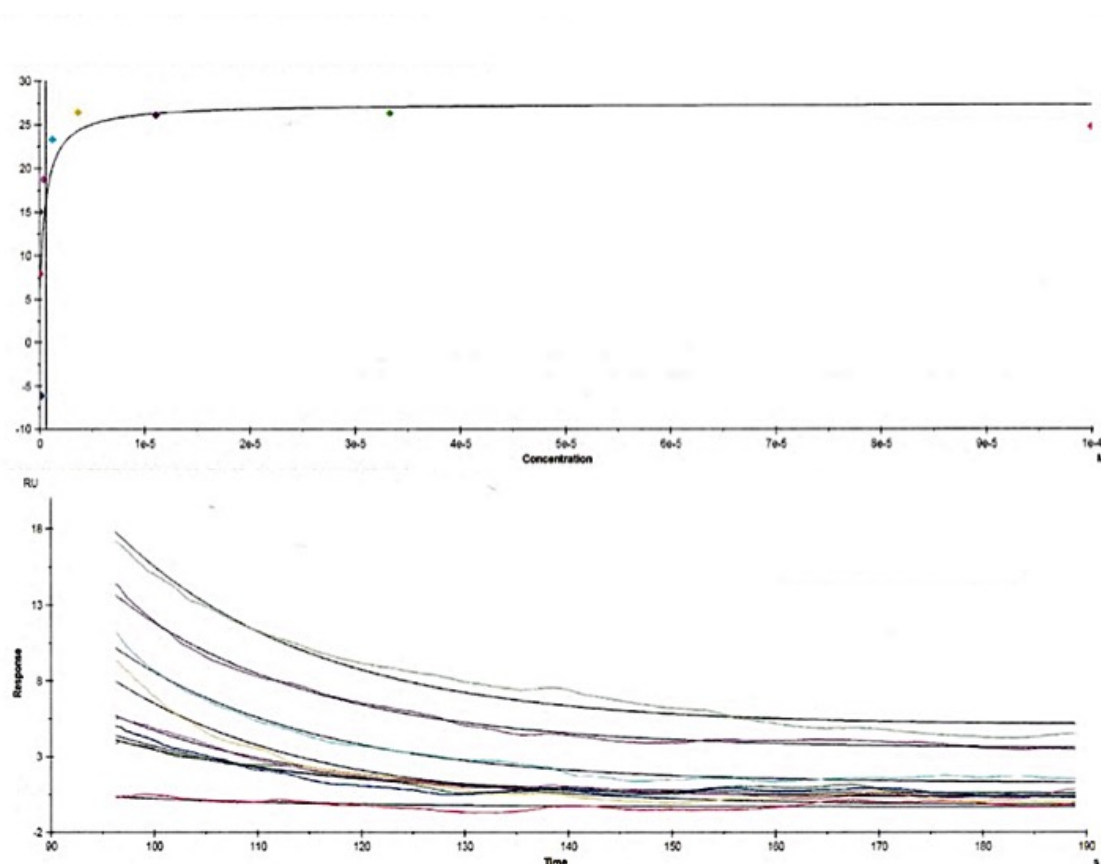| measurement number	K_d (µM)	$k_{off}(s^{-1})\times 10^{-4}$
1	0.286 $\pm$ 0.2	0.0333 $\pm$ 0.7
2	0.620 $\pm$ 0.24	0.0360 $\pm$ 0.6
3	0.079 $\pm$ 0.02	0.0416 $\pm$ 1.2
4	0.466 $\pm$ 0.37	0.0355 $\pm$ 1.2
5	0.855 $\pm$ 0.66	0.0299 $\pm$ 1.6
6	0.398 $\pm$ 0.39	0.0308 $\pm$ 2.7
7	0.271 $\pm$ 0.01	0.0315 $\pm$ 2.0
8	0.950 $\pm$ 0.39	NC
9	0.823 $\pm$ 0.38	NC
10	0.258 $\pm$ 0.03	NC
11	0.424 $\pm$ 0.24	NC

[*] NC=not calculated

Fig. A.17 Top: Representative equilibrium analysis. Bottom: representative sensorgram for compound **12** (dissociation part of the curve) K_d=0.628 $\pm$ 0.9 μM, k_{off}=0.052 $\pm$ 2.5 $\times$ 10^{-4} s^{-1}.

Table A.15 PHGDH raw data for compound **12**

measurement number	K_d (μM)	k_{off}(s^{-1})$\times$ 10^{-4}
1	0.628 $\pm$ 0.9	0.0483 $\pm$ 1.5
2	0.720 $\pm$ 0.77	0.0377 $\pm$ 1.4
3	0.197 $\pm$ 0.03	0.0393 $\pm$ 0.9
4	0.160 $\pm$ 0.09	0.0525 $\pm$ 2.5
5	0.176 $\pm$ NC	

* NC=not calculated

Appendix B

B.1 LDHA single point energies

Table B.1 LDHA theoceptor complexation energies with different functionals. Energies are in units kcal/mol.

	M06	M06-L	M06-2X	B3LYP
1	-23.2	-25.2	-25.8	-1.7
2	-13.6	-16.2	-13.1	2.7
3	-13.8	-15.9	-12.5	5.8
4	-15.0	-2.7	0.8	28.6
5	-17.7	-19.6	-19.4	6.6
6	-10.3	-11.2	-11.6	-1.0
7	-24.9	-24.4	-24.0	-3.9
8	-15.6	-17.9	-18.2	1.0
9	-18.9	-20.7	-18.7	-20.9
10	-19.7	-22.7	-20.9	-0.7
11	-10.4	-13.1	-12.5	2.9
malonate	-39.5	57.7	-50.3	468.3
oxamate	-62.6	27.4	-31.4	577.5

B.2 Lysozyme single point energies

Table B.2 The calculated complexation energies (B3LYP/6-31+G*), relative to the benzene closed conformation. Energies are in units kcal/mol.

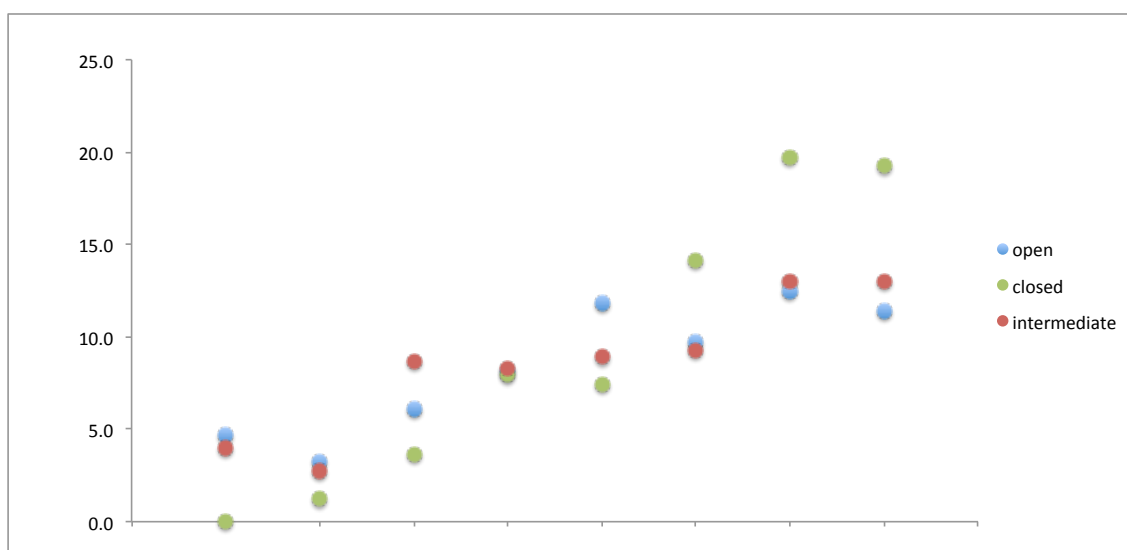|	closed	intermediate	open
benzene	0.0	4	4.7
toluene	1.3	2.7	3.3
n-ethylbenzene	3.6	8.7	6.1
n-propylbenzene	7.9	8.3	8.1
sec-butylbenzene	7.4	8.9	11.8
n-butylbenzene	14.1	9.3	9.7
n-pentylbenzene	19.7	13	12.5
n-hexylbenzene	19.2	13	11.4

Fig. B.1 Relative complexation energies (B3LYP/6-31+G*).

Table B.3 The calculated complexation energies (M06-L/6-31+G*), relative to the benzene closed conformation. Energies are in units kcal/mol.

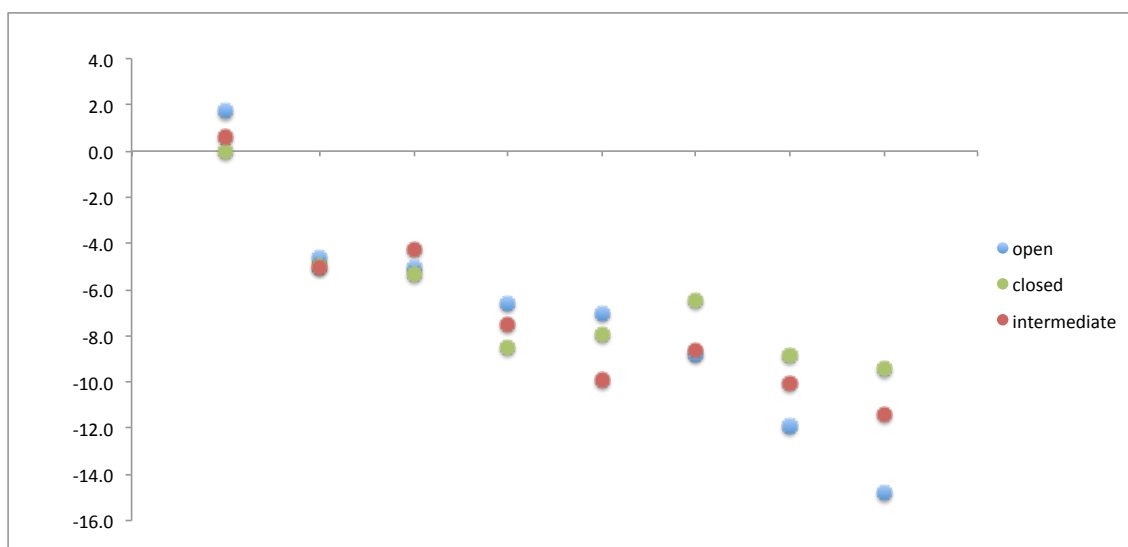|	closed	intermediate	open
benzene	0.0	0.6	1.8
toluene	-5.0	-5.0	-4.6
n-ethylbenzene	-5.3	-4.3	-5.1
n-propylbenzene	-8.5	-7.5	-6.6
sec-butylbenzene	-7.9	-9.9	-7.0
n-butylbenzene	-6.5	-8.6	-8.8
n-pentylbenzene	-8.9	-10.1	-11.9
n-hexylbenzene	-9.4	-11.4	-14.8

Fig. B.2 Relative complexation energies (M06-L/6-31+G*).

Table B.4 The calculated complexation energies (M06-2X/6-31+G*), relative to the benzene closed conformation. Energies are in units kcal/mol.

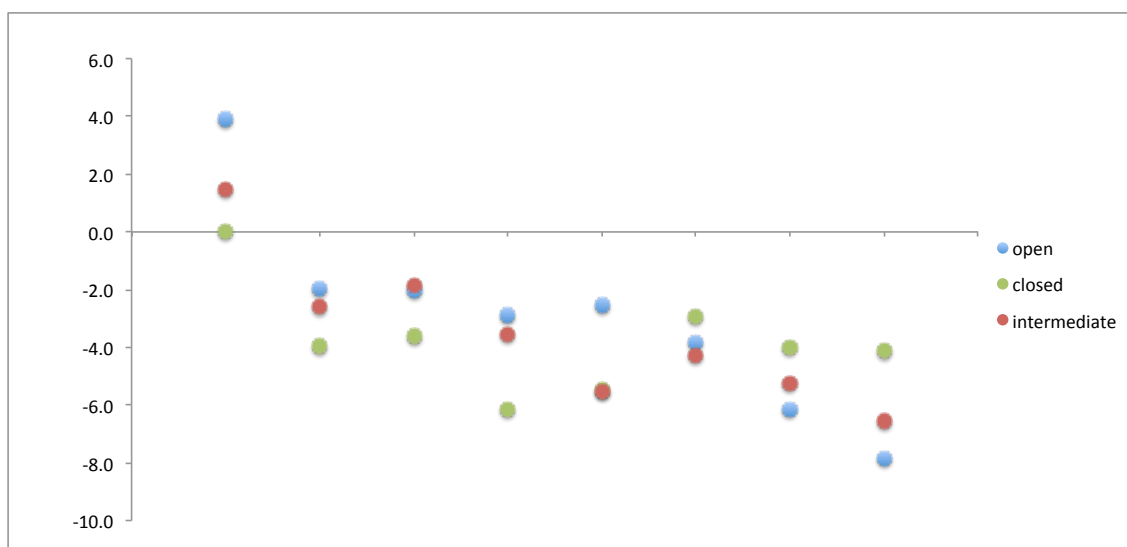|	closed	intermediate	open
benzene	0.0	1.5	3.9
toluene	-3.9	-5.6	-2.0
n-ethylbenzene	-3.6	-1.9	-2.0
n-propylbenzene	-6.2	-3.6	-2.9
sec-butylbenzene	-5.5	-5.5	-2.6
n-butylbenzene	-2.9	-4.3	-3.8
n-pentylbenzene	-4.0	-5.2	-6.1
n-hexylbenzene	-4.1	-6.5	-7.9

Fig. B.3 Relative complexation energies (M06-2X/6-31+G*).