An in-silico study: Investigating small molecule modulators of bio-molecular interactions

Girish Tampi

*Submitted in accordance with the requirements for
the degree of Doctor of Philosophy*

The University of Leeds
Astbury Centre for Structural Molecular Biology
December 2015

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

Zhuangzi and Huizi were strolling along the dam of the Hao Waterfall when

Zhuangzi said, “See how the minnows come out and dart around where they please! That’s what fish really enjoy!”

Huizi said, “You’re not a fish – how do you know what fish enjoy?”

Zhuangzi said, “You’re not me, so how do you know I don’t know what fish enjoy?”

Huizi said, “I’m not you, so I certainly don’t know what you know. On the other hand, you’re certainly not a fish – so that still proves you don’t know what fish enjoy!”

Zhuangzi said, “Let’s go back to your original question, please. You asked me how I know what fish enjoy – so you already knew I knew it when you asked the question. I know it by standing here beside the Hao.”

- Zhuang Zhou

Acknowledgments

I would like to thank Gary Thompson and Arnout Kalverda, whose help has been invaluable over the years. It was mainly through talking with them about possible avenues of research, and their perseverance in explaining difficult concepts that allowed this PhD to take shape.

I'm very fortunate that the gentleman who started his PhD a year before me was Richard Malham, as I have really benefited from the time he took to explain things in the thorough manner that only Richard can do. A lot of thanks also goes to: Sue Matthews for helping me through the experimental work I did in the first year and making life in the lab hugely enjoyable. Theo Karamonos for pretty much always being available to offer advice and assist me with lab work. Andrew Nightingale for his assistance with my initial forays into programming. Joel Dockray for interesting conversations and consultations about programming. Robin Richardson for the imagination with which we could discuss the craziest of ideas and his practical assistance with python scripting. Simon Patching for his help with optimising the MUP purification procedure.

Special thanks go to Emmanuelle Paci who dragged me into GPU computing despite my initial reluctance. I know I would have been the poorer for it if I had not started delving into Linux system administration, never mind the sampling hit my simulations would have taken. I'd also like to thank Seishi Shimizu and Jane Thomas-Oates from York, who have always supported me and taken an interest in my work even after I finished my undergraduate degree. Colin Fishwick for his advice on charge derivation. Nicola Stonehouse for her excellent advice as to how to deal with some of the problems I faced.

I would like to express my gratitude to the HPC team at Leeds, Alan Real, Dureid El-Moghraby and Mark Dixon for helping me with my questions on how the cluster worked and their assistance with the various software issues I had. Thanks also goes to Chris McLean, who sorted out my computing needs in physics and made sure that my mainstay computer ran smoothly. Many thanks also go to Aidan Richmond and the other guys in ISS, who looked after the computing facilities in FBS.

I'd also like to thank my supervisor Steve Homans and for introducing me to molecular dynamics and giving me my first lessons on the dark art of running simulations and for creating a great group with which to work with. I would also like to thank the BBSRC, and Johnson & Johnson for sponsoring this research.

Most importantly, I'd most like to thank my parents and extended family for their love, care and support over the years.

Credits

Contributions to this thesis include work by Gary Thompson who modified Chris McRaild's Order program from the box under the stairs and extensively reprogrammed it so that it could deal with modern day advancements, and added additional features that extended its capabilities. Additionally, he also wrote a script that very effectively depicts polar coordinates as a hammer projection.

I am very grateful for the python script contributed by Robin Richardson that allow the conversion of Cartesian to polar coordinates and the script that write volumetric files for the purposes of visualising densities. Many thanks also go to Irene Bassano for her help and advice with formatting this document. The page design was greatly assisted by the efforts of Theo Karamanos and Hannah Stacey. Additional thanks goes to Martin McKenna for kindly providing permission to modify and use the image of the Daemon used in the "Maxwell's Demon" graphic [501].

I'd also like to thank Gary Thompson, Arnout Kalverda, Theo Karamanos, Joel Dockray and Adam Gowland for proof reading my thesis. Additional thanks go to Steve Homans, Seishi Shimizu, Emanuele Paci and David Westhead for their comments.

Abstract

Small molecule inhibitors are commonly used to target protein targets that assist in the spread of diseases such as AIDS, cancer and deadly forms of influenza. Despite drug companies spending millions on R&D, the number of drugs that pass clinical trials is limited due to difficulties in engineering optimal non-covalent interactions. As many protein targets have the ability to rapidly evolve resistance, there is an urgent need for methods that rapidly identify effective new compounds.

The thermodynamic driving force behind most biochemical reactions is known as the Gibbs free energy and it contains opposing dynamic and structural components that are known as the entropy ($\Delta S°$) and enthalpy ($\Delta H°$) respectively. $\Delta G° = \Delta H° - T\Delta S°$. Traditionally, drug design focussed on complementing the shape of an inhibitor to the binding cavity to optimise $\Delta G°$ favourability. However, this approach neglects the entropic contribution and phenomena such as Entropy-Enthalpy Compensation (EEC) often result in favourable bonding interactions not improving $\Delta G°$, due to entropic unfavorability. Similarly, attempts to optimise inhibitor entropy can also have unpredictable results. Experimental methods such as ITC report on global thermodynamics, but have difficulties identifying the underlying molecular rationale for measured values. However, computational techniques do not suffer from the same limitations.

MUP-I can promiscuously bind panels of hydrophobic ligands that possess incremental structural differences. Thus, small perturbations to the system can be studied through various *in silico* approaches. This work analyses the trends exhibited across these panels by examining the dynamic component *via* the calculation of per-unit entropies of protein, ligand and solvent. Two new methods were developed to assess the translational and rotational contributions to $T\Delta S°$, and a protocol created to study ligand internalisation. Synthesising this information with structural data obtained from spatial data on the binding cavity, intermolecular contacts and H-bond analysis allowed detailed molecular rationale for the global thermodynamic signatures to be derived.

Abbreviations

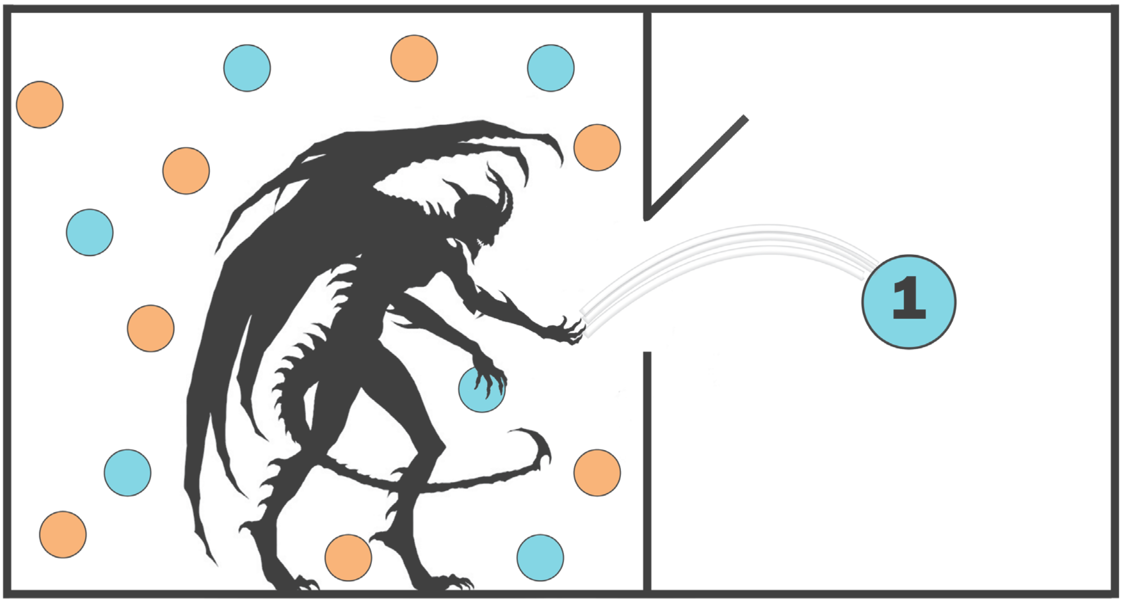
3c6	Hex-3-en-1-ol
3c7	Hept-3-en-1-ol
3c8	Oct-3-en-1-ol
3c9	Non-3-en-1-ol
3Dh	Three Dimensional Histogramming
ACE	COCH3 blocking group
ADME	Absorption, Distribution, Metabolism and Excretion
ADP	Adenosine Diphosphate
AIDS	Acquired Immunodeficiency Syndrome
AKAP	A-kinase Anchoring Proteins
AMBER	Assisted Model Building with Energy Refinement
AMOEBA	Atomic Multipole Optimised Energetics for Biomolecular Applications
AMP	Amprenavir
ASIC	Application-Specific Integrated Circuit
BAT	Bond, Angle, Torsion coordinates
BCA	Bovine Carbonic Anhydrase
BSA	Buried Surface Area
BUT	4-hydroxy-2-butanone
CAR	Conformation-Activity Relationships
CHARMM	Chemistry at Harvard Macromolecular Mechanics
COM	Centre of Mass
CPMG	Car-Purcell-Meiboom-Gill
CPU	Central processing unit
CSA	Connolly Surface Area
CSP	Chemical Shift Perturbation
CTB	B-subunit of the cholera toxin
CUDA	Compute Unified Device Architecture
DAR	Darunavir
DBH	3,4-dehydro-exo-brevicomin
DFT	Density Functional Theory
DHFR	Dihydrofolate Reductase
DNA	Deoxyribonucleic Acid
DOF	Degrees of freedom
DP	Double Precision
EEC	Entropy-Enthalpy Compensation
ESP	Electrostatic Surface Potential
FD	Free Diffusion

FDM	Finite Difference Method
FEP	Free Energy Pertubation
FLOPS	Float Point Operations Per Second
FM	Flexible Molecule
GAFF	General Amber Force Field
GB	Generalised Born
GBSA	Generalised Born Surface Area
GIST	Grid Inhomogeneous Solvation Theory
GPGPU	General-purpose Computing on Graphics Processing Unit
GPU	Graphics Processing Unit
hep	Heptan-1-ol
hex	Hexan-1-ol
HIV	Human Immunodeficiency Virus
HIV PR	Human Immunodeficiency Virus Protease
HMH	6-hydroxy-6-methyl-3-heptanone
HP	Horizontal Pose
HPC	High Performance Computing
HS	Hypothetical Scanning
HTS	High Throughput Screening
IBMP	2-methoxy-3-isobutylpyrazine
IPMP	2-methoxy-3-isopropylpyrazine
ITC	Isothermal Titration Calorimetry
IT-TI	Independent-Trajectories Thermodynamic-Integration
LIE	Linear Interaction Energy
LJ	Lennard-Jones
MD	Molecular Dynamics
MEP	Molecular Electrostatic Potential
MF	Molecule Frame
MIE	Mutual Information Expansion
MIST	Maximum Information Spanning Tree
MM	Molecular Mechanics
MUP	Major Urinary Protein
NGAL	Neutrophil Gelatinase-Associated Lipocalin
NHA	Non-Hydrogen Atom
NME	NHCH3 blocking group
NMR	Nuclear Magnetic Resonance
non	Nonan-1-ol
oct	Octan-1-ol
PB	Poisson-Boltzmann
PBSA	Poisson-Boltzmann Surface area

PCA	Principle Component Analysis
PCM	Principle Component Mode
PDB	Protein Data Bank
PDF	Probability Distribution Function
PI	Protease Inhibitor
PKA	Protein Kinase A
PL	Protein-Ligand
PMEMD	Particle Mesh Ewald Molecular Dynamics
PMF	Potential of Mean Force
PP	Protein-Protein
PS	Pocket Similarity
QHA	Quasi Harmonic Approximation
QM	Quantum Mechanical
QSAR	Quantitative Structure-Activity Relationship
RAM	Random Access Memory
RCSB	Research Collaboratory for Structural Bioinformatics
RESP	Restrained Electrostatic Potential
RMS	Root Mean Square
RMSD	Root Mean Square Deviation
RR	Rigid Rotor
RRHO	Rigid Rotor Harmonic Oscillator
RRHOA	Rigid Rotor Harmonic Oscillator Approximation
SA	Surface Area
SASA	Solvent Accessible Surface Area
SBT	2-sec-butyl-4,5-dihydrothiazole
SCR	Sequence Conserved Regions
SF	System Frame
SMD	Steered Molecular Dynamics
SP	Single Precision
SPDP	Single Precision, Double Precision (hybrid model)
SPFP	Single Precision, 64 bit Floating Point Integer model" (hybrid model)
ST	Sackur-Tetrode
T2FE	Targeted to Free Exploration
TFLOPS	teraFLOPS
ThP	Hept-6-en-1-ol
ThX	Hex-5-en-1-ol
TI	Thermodynamic Integration
TMD	Targeted Molecular Dynamics
TnO	Non-8-en-1-ol
ToC	Oct-7-en-1-ol

US	Umbrella Sampling
VOC	Volatile Organic Compound
VP	Vertical Pose
WAXS	Wide-angle X-ray Scattering

Contents



Maxwell’s Daemon

Chapter 1.0: Introduction

1.1.0. The birth of rational drug design

Paul Ehrlich hypothesised the existence of chemoreceptors based on the observation that dyes are selectively absorbed by different histological samples. The realisation that the differences between chemoreceptors of healthy cells and those found in parasites, cancers and bacteria could be exploited to effect cures to diseases began the science of drug discovery. Ehrlich examined over 600 compounds before finally discovering Salvarsan, a drug effective against syphilis [1 2].

Drug targets are molecules that modulate enzymes, receptors or other proteins that exert physiological effects on the body. Targets can broadly be divided into seven classes, with receptors falling into one of the largest subsets (**Fig.1.1**) [3].

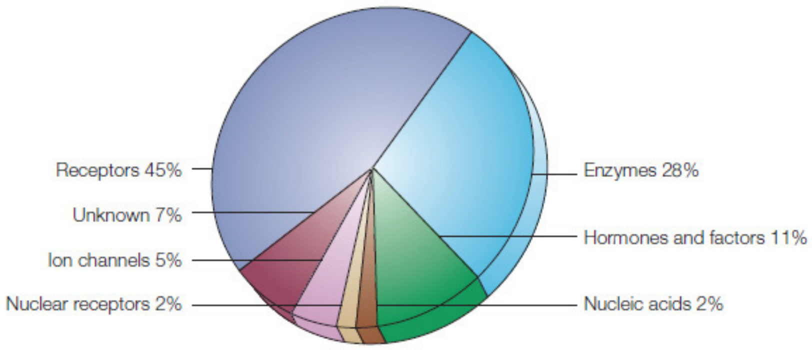


Fig.1.1. Potential therapeutic targets divided into seven different categories. Image taken from reference [4].

Subsequent to the discovery of a compound efficacious against disease, illness or microorganisms, chemists found they could modify the main chemical nucleus by altering functional groups to produce more potent analogues that overcame resistance. e.g. β-lactam antibiotics [5]. Methodologies in drug design have changed throughout the years. A key tenet of drug discovery that has evolved alongside biochemical and structural characterisation of biomolecules is that of rational design. This is the use of biochemical and structural data to engineer ligands (small, non-covalent complex-forming organic molecules) with a complementary shape to their protein targets. This can be done by improving an existing ligand or designing a completely new ligand [6–8]. Detailed, high-resolution structural information of proteins with or without complexed ligands is regularly obtainable via techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. In October, 2013, the RCSB Protein Data Bank (PDB) held the details of 62,634 structures [9].

Drug discovery is currently a multi-disciplinary field that integrates computational approaches with synthetic chemistry, biochemistry and biophysics. Computational methods are increasingly important in the search to effectively match protein targets with appropriate ligands. Once promising ligands are found, they can be optimised and assayed to determine their effectiveness. While complementarity is important for molecular recognition between interacting biomolecules, high-affinity ligand design is influenced by a complex assortment of often competing factors [10,11].

1.1.1. High throughput screening - A brute force approach

Pharmaceutical companies have spent considerable time and money creating large libraries of drug candidates that have potential activity against biological targets. When designing a new pharmacophore, a subset of these compounds is selected and approximately 500,000 compounds a week are rapidly assayed against protein targets using automated technologies such as robotics. This commonly used procedure is known as high throughput screening (HTS). Assays are normally carried out on microtitre plates containing 384 and 1,536 wells at higher capacities. There are many options for assays, including but not limited to enzyme-linked immunosorbent assays, optical sensing, mass spectrometry and reporter based assays. These procedures can be cell based or cell free. A preliminary assay often identifies potential candidates. These "hits" are subjected to more stringent subsequent rounds of testing to identify whether they qualify as a promising lead that is suitable for pre-clinical trials [12,13].

In order to maximise the generation of successful drugs in a cost-effective manner, the libraries must sample as much chemical space as possible. Most drug libraries contain around 10^6 compounds, whilst a moderate estimate of chemical space contains an excess of 10^{60} entities [2 14].

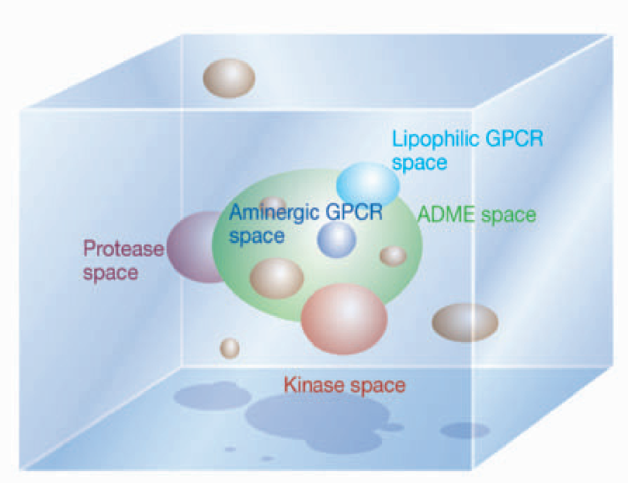


Fig.1.2. A representation of chemical space containing regions occupied by clusters of chemical compounds possessing biological affinity. Examples include proteases (purple), lipohillic GPCRs (blue) and Kinases (red). Those compounds with the shared characteristics of orally administered drugs (i.e. absorption, distribution, metabolism

and excretion) are shown to share space with the region shaded green. Image and information taken from reference [2].

The higher dimensionality of chemical space can be reduced to a map of chemical compounds that populate a 3D space. Chemographing is a technique by which these compounds are positioned relative to each other within this space according to their most significant chemical characteristics. These can be various physiological or topological properties such as size, lipophilicity, structural motifs or other physio-chemical characteristics. If a very accurate map of chemical space could be generated, compounds neighbouring each other should possess high structural similarity and thus have a greater chance of displaying similar biological activities (**Fig.1.2**). The task facing drug designers is to map and ascertain the biologically relevant regions of chemical space [2,15,16].

Early libraries contained many natural compounds but their contents represented a very small region of chemical space. The advent of combinatorial chemistry allowed library expansion via the automated synthesis of large numbers of diverse compounds from chemical "building-blocks". This approach can be problematic as increasing the quantity of possible candidates does not necessarily increase the quality of the library. Compounds were characterised by excessive complexity and a lack of "drug-likeness" (§1.1.2). Despite exponential increases in library size, the discovery of new drugs remained approximately constant [4,15,17].

1.1.2. The rules of 5 - A set of exclusionary criteria

An influential concept pioneered by Lipinski was that of "drug-likeness". In 1996, a study demonstrated that in contrast to the large dimensions of chemical space, the actual target space at that time was limited to ~500 drug targets (**Fig.1.2**). Lipinski argued that the methods adopted by pharmaceutical companies to explore chemical space were grossly inefficient, as the diversity generated by large combinatorial libraries hid the fact that only a smaller subset of those compounds actually had the requisite physio-chemical properties to be effective in humans. As drugs with similar properties are clustered together in the vastness of chemical space, it would be wiser to search target space by filtering out compounds that did not have drug-like properties. This argument is substantiated by a calculation that demonstrated that, if the hypothetical number of drug targets of molecular weight 500 in a human being was 10^{26}, searching a chemical space containing (the lower estimate) of 10^{40} chemical entities generated only a 1 in 10^{14} probability of a hit. This problem is exacerbated by the fact that chemical libraries lack true diversity. They have a surfeit of molecules with agrochemical or pharmaceutical applications but are deficient in truly diverse compounds due to limitations in synthetic techniques [1,14].

In practical terms, Lipinski advocated filtering potential drug candidates based on principles of Absorption, Distribution, Metabolism and Excretion (ADME), versus affinity alone. The rules of 5 is a series of exclusionary criteria based on ADME, whose purpose is to reduce the number (and associated cost) of drug candidates selected for HTS. This rule set for orally administered drugs was derived from examination of a selected subset of the World Drug Index (coined the USAN library) and contains 2,245 drugs. The rules are based on the observation that successful orally administered drugs (i.e. those that had passed Phase I-IV clinical trials) display a distinct subset of physio-chemical properties that are defined by solubility and intestinal permeation. However, compounds that did not enter preclinical or Phase I trials typically did not possess these essential characteristics. Thus, if a compound violates the rules of 5 (summarised below), it is very likely that it will fail. The strength of these rules is their ability to be successfully applied to lead compounds at a very early stage in their development.

1. Compound should have < 5 hydrogen bond donors (expressed as the sum of OHs and NHs). Compounds capable of making too many hydrogen bonds tend to suffer permeability issues as they cannot cross the membrane bilayers easily.

2. Compound molecular weight should be < 500 Daltons. A higher molecular weight is also associated with poor permeability due to difficulties in passing intestinal and blood-brain barriers.

3. The logP is < 5. This measurement is the octanol-water partition coefficient and can be taken as a measure of lipophilicity (**Table.1.1**).

	logP
Optimum **CNS penetration**	~2 +/- 0.7
Optimum **Oral absorption**	~1.8
Optimum **Intestinal absorption**	~1.35
Optimum **Colonic absorption**	~1.32
Optimum **Sub lingual absorption**	~5.5
Optimum **Percutaneous** (& low mw)	~2.6

Table.1.1. Typical logP values. Data taken from reference [18]

4. Compound should have less than 10 H-bond acceptors (calculated from the sum of Ns and Os). Once again, compounds with too many acceptor groups have lower permeability.

5. Actively transported compounds (e.g. antibiotics, antifungals, vitamins and cardiac glycosides) are excluded from the above rules.

Further analysis indicated that violations of any two rules did not exist in excess of 10% of successful compounds. The combination of high molecular weight and logP had the lowest rate (~1%) of passing clinical trials. The author noted that it is this combination that was usually enhanced by HTS [19]. A detailed statistical study validated Lipinski's findings by comparing the physio-chemical property profiles of orally administered drug candidates at various stages in the developmental cycle to that of successfully marketed drugs [20].

1.1.3. The disassociation constant and residence time

Besides consideration of ADME properties, ligand effectiveness is often ranked by binding affinity (K_a). It is also measured in terms of its reciprocal, the disassociation constant (K_d) via the use of Michaelis-Menten or pre-steady state kinetics [21,22].

For the bimolecular reaction in **eqn.1.1**, K_d is represented by **eqn.1.2**, where the values in square brackets denote concentrations of Protein [P], Ligand [L], and Protein-Ligand complex [PL]. Consequently, K_d has Molar units and corresponds to the concentration at which the total population of protein active-sites are half occupied by ligand. Thus, a ligand with a K_d in the nanomolar range has a tighter binding constant than one in the micromolar range.

$$P + L \underset{K_d}{\overset{K_a}{\rightleftharpoons}} PL \qquad \textbf{(eqn.1.1)}$$

$$K_d = \frac{[P][\mathrm{L}]}{[PL]} \qquad \textbf{(eqn.1.2)}$$

K_d can alternatively be expressed as a ratio of ligand disassociation to association rate (**eqn.1.4**).

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL \qquad \textbf{(eqn.1.3)}$$

Where:

$$K_d = \frac{k_{off}}{k_{on}} \qquad \textbf{(eqn.1.4)}$$

Binding affinities (K_a) between proteins and ligands allows broad classification into strong, medium and weak binding categories (**Fig.1.3**). Examples of strong binding complexes (sub-picomolar) include enzyme transition-state complexes formed using mechanisms such as electrophilic, nucleophilic and acid/base catalysis. The medium binding range (micromolar – nanomolar) include antibodies, drug-receptor complexes and enzyme-inhibitor complexes. Lastly, the weak binding category (decimolar to weak micromolar) contains entities such as cyclodextrins, albumins and some weaker enzymes. Typically, initial searches for lead drug-like molecules generate "hits" around the micromolar range and successful optimisation should generate compounds that possess nano to picomolar binding affinities [23,24].

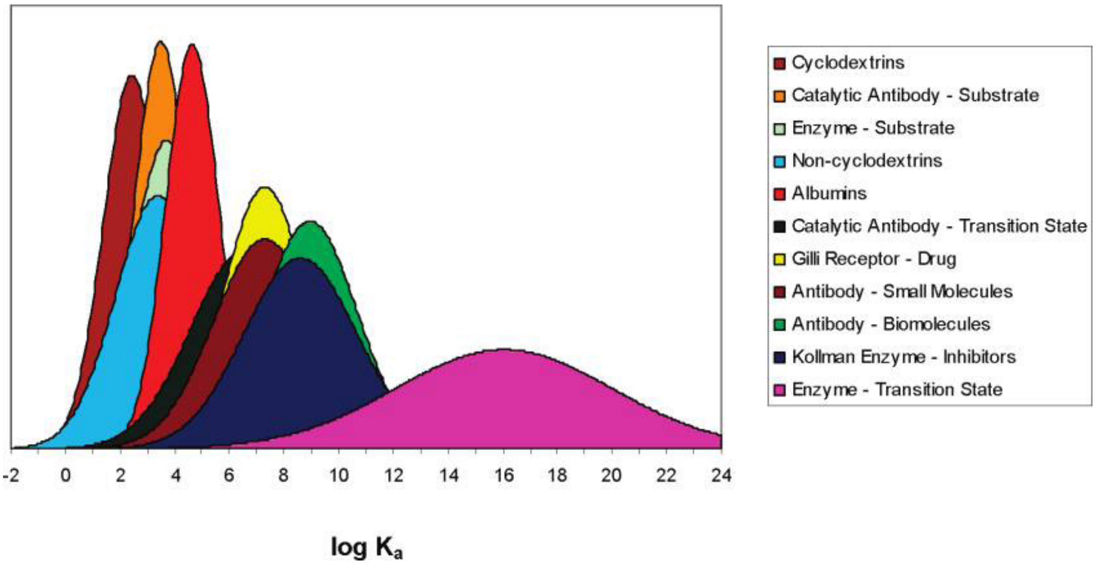


Fig.1.3. Typical Protein-Ligand binding interactions measured in log K_a. Each area is an idealised normalised binding distribution with the maxima denoting the group average. Image and information taken from [23,24].

A ligand attached to a biological target usually exerts its effect for the duration it is bound. This concept is known as residence time and is primarily reflected in the disassociation constant. Enzyme catalysed substrates have a relatively low residence times (50 ms^{-1}), whereas effective inhibitors generally have much longer residence times e.g. Basic pancreatic trypsin inhibitor inhibits trypsin with a residence time of ~6 months [25]. A longer residence time is desirable for inhibitors as it decreases the requirement for sustained high doses and the possibility of resistance. Whilst the formation of strong covalent bonds favour long residence times, non-covalent inhibitors can also be effective. Some of the most stable Protease Inhibitors can achieve residence times > 4 months. Small molecule ligands face a greater challenge compared to larger compounds due to the more limited number of possible interactions with their target receptors. Maintaining a lengthy residence time thus involves a more precarious balancing of positive and negative contributions to the disassociation constant [24,26].

Proteins and receptors encounter each other in two types of system: open and closed. *In vitro* assessment of binding is usually carried out in a closed system, in which the total [P] and [L] is static, only changing with respect to the proportion of bound (PL) and free species. The cell is a good example of an open system, in which [P] and [L] are in flux. In vivo, k_{on} can be influenced by many biochemical parameters such as ligand concentration, metabolic clearance, and the limits imposed by the rate of diffusion. k_{off} is not affected by these factors, although it is more susceptible to considerations pertaining to the binding of protein and ligand. e.g. van der Waals interactions, hydrogen-bonding, protein conformational changes and solvent effects. For example, a study on binding of a variety of protease inhibitors (PIs) to resistant variants of HIV-I protease used data from surface plasmon resonance to compare K_d, k_{on} and k_{off} values. Resistance to protease inhibition correlated with an increase in disassociation rate, coupled with a smaller decrease in association rate. The authors proposed that increases in k_{off} were due to structural changes around the binding pocket reducing inhibitor residence time. Thus, using the K_d as the primary metric for ligand affinity can mask problems that could otherwise be optimised, as compounds can possess similar K_d values whilst having different ratios of component k_{off} and k_{on} terms [26–28].

1.1.4. Gibbs free energy and the interplay of enthalpy and entropy

For a closed system at constant pressure and at equilibrium, K_d can be related to energy change in the system via the Gibbs free energy ($\Delta G°$) as shown in **eqn.1.5**. The term R is the ideal gas constant ($8.314 \, J \, K^{-1} \, mol^{-1}$) and T is the temperature in Kelvin (K). The superscript ° denotes that the measurements are made at standard states that usually correspond to a pressure of 1 atmosphere, 1 M concentration and a temperature of 298 K. Standard states are common reference points that allow comparison of the relative thermodynamic values generated by different reactions via the use of a common experimental environment.

$$\Delta G°_{binding} = -RT \ln(Ka) = RT \ln(Kd) \qquad \textbf{(eqn.1.5)}$$

The more negative the free energy ($\Delta G°$), the more strongly a reaction (e.g. L + P –> PL) will be spontaneous/favourable and move to completion. Hence, it is commonly used in drug scoring systems to evaluate the binding affinity of chemical compounds. $\Delta G°$ can be further broken down into contributions from enthalpy ($\Delta H°$) and entropy ($\Delta S°$) (**eqn.1.6**).

$$\Delta G° = \Delta H° - T\Delta S° \qquad \textbf{(eqn.1.6)}$$

An examination of the relationship between both these component forces reveals that the entropy must be maximised whilst the enthalpy is simultaneously minimised to achieve highly favourable free energies of binding. This has led to attempts to define a master equation that allows the computational determination of $\Delta G°$ by taking into account the two main contributors that compose it.

Enthalpy can be thought of as the amount of heat absorbed or emitted by the system whilst entropy is a measure of the dispersal and spread of energy during that process at a defined temperature. The definition of ΔG° is based on the second law of thermodynamics which states that a spontaneous process occurring within the universe is driven by the overall dispersion and dissipation of a very concentrated source of energy ($\Delta S° > 0$). In a biochemical context, making bonds and non-covalent interactions are exothermic processes that are characterised by the reaction having a negative enthalpic value. The inverse is true for the process of breaking bonds, as an input of energy is required to pull apart molecules. A positive increase in entropy can be due to various solvent effects or because regions of the protein or ligand have greater freedom of motion at the end of a process (**Table.1.2**). Thus, negative enthalpic contributions in conjunction with positive entropic contributions are favourable to a spontaneous reaction occurring at a specific temperature ($\Delta G° < 0$) [21,29,30].

Whilst strong covalent, ionic and metallic bonds are mainly responsible for the chemical bonds (intramolecular interactions) holding a molecule together, there exist a host of weak interactions (intermolecular interactions) that exhibit disparate effects in terms of attractive and repulsive forces. Some of these are tabulated in **Table.1.2**.

Enthalpy	Entropy
van der Waals interactions	Ligand degrees of freedom(rotatable bonds)
Hydrogen bonding	Protein configurational disorder
Electrosatic complementarity	Expulsion of water molecules from protein active site on binding ligand
Salt bridges	Desolvation of hydrophobic groups
	Ligand desolvation

Table.1.2. Typical factors that contribute to enthalpy and entropy values.

The entropy of a simple molecule such as water can be subdivided into contributions from translational, rotational and internal vibrations (**Fig.1.4**). Furthermore, the aqueous milieu within which most biochemical reactions occur also affects the entropic term (§1.1.6). The interplay between entropic and enthalpic forces means that it is very difficult to rationally design drug-like molecules on the basis of structure alone, as any physical modifications have to account for the inevitable impact on dynamics (§1.2.0).

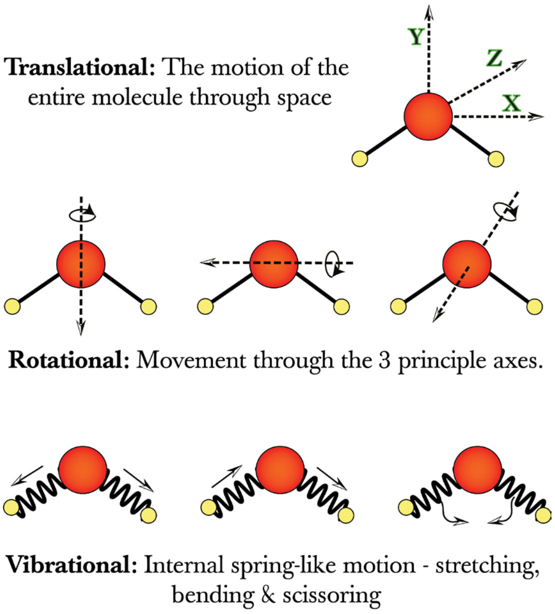

Translational: The motion of the entire molecule through space
Y
Z
X
Rotational: Movement through the 3 principle axes.
Vibrational: Internal spring-like motion - stretching, bending & scissoring

Fig.1.4. The main sources of entropy for a water molecule.

Figure adapted from reference [31]

1.1.5. Decomposition of global thermodynamic terms

This section describes an experimental decomposition method and also serves as a conceptual framework whereby the interplay between entropy and enthalpy can be illustrated. Experimental techniques such as Isothermal Titration Calorimetry (ITC) provide global values for the three key thermodynamic values. This technique directly measures successive changes in ΔH°, as the ligand is titrated in aliquots into a solution of protein receptor. It allows ΔH°, K_d and stoichiometry to be measured in a single experiment. Values for ΔG° and ΔS° are calculated from the relationship in **eqn.1.7** [32]:

$$RT\ \ln(K_d) = \Delta G° = \Delta H° - T\Delta S° = -RT\ \ln(K_a) \qquad \textbf{(eqn.1.7)}$$

While these thermodynamic values give global information about binding reactions, it is more difficult to ascribe rationale to the structural interactions or dynamic components with any specificity, due to the complexity of the system. i.e. The *trees* cannot be seen for the forest. *Per contra*, computational methods such as molecular dynamics (MD) simulations provide atomistic-level models that contain detail about the dynamics and structure of a binding event. However, accurately describing the macroscopic thermodynamics of the system can be a challenge as validating the model and achieving sufficient statistical averaging is not trivial. MD is akin to a single molecule technique that requires simulations of sufficient length, coupled with multiple repeats to substantiate the credibility of any observations. i.e. The *forest* cannot be seen for the trees. The reductionism prevalent in science is applied in both MD and experimental methodology. Experimentalists tend to apply techniques (*vida infra*) that decompose the system to its fundamental parts, whilst MD takes an approach that “builds up” the system from the smallest, tractable component interactions possible (§1.4.0). Though these two approaches are currently (§1.4.3) separated in terms of size and timescale, the synergistic application of the two allows for the validation of theories and a better understanding of systems.

There are three main players in a bimolecular binding reaction that must be taken into account to fully understand the behaviour of the system. These are the protein, ligand and the solvent, each with decomposable enthalpic, entropic and free energy aspects. Capturing the thermodynamic facets of any one of these is a daunting task, made all the more dire by the fact that all of them must be quantified in order to ascertain whether the decomposition is an accurate reflection of the system. The most difficult component to calculate is the solvent thermodynamics due to the sheer number of molecules, their associated degrees of freedom (DOF) and the complexity of their hydrogen bonding interactions. Thus, the experimental decomposition of the global free energy ($\Delta G°_{obs}$) begins by separating contributions into (Protein + Ligand) solute-solute free energy ($\Delta G°_{intrinsic}$), and the solvation contribution using equation **eqn.1.8**.

$$\Delta G°_{obs} = \Delta G°_{i} + \{\Delta G°_{sb} - \Delta G°_{su}\} \qquad \textbf{(eqn.1.8)}$$

The latter is calculated from the difference between the solvation free energy of the bound ($\Delta G°_{sb}$) and free ($\Delta G°_{su}$) species. Construction of a Born-Haber cycle takes advantage of the fact that ΔG° is a state function whose final energetic value is independent of the path taken to reach it (**Fig.1.5**). Thus, construction of an appropriate form allows the elucidation of quantities that are not directly measurable. This treatment can be extended to other thermodynamic terms such as the entropy and enthalpy (**eqn.1.9-10**) [33,34].

$$\Delta H°_{obs} = \Delta H°_{i} + \{\Delta H°_{sb} - \Delta H°_{su}\} \qquad \textbf{(eqn.1.9)}$$

$$\Delta S°_{obs} = \Delta S°_{i} + \{\Delta S°_{sb} - \Delta S°_{su}\} \qquad \textbf{(eqn.1.10)}$$

Air-solvent partition equilibrium experiments are utilised to calculate the solvation term required for this type of decomposition and can only be conducted with ligands that possess sufficient volatility at room temperature [35]. $\Delta G°_{i}$ encompasses the difference in solute-solute free energy between products (protein-ligand complex) to reactants (free ligand and protein) in a solvent free environment. The enthalpic contributions to $\Delta G°_{i}$ arise principally from the formation of new dispersive interactions, salt bridges and hydrogen bonds in the complexed state. Naturally, these interactions must be offset against the non-bonded solute-solvent interactions held prior to binding and is encompassed in the solvation term (§1.1.6). Typically, van der Waals interactions are thought to make a negligible contribution, because the small size of water molecules surrounding the ligand in the free state provides excellent shape complementarity. However, new solute-solute hydrogen bonds have the potential to be much stronger than exchanged solute-solvent bonds if positioned favourably. Unfavourable $\Delta H°_{i}$ contributions can occur if

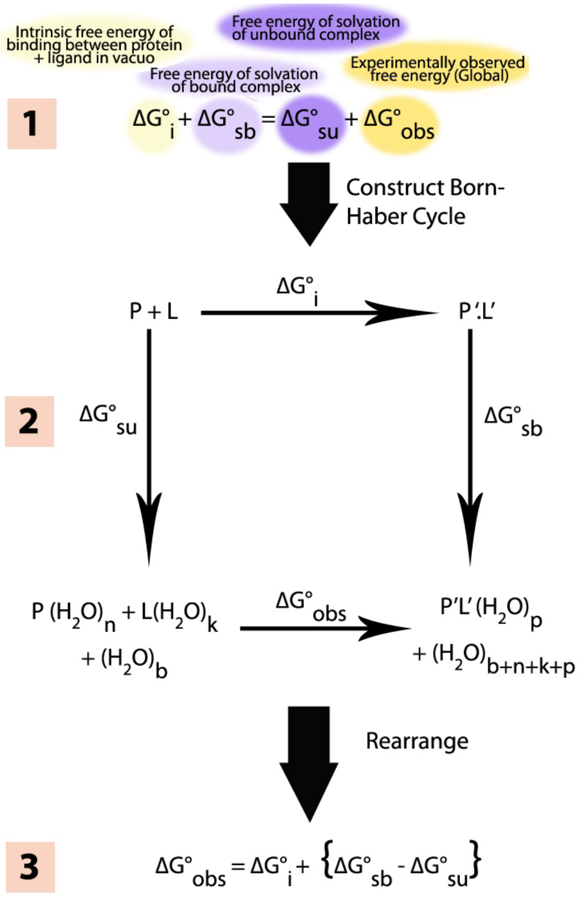

Intrinsic free energy of binding between protein + ligand in vacuo
Free energy of solvation of unbound complex
Free energy of solvation of bound complex
Experimentally observed free energy (Global)
1
ΔG°i + ΔG°sb = ΔG°su + ΔG°obs
Construct Born-Haber Cycle
P + L
ΔG°i
P'.L'
2
ΔG°su
ΔG°sb
P (H2O)n + L(H2O)k + (H2O)b
ΔG°obs
P'L'(H2O)p + (H2O)b+n+k+p
Rearrange
3
ΔG°obs = ΔG°i + {ΔG°sb - ΔG°su}

Fig.1.5. Thermodynamic decomposition facilitated by construction of a Born-Haber cycle. Image adapted from reference [33].

the ligand or protein residues have to adopt energetically strained conformations upon binding [33,34,36].

Entropic contributions to ΔG°_{i} on binding are usually considered unfavourable due to restriction of rotational, translational and internal DOF of the ligand upon binding (**Fig.1.4**). Protein residues in close proximity to the bound ligand are generally thought to rigidify and are thus also entropically penalised [33]. The solvation contribution is discussed in more detail in the next section.

1.1.6. The effect of solvent and the hydrophobic Effect

The aqueous medium in which most biomolecules interact is characterised by unusual properties such as the hydrophobic effect; a phenomena best illustrated by the observation that oil and water do not mix. Most students are familiar with the ordered tetrahedral geometry adopted by water molecules in ice. Each molecule forms 4 hydrogen bonds with its neighbours and the crystalline phase ensures translational and rotational entropy is close to zero, whilst the enthalpic term is maximised. However, liquid water exists in a far more dynamic environment, in which the hydrogen-bonding network is in a constant state of flux and a single water molecule has the possibility of orientating itself in multiple ways with respect to its neighbours. The scales balancing enthalpic and entropic components in this medium are more even compared to that in the solid phase. The classical hydrophobic effect qualitatively explains the segregation and aggregation of non-polar solutes in solution using the following description. Apolar solutes are incapable of forming strong hydrogen bonds with each other, or with the solvent and predominantly interact via weak dispersion interactions. On solvation, water molecules are unable to form the maximum number of hydrogen bonds possible when adjacent to such a solute. This equates to a loss of entropy due to the unavailability of previously available configurations. To maintain maximal hydrogen bonding interactions (and thus maximise configurational entropy), water molecules reorient themselves to form ordered "clathrate-like" structures around the solute. The force that drives the system is entropic and apolar solutes aggregate so as to minimise their exposed surface area. This allows surrounding water molecules to preferentially hydrogen bond with each other. Though water loses a small measure of entropy in the formation of these ordered structures, the expulsion of a proportion of these ordered water molecules into bulk solvent (upon minimisation of hydrophobic surface area) results in the overall process being highly entropic at physiological temperatures. On the other hand, despite the loss of some solvent-solvent hydrogen bonding, the enthalpy change of solvation is very small and negative: it is only just exothermic and favourable. This is because expelled interfacial water molecules renew hydrogen bonds with bulk solvent and the hydrogen bonds of waters involved in the clathrate structure possess greater strength.

In the case of a hydrophobic protein-ligand binding interaction, binding is considered a desolvation event as water molecules enveloping the ligand are stripped off on entry into the pocket. The enthalpic term is thus unfavourable, because the sign is the inverse of that observed for hydrophobic solvation. As detailed above, the associated entropic term dominates and is favourable due to a proportion of waters hydrating the protein's binding cavity being displaced upon ligand entry.

The specific heat capacity at constant pressure (ΔC_p) measures the quantity of heat energy required to change the temperature of a defined mass by one degree Celsius. Another perceived hallmark of the hydrophobic effect is its association with a negative change in ΔC_p at physiological temperatures. This follows from the observation that transferring apolar solutes (such as small aliphatic hydrocarbons) from a hydrophobic to aqueous phase is accompanied by a positive change in heat capacity and that the unfolding of proteins is also accompanied by a heat capacity change of the same sign. As water molecules participating in ordered clathrate structures have a higher heat capacity than those in bulk, they can absorb a greater amount of thermal energy by virtue of possessing low kinetic energies. Hence, the expulsion of waters from an aggregating apolar interface back into the bulk medium yields a concurrent decrease in the heat capacity. However, other factors for the decrease in heat capacity have also been proposed such as protein tightening on ligand binding [37].

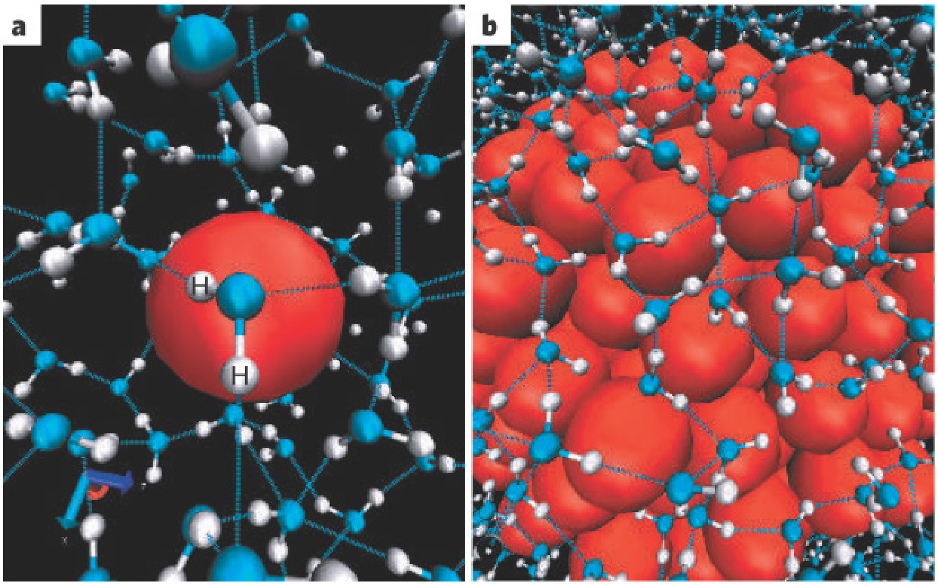


Fig.1.6. Red spheres are representations of methane-like molecules, whilst water molecules are depicted as blue ball and stick structures. Hydrogen bonds are coloured as blue dashed lines. Panel (a) shows that waters surrounding the single methane-like molecule participate in a hydrogen bond network is not greatly perturbed. However, in panel (b) waters surrounding the 135 member cluster (formed by aggregating methane-like particles) typically possess less than three hydrogen bond interactions. Image taken from reference [38].

The crystalline-like explanation given by the clathrate model for the aggregation of non-polar solutes is controversial. It is now considered to be a simplification, as MD simulations and neutron scattering experiments question the level of ordering and amount of hydrogen bond enhancement. A number of other models have been proposed to account for these discrepancies.

An interesting alternative theory proposed by Chandler (2005) regarding the hydrophobic driving force, states that very small non-polar solutes under 1 nm^2 in surface area (e.g. methane) are so small they do not consequentially interrupt water's hydrogen bond network. i.e. Water molecules surrounding the solute are still capable of maintaining 4 hydrogen bonds with their neighbours, despite the excluded volume taken up by the solute. Hence, there is no impetus for a few modestly separated hydrophobic molecules to coalesce in a very dilute solution. It is only in very close physical proximity to each other that the thermodynamic cost of association is naturally overcome. On the other hand, the formation of larger non-polar structures does disrupt the ability of water molecules to hydrogen bond with each other and this leads to the formation of a dewetted (intermediate phase between vapour and a liquid) interface. As the cost of creating the new interface scales with surface area whilst the thermodynamic forces opposing this scale with volume, a critical size greater than 1 nm^2 must be amassed before these apolar assemblies become metastable. The total solvation free energy for very small, dispersed apolar solutes equates to the algebraic sum of their solvation energies. Larger apolar assemblies possess total energies smaller than the sum of their parts as they possess larger volume to surface area ratios. This energetic discrepancy is the driving force for hydrophobic association which is dominated by enthalpy for large assemblies and the entropic term for small molecules. Unlike the clathrate theory, evidence from experiments and MD simulations do provide supporting evidence for this model (**Fig.1.6**) [33,39,40] [41,42] [33,34,37,43,44] [38].

1.1.7. Surface area burial

The Scorpio database is one of several that contains ITC derived thermodynamic data. It contains over 254 protein-ligand complexes from peer reviewed sources. On scrutinising this database, Olsson et al. (2008) found that burial of hydrophobic surface area in protein-ligand complex formation (i.e. difference in hydrophobic surface area of complexed ligand to that of free protein and ligand) contributed favourably to free energy [45]. This equated to a burial of 20 Å^2 of Connolly surface area (CSA) reducing ΔG° by -1 kJ/mol (**Fig.1.7**).

The authors noted that smaller ligands buried a roughly equal proportion of polar and apolar groups, whilst larger ligands tended to bury a greater proportion of apolar surface area. Polar groups afford ligand specificity as they are capable of making high affinity interactions (e.g. H-Bonds) with the protein. However, these are sensitive to spatial positioning, unlike apolar interactions that are "amorphous and nonselective by nature" [45]. Increasing polar surface area burial increases the complexity of interactions and is therefore harder to achieve. Thus, smaller ligands have to bury a greater proportion of polar groups to achieve minimum specificity, compared to larger ligands that maximise apolar surface area burial to avoid excess complexity.

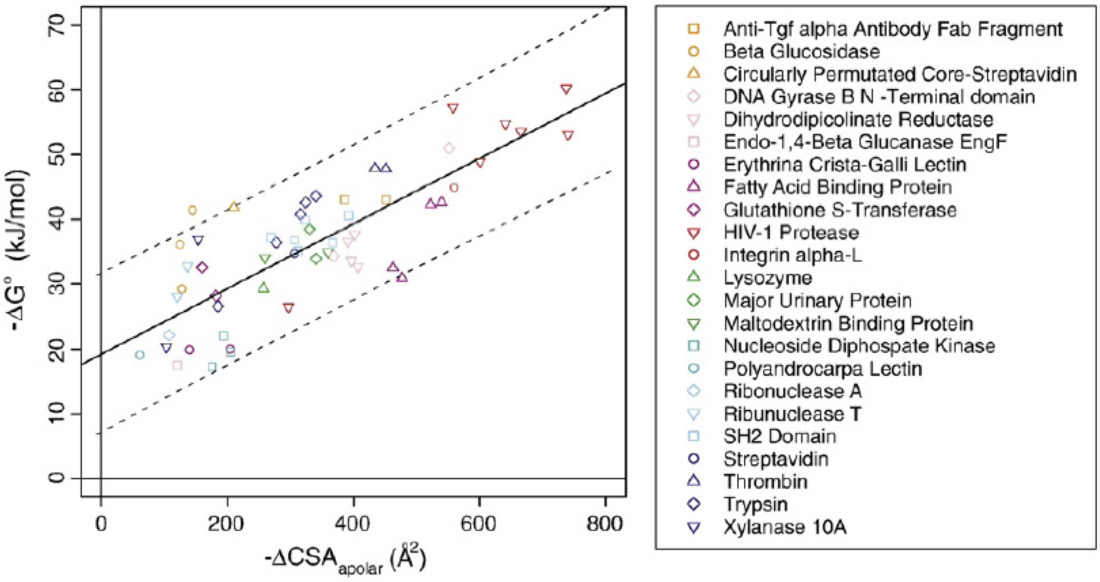


Fig.1.7. Correlation of ΔG° with apolar surface area burial. Dotted line gives 95% confidence interval of the fit. Image and information taken from reference [45].

The compounds in the Scorpio database do not show a clear relationship between molecular weight and affinity, but do show strong correlations between apolar surface area burial and affinity. Thus the latter is a better indicator of affinity as it accounts for the binding interface only. This is more specific than looking at the whole molecule which may possess groups uninvolved in the interaction [45].

1.1.8. Maximal ligand affinity

Analysis of a variety of natural & synthetic ligands by Kuntz et al. (1999) found the maximal free energy possible in non-covalent interactions (with a few exceptions such as Biotin) to be ~ -62.7 kJ/mol (10^{-11} M) [46]. The contribution per non-hydrogen atom (NHA) is -6.27 kJ/mol for the first 15 atoms, after which the contribution made to the free energy of binding sharply dropped to ~0.04 kJ/mol (**Fig.1.8**). The authors examined contributions to ΔG° from hydrogen bonds, electrostatic, van der Waals and hydrophobic interactions, concluding that the effect was primarily due to the latter two. They proposed a solvent-free model that reproduced similar free energy contributions per NHA. Adding more atoms increased "self-shielding", resulting in a reduction of the energy contribution per NHA. This phenomenon was geometry dependent, with linear ligands experiencing less shielding than a planar or cubic ligand upon burial in the receptor. Though their model did not directly take solvent into account, they suggested that entropically based phenomena such as the hydrophobic effect is linked to and would augment van der Waals interactions. There is the additional caveat that the dataset may not represent compounds of extremely high (femtomolar) binding affinity as such ligands would take years to disassociate and thus might be evolutionarily selected against [46].

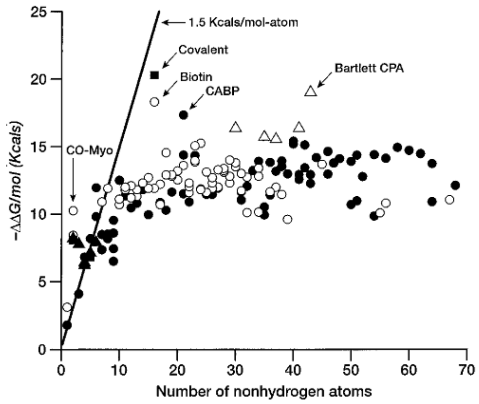


Fig.1.8. Half maximal inhibitory constants (IC_{50}) or disassociation constants (Kd) were used to calculate the change in free energy ($\Delta\Delta G°$) versus number of NHAs for a variety of strongly binding ligands. Image taken from reference [46].

Another study containing a larger dataset of 1,012 enzymes suggests that strong binders with picomolar (10^{-12} M) to ten zeptomolar (10^{-20} M) affinities utilise enzyme transition states involving the formation of covalent intermediates. However, enzymes with affinities less than 10^{-11} M (corresponding to the -62.7 kJ/mol affinity limit) operated via weaker non-covalent mechanisms such as hydrogen bonding and electrostatic interactions. e.g. Chorismate mutase ($10^{10.6}$ M^{-1}) and Cyclophillin ($10^{8.7}$ M^{-1}) [47]. The maximum free energy obtainable for non-covalent interactions was also observed by Gilli et al. (1994) in a study of the binding of 136 ligands to 10 receptors [48]. The authors observed that the smallest observable K_d was 10^{-11} M for high-affinity drugs. They suggested that this behaviour was caused by compensation between enthalpic and entropic forces primarily due to the differences in hydrogen bonding, prior to and after binding.

These experimental data and conflicting conclusions raise questions regarding the physical basis for the observation of a free energy ceiling of -62.7 kJ/mol. A proposed answer is that most non-covalent interactions are subject to entropy-enthalpy compensation (EEC) and thus cannot easily attain free energy values that exceed this level of affinity. However, it is clearly possible to design drugs that break this apparent barrier, as there are binding interactions that possess exceptionally high non-covalent affinities. e.g. the avidin-biotin complex (~10^{-15} M) [49].

1.2.0. Entropy-Enthalpy Compensation

This phenomenon occurs when one of the two components constituting ΔG° (enthalpy or entropy) compensates a change in its counterpart with an equivalent change possessing the same sign (**eqn.1.6**). The net result is a ΔG° change at, or close to zero. It has been suggested to arise "because bonding opposes motion and, also reciprocally, motion opposes bonding" [30]. For example, the formation of a hydrogen bond between a ligand and protein generates a favourable enthalpic contribution to ΔG°, but also restricts the movement of the ligand, thereby reducing entropy and negating any enthalpic gain. This issue is particular to non-covalent interactions in solution, as the weaker bond strengths are easier to disrupt by comparable thermal energies at 298 K [30,50,51]. Though this definition neglects to account for potential solvent effects as a source of compensation, it is a useful concept to begin understanding EEC.

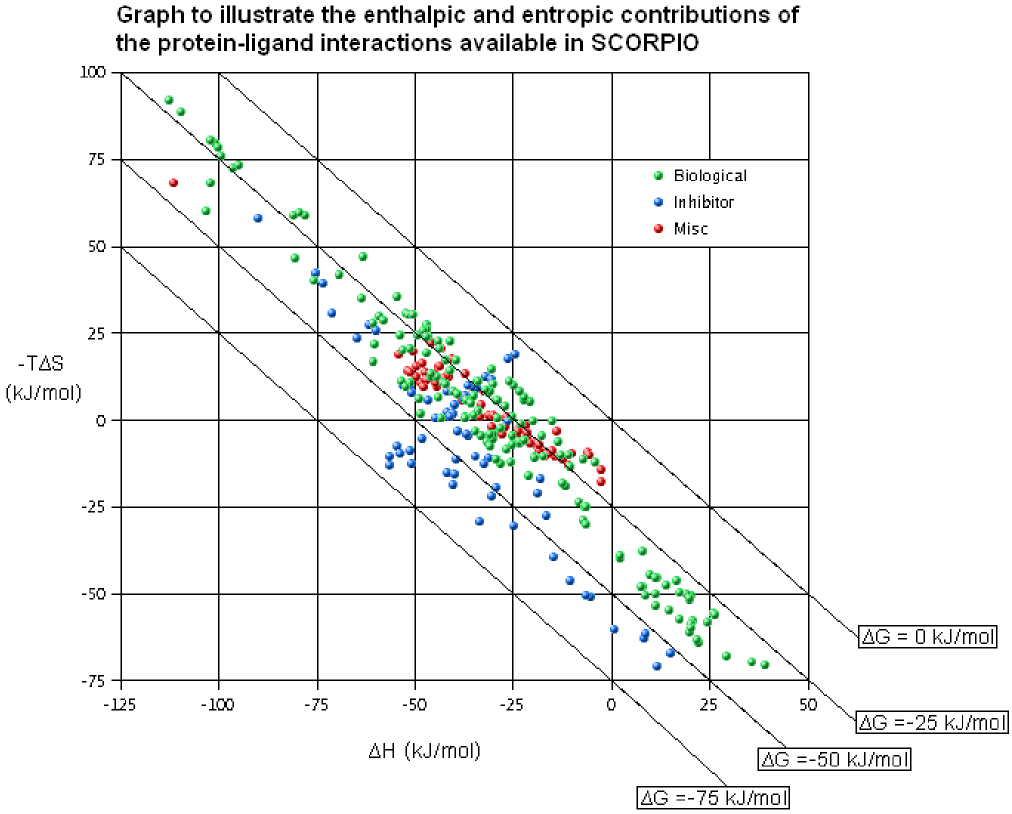


Fig.1.9. Entropy-enthalpy plot of compounds in the Scorpio database derived from ITC data. Image taken from reference [54].

When plotting entropy vs. enthalpy, the relationship sometimes displays a linear correlation (coefficient > 0.95) and the data points fall within a narrow diagonal band. The data can be derived from databases of binding data, and/or examination of a homologous series of ligands with small incremental alterations in structure. e.g. The linear alkane series: pentane to hexadecane. This has led to heated debate as to whether this has extra-thermodynamic significance (i.e. the consequence of a real physical relationship that cannot be defined by statistical thermodynamics) or is in fact a statistical anomaly [52,53]. Data from the Scorpio database shows a characteristic EEC plot with ΔH° and ΔS° values spanning a range much greater than those observed for ΔG° (**Fig.1.9**).

The thermodynamics of EEC has been theoretically described in terms of bonding between two molecules: A + B. The formation of the A-B complex can be represented by a potential energy well, in which the deepest part of the well represents a stronger bond than the shallower part. As the interatomic distance is reduced, the bond strengthens until it reaches a point where repulsive forces dominate. The potential energy (*V*) can be used to measure the enthalpy. The entropic contribution can be calculated from the vibrational frequency (*v*), using statistical mechanics. The entropy is very small at vibrational frequencies observed for covalent bonds (v > 1000 cm^{-1}). As the frequency drops, entropy increases rapidly [50]. This can be represented by an idealised entropy-enthalpy plot, in which increased bonding strength is linked to a loss of vibrational entropy. Weak interactions are characterised by greater entropic penalties, whilst stronger (covalent) interactions are penalised less because the majority of the entropic cost has already been overcome. In-between both these extremes, the entropy roughly equals enthalpy and a linear relationship is observed (**Fig.1.10)** [55].

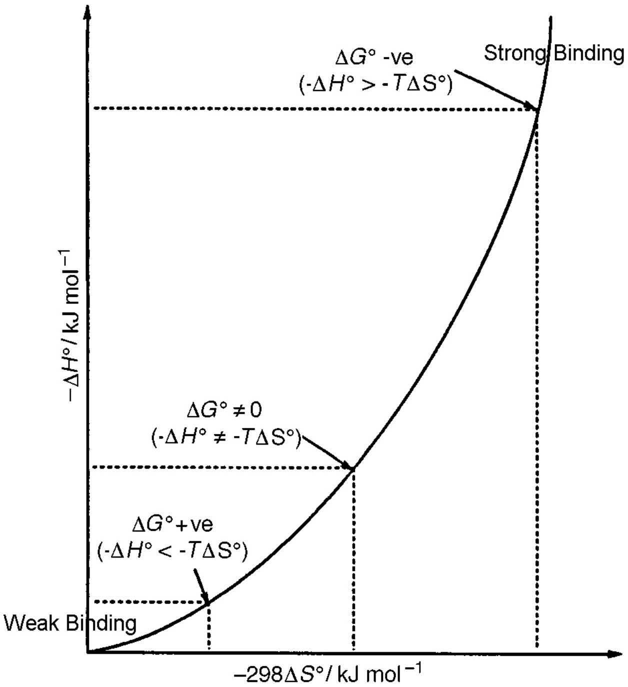


Fig.1.10. Exothermicity in terms of the entropic cost for the reaction A + B --> AB. Image taken from reference [55].

The observation of EEC effects in datasets derived from disparate chemical compounds suggests that EEC is the result of “a limited Gibbs free energy window” that illuminates only the linear portion of **Fig.1.10.** This is likely to be due to practical limitations in the experimental measurements of thermodynamic values and lack of true diversity in the databases of compounds examined [56].

Many examples of EEC have been discredited, due to the use of a vant Hoff or Arrhenius plot to determine the entropy (intercept) and enthalpies (slope) from differences in experimental binding constants measured at various temperatures. As most biological reactions are only feasible within a restricted temperature range, the number of experimentally obtainable measurements is likewise limited. Thus, determination of the ordinate intercept often involves an extrapolation many times the range of the data (**Fig.1.11**).

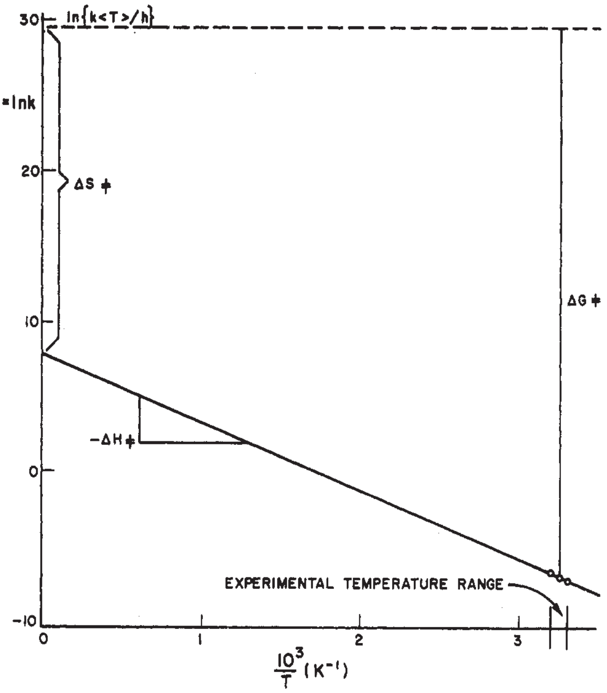


Fig.1.11. Extrapolation of data from a small temperature range to calculate enthalpy and entropy. Image taken from reference [57].

Thus, the resulting entropy and enthalpy values have large systematic errors that are characterised by a high correlation coefficient [52,53,57]. While these considerations may seem like a damning indictment of EEC, they should not extend to ITC analysis which directly measures ΔH°, K_d and stoichiometry within a single experiment. Olsson et al. (2008) caution that this does not necessarily qualify as prima facie evidence for EEC, as error margins of calorimeters cause difficulties in measuring ΔG° > -20 kJ/mol [45]. Additionally, caveats regarding binding databases used in studies (such as Gilli, Kuntz and Lipinski's rules of 5) are also relevant here. There are limitations in the experimental measurements that can be gleaned for very small or large ligands. The binding affinities of the former are too small to detect, whilst larger ligands tend to be insoluble and are thus intractable to study [24,45,46].

1.2.1. Entropy-enthalpy compensation case studies

Successful drug design involves knowledge of how modifications to structure affect the relationship between ΔH° and ΔS°. A deeper understanding of the forces that govern EEC can potentially guide rational design to maximise ligand efficacy. Consequently, the following sections detail real-world case studies that illustrate the molecular basis of a variety of EEC effects.

1.2.2. HIV-1 Protease

Several drugs have been developed and put on the market to mitigate the problem of acquired immunodeficiency syndrome (AIDS), which is caused by the human immunodeficiency virus (HIV). The viral protein known as HIV-I protease cleaves newly synthesised polyprotein precursors, thus facilitating the creation of multiple mature proteins essential to the life cycle of the virus. Freire (2008) noted that the binding signatures of HIV protease inhibitors (PIs) have evolved from being entropically to enthalpically dominated over the last 11 years. He suggested that the reason for this evolution is that, entropically favourable compounds are easier to design as this merely involves the addition of non-polar groups. This results in a positive contribution from the hydrophobic effect and a compound with an entropically favourable binding signature. However, enthalpic gains take more time and work to engineer because of the complexities posed by EEC [58]. For instance, hydrogen bonds range in strength from weak to very strong. A typical interaction can contribute ~6.0 kJ/mol of favourable enthalpy, which roughly translates to a tenfold increase in binding affinity. A weak hydrogen bond interaction caused by poor spatial positioning may even be unfavourable. Mutational studies on tyrosyl-tRNA synthetase have indicated that hydrogen bonds with charged groups have enthalpies of -14.2 to -18.8 kJ/mol and those with uncharged groups have -2.1 to -6.3 kJ/mol, whereas bonds with poor spatial positioning contributed negatively (~ -1.8 kJ/mol) [59]. For example, the HIV-1 protease inhibitor, KNI-10033 was engineered with a sulfonyl group that made the binding enthalpy more favourable by -16.3 kJ/mol via the formation of a strong hydrogen bond. However, an entropic penalty of 17.1 kJ/mol was incurred as a result of decreased ligand conformational flexibility and reduced desolvation entropy. The proposed strategy to mitigate entropic loss, was to design inhibitors with hydrogen bond acceptors/donors placed adjacent to well structured regions of the protein [60].

Constraining the ligand to minimise entropic losses and maximise favourable enthalpic contacts may seem like the way forward to design effective drugs. However, the resulting rigid body inhibitors are optimised for shape complementarity and thus engineered binding affinities can be severely affected by binding-site mutations. This is commonly seen in HIV-1 protease, resulting in ~40% of infected patients possessing mutant variants that are resistant to at least one of the currently available PIs. An interesting example highlighting the subtlety of entropic compensation was observed on adding a phosphonate moiety to the PI, TMC-126. The phosphonate variants displayed similar inhibitory activity (2 to 7 pM) to the parent molecule when complexed with wild-type HIV-1 protease. However, TMC-126 experienced a reduction in activity vs. HIV-1 protease mutants M461/147V/150V (41.3 pM), and I84V/L90M (12.3 pM) compared to K_i's of only 0.8 to 1.6 pM for the phosphonate derivatives. ITC analysis revealed that whilst all the PIs had suffered reductions in binding enthalpy (ΔΔH°: +8.4 to +10.9 kJ/mol), the PIs with a phosphonate moiety entropically offset the free energy loss (ΔΔS°: 8.8 to 9.2 kJ/mol). The entropic increase was not due to desolvation or increased ligand torsional freedom, but was instead proposed to be caused by a "solvent anchoring mechanism."

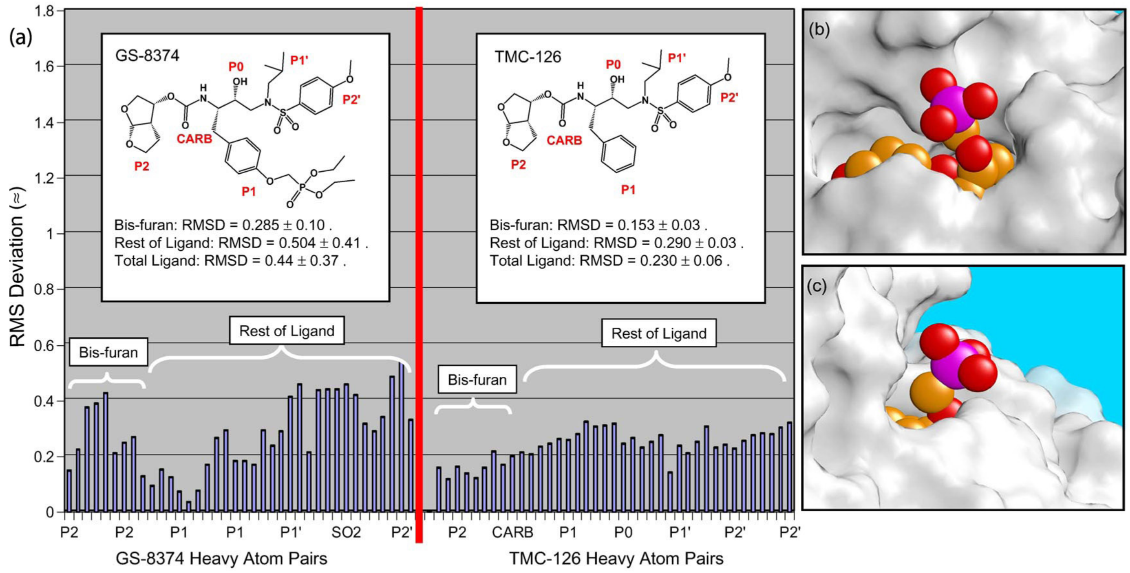


Fig.1.12 (a) 2D structure of TMC-126 and the phosphonate derivative, GS-8374. RMS deviations (indicating mobility) of the ligands bound to the I84V/L90M mutant are plotted as a histogram beneath the relevant part of the molecule. **(b)** and **(c)** show structures of GS-8373 bound to HIV-1 protease. Note the phosphonate juts out of the binding site. Image and information taken from reference [61].

The I84V/L90M mutant is successful against inhibitors because the size of the binding site is enlarged compared to wild-type. Crystal structures of the complex showed that the solvated phosphonate moiety protrudes out of a hydrophobic channel at the enzyme surface. The electron density from the crystal structure indicated that this group sampled multiple conformations and was unlinked with the binding site. The relatively immobile P1 aromatic ring (**Fig.1.12.a inset**) is wedged in the hydrophobic channel and is thought to act as a "molecular fulcrum" that transfers the dynamics of the exterior phosphonate moiety to ligand functional groups located within the heart of the binding pocket. This results in an increase in their conformational mobility. TMC-126 also exhibits a measure of mobility in the enlarged pocket, but phosphonate PIs are better at adapting to (mutation induced) alteration of the binding cavity volume/shape. Thus, they sample alternative binding poses and avoid losses in potency (**Fig.1.12**). This directly translates into the phosphonate analogues having a better resistance profile than TMC-126 [61].

1.2.3. Carbonic Anhydrase

This enzyme reversibly catalyses carbon dioxide and water to protons and bicarbonates so that pH balance of blood and tissues are maintained. A review by Krishnamurthy et al. (2008) details the many reasons why bovine carbonic anhydrase (BCA) is considered a model system [62]. It is a very simple system through which the various biophysical aspects of binding interactions can be studied. e.g. ligand design and protein-ligand selectivity. It is easily purified and amenable to a variety of experimental techniques. Despite not being an important medical target, it is very well characterised and has aspects that are representative of features possessed by a larger number of other enzymes. Additionally, its binding site can bind a series of inhibitors that differ incrementally from each other in terms of structure. Consequently, it functions as a useful test system in which perturbations to binding can be correlated to changes in ligand structure. Histidine residues coordinate a catalytic zinc ion that is situated at the base of a conical cavity that forms the hydrophobic binding site [62]. ITC investigations into the binding of a series of para-substituted benzene-sulfonamides of various chain lengths (n = 1-5) demonstrated almost perfect EEC. As chain length increased, the enthalpy became unfavourable and the entropy favourable, whilst ΔG° remained relatively constant (**Fig.1.13**).

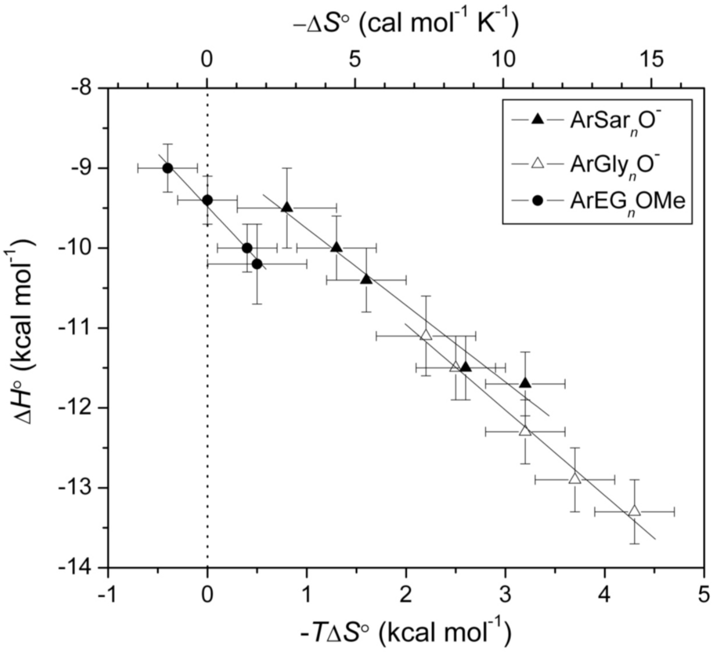


Fig.1.13. Entropy-enthalpy compensation plot of three separate series of ligands. Image taken from reference [63]

As the effect was independent of heat capacity, the hydrophobic effect was thought to be an unlikely cause. To account for the decrease in enthalpy with increasing chain length, a model was proposed whereby distal residues in the ligand exerted a destabilising effect on the proximal residues adjacent to the zinc ion (**Fig.1.14)** [63].

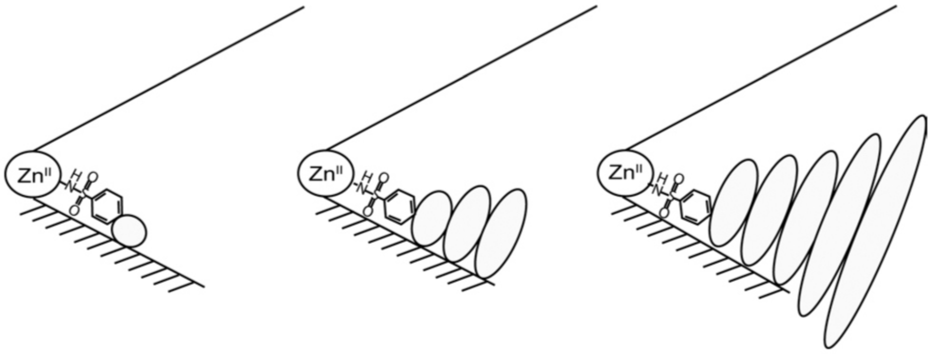


Fig.1.14. Model illustrating how addition of successive residues decreases the stability of previous residues. Ellipse size represents the mobility of the residue. The ligand is bound within the conical binding site of carbonic anhydrase. The hydrophobic wall is represented by cross-hatching. Image and information taken from reference [63].

In a subsequent independent NMR spectroscopy study, 2 sets of 6 para-(glycine)$_n$-substituted benzene-sulfonamide (ArGlynO-) ligands complexed with BCA were examined. Series-1 contained incremental additions of GLY residues (1-6), ^{15}N labelled at the terminal glycine. All ligands in series-2 were composed of six residues and differed from each other in terms of which residue was labelled (**Fig.1.15**).

Comparing adjacent ligands in a series allowed the intrinsic entropic contribution of each additional glycine residue to be measured. This decomposition revealed a different picture to that afforded by global ITC values. Increasing chain length *decreased* the mobility of preceding residues (**Fig.1.15**). This model is subtly different and can be summarised in 4 points:

1. Series-1 ligands are labelled on the terminal residue and thus have a larger measured S^2 mobility relative to series-2 ligands, as they are attached to only one vincal neighbour.

2. Series-2 ligands demonstrate that the glycine residue nearest the zinc ion has the lowest mobility (high S^2).

3. Adding successive glycine residues to series-1 ligands does not greatly increase the mobility of the terminal residue compared to the other ligands in series-1 that possess fewer residues (S^2 values roughly the same).

4. Comparison of residues 1-4 of both series indicates that the residues in series-2 ligands have lower mobility compared to series-1. Thus, the addition of glycines restricts mobility of residues proximal to zinc.

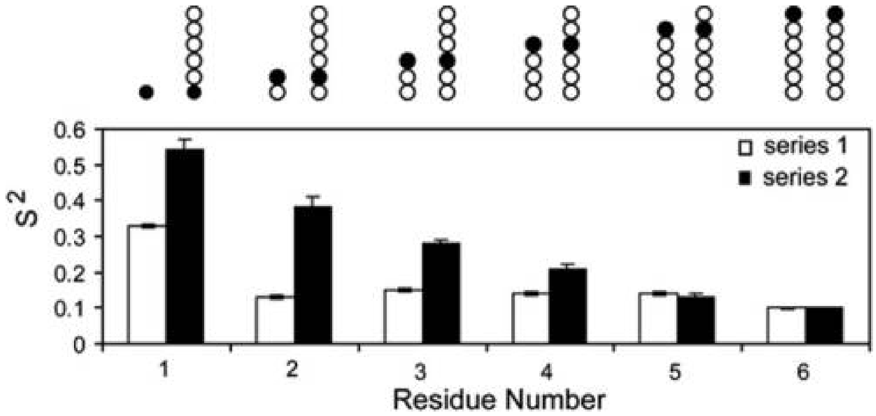


Fig.1.15 NMR spectroscopy derived order parameters for ^{15}N-^{1}H bond vectors in series-1 and 2 ligands bound to carbonic anhydrase. The number of glycine residues is represented above the histogram. Filled circles represent the labelled residue. Note that high values of S^2 equate to lower mobility. Image taken from reference [64].

This demonstrates that ligand destabilisation is not the reason for the observed entropic binding signature. The conclusion drawn from further MD calculations was that increased enzyme side-chain mobility (particularly near the binding pocket) was primarily responsible for the globally observed favourable entropic term [64].

1.2.4. High-affinity non-covalent Interactions

A discussion of non-covalent interactions would be incomplete without examining the avidin-biotin complex. Biotin is a soluble vitamin of ~240 Da that displays strong co-operative binding to avidin ($\Delta G° = -85.9$ kJ/mol), a component of egg white. The complex is 319 to 320 K more heat stable than avidin alone. Biotin binds to a β-barrel that displays a high degree of shape complementarity to the ligand. Tryptophan and phenylalanine residues form a rigid "hydrophobic box", whilst polar residues form a stabilising hydrogen bond network with the ligand (**Fig.1.16**).

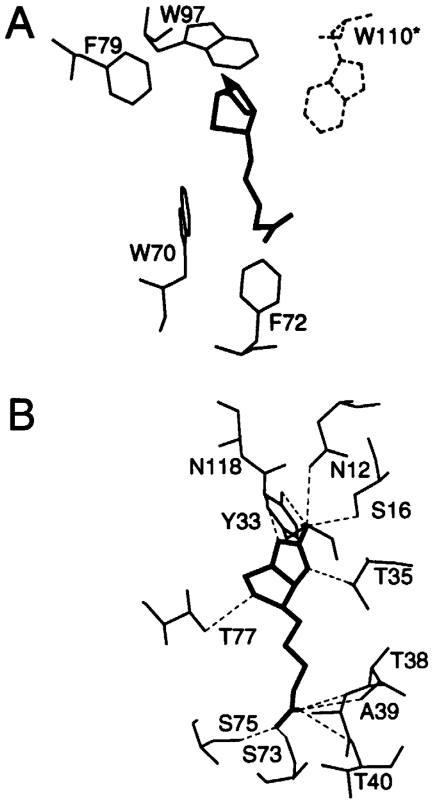
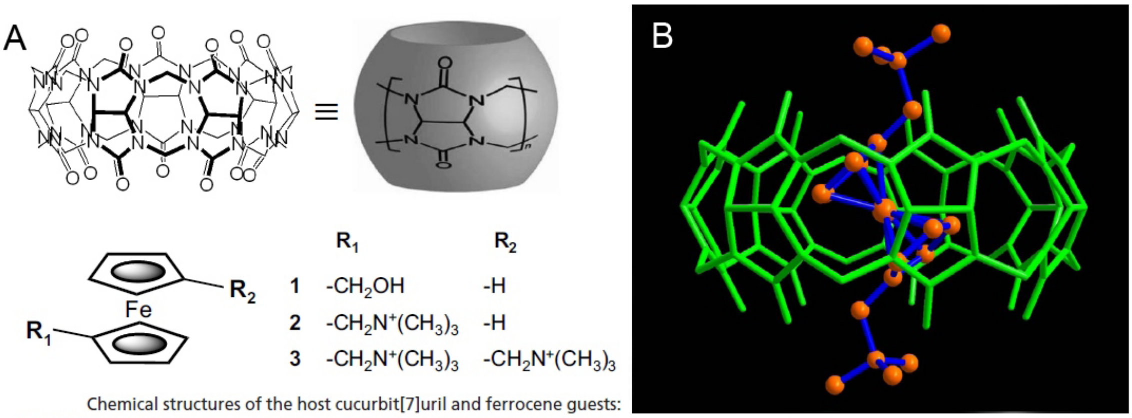


Fig.1.16. Binding site of avidin-streptavadin complex. **(a)** Hydrophobic binding site residues. **(b)** Hydrophilic binding site residues. Image and information taken from reference [49].

Upon binding, water molecules are expelled from the binding site and a mobile loop stiffens and "locks" biotin into the binding site [49,65]. ITC Calorimetry indicated that biotin binds with a strong enthalpic signature (ΔH° = -84.9 kJ/mol) as a consequence of favourable hydrogen bonding and van der Waals interactions. Loss of conformational DOF is minimised because the binding site residues are already rigid, whilst an additional favourable desolvation term results in a net entropy change of zero [66,67].

This interaction was taken as the inspiration for the design of a synthetic host-guest interaction. Cucurbiturils are nanoscale macrocycles that are synthesised from the condensation of glycoluril and formaldehyde. They are abbreviated as CB[n], where n refers to the number of repeat units (5-10) [68]. Rekharsky et al. (2007) investigated non-covalent interactions in-between CB[7] and 1.1'-bis(trimethylammoniomethyl)ferrocene to achieve an extremely high K_a of $3x10^{15}$ M^{-1} (ΔG° ~ -87.8 kJ/mol) [69]. The guest is characterised by two cationic sidearms that can interact with one of the carbonyl oxygen atoms lining either side of the hosts ring system (**Fig.1.17**).



Chemical structures of the host cucurbit[7]uril and ferrocene guests: 1-hydroxymethylferrocene (1), 1-trimethylammoniomethylferrocene (2), and 1,1′-bis(trimethylammoniomethyl)ferrocene (3).

X-ray Crystal Structure of CB[7] complex

Fig.1.17. (a) Structure of CB[7] host and binding three ligands (1-3). **(b)** 3D structure of ligand 3 bound to CB[7]. Image and information taken from reference [69].

Competition ITC experiments were carried out to accurately quantify the binding enthalpies. They indicated binding was driven by a favourable enthalpic contribution (-90 kJ/mol). Comparison with ferrocene-based guests with single or no sidearms established that this feature did not increase the enthalpy, but instead increased the entropy by 16-18 kJ/mol so that it approached zero.

There are several reasons behind such tight binding:

1. The ligand displays rigidity and shape complementarity with the binding cavity and this resulted in hydrophobic interactions being maximised whilst losses in configurational entropy were attenuated.

2. The surface area of the guest that is buried upon complex formation is large (~55% of host cavity volume).

3. Reduction in configurational entropy is mainly affected by restriction of the ferrocene core. The cationic sidechains do not perturb this value greatly and the majority of the favourable entropic term is obtained from the expulsion of waters hydrating CH[7]'s cavity back to bulk.

When placing this interaction in context of the Scorpio binding data, this example can be taken as an example of entropy-enthalpy reinforcement as both entropic and enthalpic terms are favourable (**Fig.1.18**) [69]. Thus, by subtle optimisation of disparate entropic and enthalpic contributions, it should be possible to overcome the apparent ceiling on maximal ligand affinity and design inhibitors with exceptional affinity.

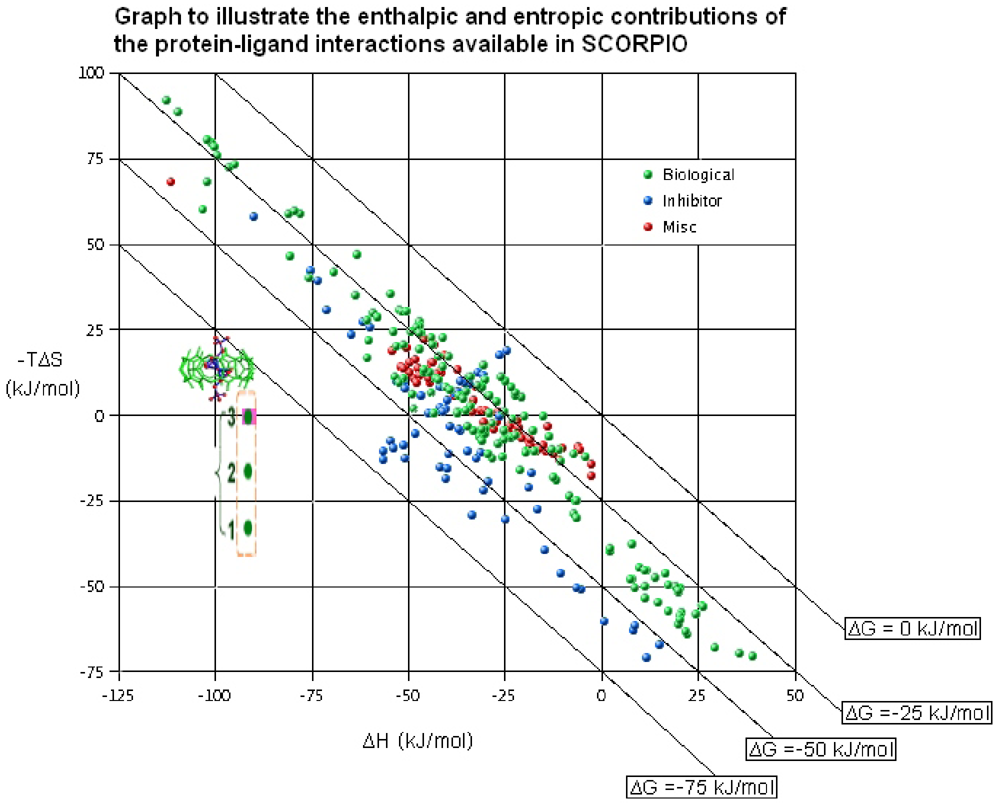


Fig.1.18. Points marked 1-3 correspond to ferrocene ligands: Ligand 1 has no cationic arms, whilst ligands 2 and 3 have one and two cationic arms respectively. A putative entropy-enthalpy reinforcement plot would move from the bottom left quadrant to the top right. Image adapted from references [54,69].

1.3.0. The Mouse Major Urinary Protein (MUP)

MUP-I is one of several isoforms of a protein found in mouse urine and is estimated to form as much as 99% of the total protein content in that medium. MUPs are typically expressed in areas associated with pheromone excretion such as the liver and kidneys at such high concentrations (5-10 mg/ml) that a significant metabolic cost is incurred.

MUPs are ubiquitous and are also found in many other tissues and secretions. For example, MUP-IV is found in the vomeronasal mucus and has substantial sequence divergence and much higher binding specificity compared to the urinary isoforms [70 71].

As a member of the lipocalin family, MUPs possess a characteristic beta-barrel motif that forms a hydrophobic cavity which can promiscuously bind a variety of small volatile pheromones. These are also collectively known as volatile organic compounds (VOCs) and have been linked with individual recognition, kin recognition, inbreeding avoidance, inter-male aggression, onset of female puberty, pregnancy termination courtship mating, and other behavioural and physiological effects in mice [72,73 74–76] .

MUP proteins possess significant sequence polymorphism as a consequence of the large number of MUP genes found in mouse chromosome 4 and, to date, more than 2,000 MUP sequences can be found in GENBANK. Most of the amino acid differences can be tracked to substitutions on the surface of the protein with a more limited number involving modifications to the binding cavity. The latter would ostensibly have a greater impact on ligand specificity and binding [77,78]. Studies indicate that mouse urine contains a heterogeneous population of MUPs whose structural differences allow them to differentially bind various populations of VOCs. Analysis demonstrates that the concentration and composition of VOCs within urine produces a unique “scent signature” that can vary not only by individuals but also according to mouse strain and gender. This signature changes naturally over time as urine deposited in the environment ages, and one of the predominant theories regarding MUP’s structural function is its facility to protect these volatile chemical messengers from premature decomposition and evaporation. MUPs are extremely stable in an aqueous milieu and can withstand melting temperatures in excess of 70°C. Experiments that involve the addition of guanidine hydrochloride to aged mouse urine indicate that the VOC profile dramatically changes upon denaturation of the protein, releasing pheromones that should have evaporated almost immediately. This serves to reinforce the thesis of a time-delay role for MUPs in scent communication [74,79].

As well as having a detailed and interesting physiological role, the ability of some MUPs to promiscuously bind a series of chemically related ligands makes it an ideal model system to study the binding of hydrophobic ligands that structurally vary from each other in an incremental manner. Thus, there exists a significant corpus of research that seeks to elucidate the thermodynamic and structural details on the binding of “natural” ligands such as 3,4-dehydro-exo-brevicomin (DBH); 2-methoxy-3-isopropylpyrazine (IPMP); 2-methoxy-3-isobutylpyrazine (IBMP); 2-sec-butyl-4,5-dihydrothiazole (SBT); and 6-hydroxy-6-methyl-3-heptanone (HMH) [34 74 78] (**Fig.1.19**). The binding of primary alcohols to MUP are of particular interest, as these ligands form a panel whose members differ from one another in terms of carbon chain length. Thus, this allows analysis of the thermodynamics of binding and how it is perturbed by modifying the length of the ligand by a methylene [36]. Investigations into these panels and their structural analogues form the focus of this thesis.

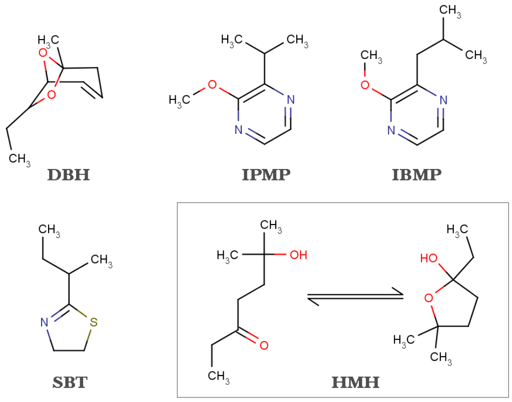


Fig.1.19. Structures of some of the different ligands that bind to MUP-I. HMH exists in equilibrium between the form of a closed furan ring and an open hydroxyketone tautomer that both bind to MUP-I [72].

An improved theoretical understanding of binding practically assists in the engineering of purpose built proteins that take advantage of the hyperplasticity of the MUP binding site to perform functions such as drug delivery. As the excessive hydrophobicity of many promising drug leads violates Lipinski's "rule of five", MUP's hydrophobic cavity can be leveraged to contain and transport these poorly soluble compounds to sites of activity [80] [81]. Members of such engineered lipocalins form a protein family known as anticalins. As they maintain structural stability despite sequence modification, they are designed to act as drug delivery systems, scavenging systems that assist in toxic compound removal, and as antibody mimetics that can bind other protein targets with high specificity [82–84].

1.3.1. Brief structural dissection of MUP-I

Lipocalins are generally 18-20 kDa in weight and, along with fatty acid binding proteins (FABPs), triabins, avidins, and a subset of metalloprotease inhibitors, form a larger superfamily known as Calycins. The etymology of the label 'lipocalin' comes from a conjugation of "lipo" and the Latin word "*calyx*", to describe the hydrophobic cuplike structure of the binding cavity which forms the key structural motif within this family. The calyx consists of a β-clam fold fashioned by the arrangement of eight antiparallel β-strands labelled *a-h*. Strands *b-d* are orientated orthogonally with respect to *e-h*. The series of loops (denoted L2-L7) that join these strands are β-hairpin +1 connectors, whilst the L1 forms a large Ω loop. Within the MUP family, the bottom portion of the calyx is closed, whilst the L1 loop forms a lid that caps the opposite end of the binding site (**Fig.1.20**). However, this is not the rule for all lipocalins and an example of a notable exception is neutrophil gelatinase-associated lipocalin (NGAL), which possesses a funnel-like cavity exposed to solvent.

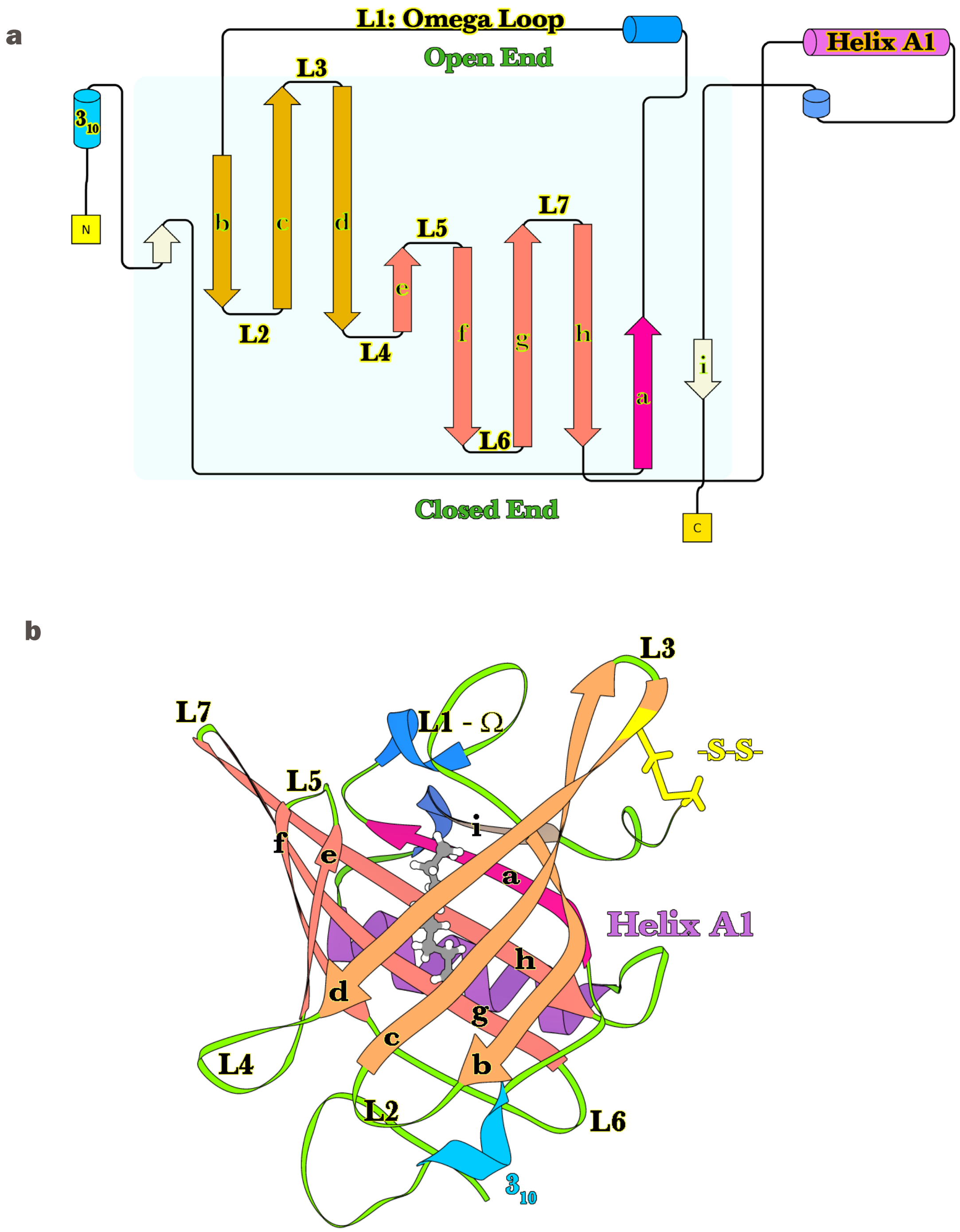


Fig.1.20. (a) Protein topology map created with pro-origami [85]. (b) 3d structure of MUP-1 bound to octanol (ball & stick). Structural features are assigned canonical names. Secondary structure elements are colour coded to match features in 2D and 3D representations.

Lipocalins share low pairwise sequence similarity (< 20%) and one of the essential criteria for membership within this group is the number of sequence conserved regions (SCRs). There are three main, short SCRs and assignment of a putative lipocalin into kernel or outlier lipocalin subfamilies is accomplished via quantification of the number of SCRs the protein possesses. The first SCR encompasses the 3_{10}-helix and strand *a* near the N-terminus. It maintains the highest degree of sequence and structural conservation, whilst SCR2 corresponds to the bottom portion of strands *f*, *g* and the

L6 loop. The SCR3 motif consists of strand *h* in addition to a few flanking residues and tends to maintain greater sequence versus structural conservation [83,86,86–88].

Lipocalins can have a variable number of disulphide bonds. MUP-I possesses one such bond that acts to directly connect the C-terminus to the base of strand *d*, near the L3 loop. The main sidechain available for a directional hydrogen bond belongs to TYR120 and is situated deep in the calyx core. Due to the conformational flexibility of the *c*, *d* strands, and the L3 loop region, ligand entry into the occluded pocket has been postulated to be mediated by these regions. Furthermore, examination of crystal structures indicate that the most direct route out of the pocket is by passing over the methionine situated in the cleft formed by the *d* and *e* strands [72].

Ligands binding to the MUP-I cavity are caged by a mixture of aliphatic and aromatic residues (**Fig.1.21**). The arrangement of these hydrophobic sidechains along the sides of the calyx can be broadly categorised into three levels. The predominant residue, leucine lines the upper rim of the calyx, and is overshadowed by a dyad of phenylalanines that act in concert with other Ω loop residues to cap the open mouth of the protein. A further two leucines are located in the saddle of the calyx, interspersed betwixt a triumvirate of isoleucines and function to block ligand exit

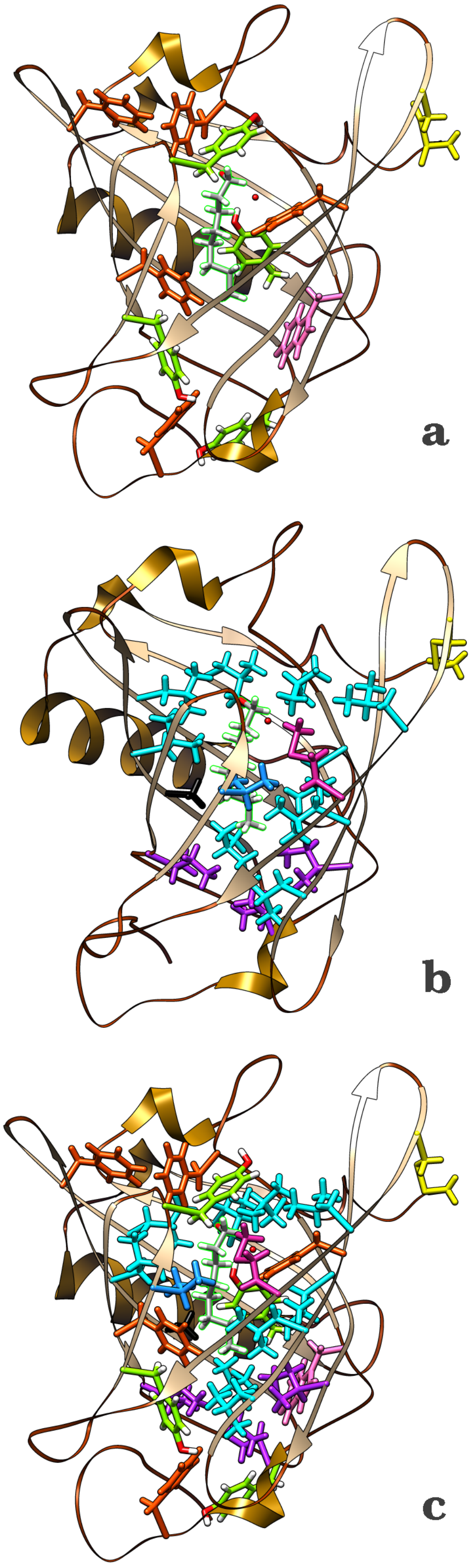


Fig.1.21. The architecture of the MUP binding pocket **(a)** Aromatic residues that line the cavity; **(b)** Aliphatic residues; **(c)** All aliphatic and aromatic residues. Red sphere marks a binding site water molecule. Phenylalanines are coloured orange-red; tyrosine - green, whilst TYR120 is additionally coloured by heteroatom; alanine - black; cysteine - yellow; isoleucine - purple; leucines - cyan; methionine - violet-red; and tryptophan - pink.

and egress. The bottom of the pocket is padded by an additional two counterforts in the form of a tryptophan and tyrosine residue. The sidechains lining the central tier are more varied in identity and distribution - a matrix of aliphatic and aromatic residues that act to corral any bound ligand within the pocket.

1.3.2. An atypical binding signature

An early thermodynamic study examined the binding characteristics of several SBT analogues varying in the size of their alkyl functional group. Despite the negative ΔC_p obtained for ligands binding to MUP-I (henceforth referred to as MUP), the expected entropic signature associated with the hydrophobic effect was not observed. Instead, the enthalpic term dominated binding and was approximately 50% more favourable than the entropy. This was unexpected, as the association of hydrophobic solutes in solution is usually believed to be an entropic process, abetted by ligand desolvation and the expulsion of ordered water molecules from the binding pocket (§1.1.6). Solvent transfer of solvated SBT ligands into cyclohexane confirmed that the desolvation enthalpy should indeed be unfavourable [70]. This unusual behaviour was also corroborated by studies on the binding of a panel of n-alkanols and pyrazine derivatives (IPMP and IBMP) to MUP [36,89,90]. The exact reason for the discrepancy between ΔC_p values and the binding signature could not be deconvoluted by examining global ITC values alone and various other techniques were utilised.

High resolution crystal structures indicated that there are very few waters present in the occluded binding site. The apo protein only contains 4 waters clustered near TYR120, whilst complexed structures contain 0 to 3 water molecules, depending on the identity of the ligand bound [36,72,90]. Short ~10 ns MD simulations indicated that the density of water within the occluded binding site is very low (0.2 to 0.3 g/cm^3). Artificially flooding the calyx with water resulted in their expulsion and a return to a state of minimal hydration as the simulations progressed [90]. Low levels of water density were also observed in longer MD simulations of 1.2 μs [91].

Published crystal structures indicate that bound ligands usually form a single hydrogen bond via a bridging water molecule (e.g. SBT and n-alkanols) or directly to TYR120 (e.g. IBMP) [36,70,90]. The lack of suitable hydrogen bond donors and acceptors in the dewetted cavity means that waters are disordered because they cannot maintain a stable hydrogen bond network. Computational work on the binding of n-alkanols indicated that increasing ligand size merely resulted in the waters repositioning themselves within the calyx [36,90]. The possibility that the low dielectric environment in the sub-solvated cavity could strengthen electrostatic interactions such as hydrogen bonds was investigated by Barratt et al. (2005) by mutating TYR120 to a phenylalanine [90]. It was found that the binding enthalpy was reduced by ~12 kJ/mol, whilst the entropic term partially compensated this loss by ~7 kJ/mol. As the dominance of enthalpic binding signature (-31.44 kJ/mol) was undiminished, this hydrogen bond was not thought to be a predominant factor for the observed phenomenon [90,92]. However, given that the enthalpy linearly scales with ligand surface area and hydrogen bonding is relatively consistent between different ligands, a strong dependence on van der Waals interactions was thought likely [36,70].

Ross et al. (1981) put the "classical" hydrophobic effect to question through the observation that the binding of some protein-ligand complexes are driven by the enthalpic term. e.g. α-chymotrypsin and lactic dehydrogenase [93]. They advance a two stage model, the first of which has thermodynamic characteristics directed by solvent reorganisation effects, while the last stage is more influenced by solute-solute interactions. Whereas the hydrophobic effect posits that the exchange of solute-solvent for new solute-solute interactions during ligand binding roughly cancels; the model provides examples why this may not be the case. Examples include the optimisation of van der Waals contacts via aromatic stacking interactions, strong hydrogen bonds, ionic interactions and salt bridges. Additionally, these effects could all potentially be amplified by a low dielectric environment [93].

As MUP is suboptimally hydrated, favourable dispersion interactions at the protein-ligand interface are not offset by the compensating enthalpic and entropic terms associated with displaced binding site waters rejoining bulk. Thus, it was proposed that the enthalpic binding signature must arise from new interactions made during the second stage [36,90].

1.3.3. Thermodynamic Decomposition: The binding of n-alkanols to MUP

The binding of a simple n-alkanol panel (pentan-1-ol to nonan-1-ol) was investigated by Malham et al. (2005) using ITC and illustrates how the entropy and enthalpy of binding vary in a linear fashion upon addition of a methylene group (**Fig.1.23.a**) [36]. Using the decomposition methodology outlined in §1.1.5, the solvation term was factored out via the use of available experimental solvation data for primary alcohols to yield values for $\Delta G°_i$, $\Delta H°_i$ and $T\Delta S°_i$ [94]. In the absence of any solvent effects, additivity is observed upon hydrocarbon chain extension, yielding a $\Delta H°_i$ of -8.4 kJ/mol; and a $T\Delta S°_i$ of -5.5 kJ/mol (**Fig.1.23.b**). Notably, $\Delta\Delta H$ and $\Delta\Delta S$ values were relatively constant between different members of the panel. Caveats associated with this method are that the number of bound waters displaced on ligand binding must be the same between all ligands within the panel. Additionally, the protein is assumed to make the same contribution to the thermodynamics of binding for all ligands in the panel and thus its contribution is considered to be zero. Despite these reservations, the favourable $\Delta H°_i$ term correlates well to the value of -6.9 kJ/mol per methylene returned by theoretical bead models of alkanes projected on a 3D-lattice. Furthermore, experimental values (-6 to -7.5 kJ/mol per methylene) show a relationship between the length of linear molecules and their molar cohesive energy [95]. In entropic terms, the intrinsic value is comparable to the ~6 kJ/mol energetic cost associated with restriction of another torsional rotor upon extending carbon chain length, Thus, the working hypothesis codified by this work stated that increasing the size of a small n-alkanol by a single methylene, results in an $\Delta H°_i$ gain on binding due to increased protein-ligand van der Waals contacts. This favourable term is partially compensated by the accompanying $T\Delta S°_i$ associated with restricting an additional rotor [36].

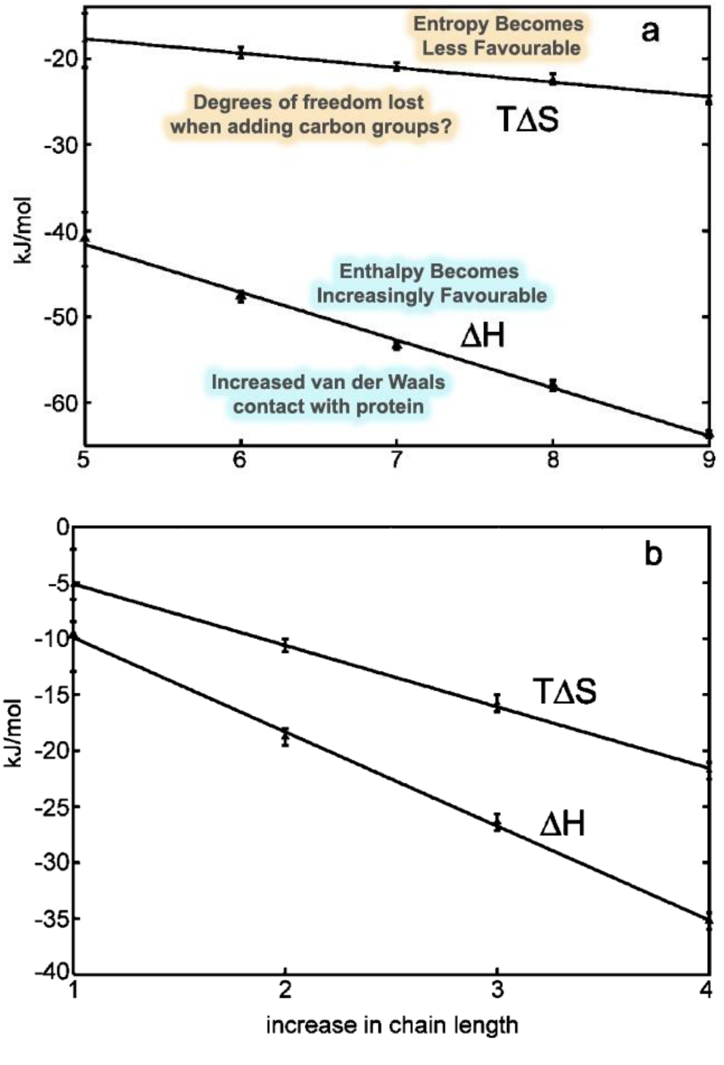


Fig.1.23. (a) Global ITC Enthalpies and entropies of binding of pentan-1-ol to nonan-1-ol to MUP, plotted against length of carbon chain. Figure adapted from reference [36]. **(b)** Difference in intrinsic enthalpies & entropies between successive ligands in primary alcohol series. Figure taken from reference [36].

1.3.4. Protein contribution to binding and distal residue dynamics

The protein is formed from the folding of a linear polypeptide chain into a complex folded structure. Its dynamics are subtly affected by motions propagated between amino acid units close to one another in terms of both sequence and their spatial positioning within 3D space. In order to deconvolute the protein contribution to the global entropy of binding, the NMR spectroscopy derived order parameters (S^2) can be converted to yield per-residue conformational entropies using the method described by Yang et al. (1996) [96]. The order parameter gives a measure of the dynamics of methyl and amide bond vectors from relaxation data and is subject to the following caveats. Firstly, this is a 1st order estimate as no account of correlated motions between bond vectors is taken into account and thus the entropies obtained represent an upper limit. Only timescales in the picosecond to nanosecond timescale can be reliably measured. Finally, the method assumes the motional model for the configurational entropy is invariant between protein free and bound states [34,97].

Bingham et al. (2004) used NMR spectroscopy relaxation measurements to assess configurational entropy differences in backbone amide and side-chain methyl groups between the apo and complexed state [89]. Data for the ligands, IPMP and IBMP are tabulated in **Table.1.3**.

	IPMP	IBMP
ITC TΔS°	-9.4 ± 0.9	-10.7 ± 0.5
Backbone Amide	-	-7.4 ± 6.5
Sidechan Methyl	-0.8 ± 3.8	-3.4 ± 2.8
Amide plus Methyl	-	-10.8 ± 7.1

Table.1.3. Comparison of ITC data to summed entropies obtained from NMR spectroscopy order parameters [89].

Despite being unable to obtain order parameters for all the bond vectors, the entropic sum of IBMP methyl's and amides comes close to the global ITC values. Obtaining entropies on a per-site basis indicated that the expected loss of movement on binding a ligand was accompanied with an increase in mobility of residues distal from the binding site. This phenomenon was described as a "conformational relay" (or entropy-entropy compensation) and served to offset part of the entropic penalty typically associated with binding. Increases in mobility were primarily associated with the flexible loop regions [89]. This is not entirely surprising as there are numerous examples from directed evolution of how mutating residues distal to the binding site improves catalytic efficiency dramatically [98,99]. It is not immediately obvious what roles residues distant from the binding site play in structural reinforcement, and assessing their role when rationally designing ligands is important.

The experimental results for IPMP and IBMP are at variance with relaxation data and computational studies obtained for SBT, as these show an increase in protein dynamics on ligand binding [100–102]. The reason for this may be due to differential protein dynamics dependant on the identity of the ligand bound. Additionally, the caveats associated with calculating order parameters, coupled with the fact that only subsets of all possible protein bond vectors were measured, render this an approximate method subject to considerable error.

1.3.5. Overview of why MUP is considered a model system

MUP is considered a good model system to study EEC and other thermodynamic effects because of its ostensible simplicity.

1. It binds promiscuously to a variety of hydrophobic ligands. Hence, it is possible to assess the binding characteristics of panels of ligands, whose members are designed to possess incremental differences in structure. e.g. a panel of n-alkanols allows an evaluation of the thermodynamic cost of adding a single methylene group.

2. Binding is dominated by apolar interactions and only a single directional hydrogen bond donor is thought to be present in the calyx. The former allows recognition of multiple ligand partners, whilst the latter simplifies thermodynamic assessments as hydrogen bond strengths are distance and orientation dependant.

3. The binding pocket of MUP is suboptimally hydrated and this feature allows easier appraisal of the solvation term.

4. MUP is a relatively small, well-behaved protein (< 20 kDa) which is easily overexpressed in *E.Coli.* Its characteristics can be probed via MD simulations and a variety of experimental techniques. e.g. NMR spectroscopy, ITC, X-ray crystallography, etc.

1.4.0. Molecular Dynamics: Quick or accurate?

Molecular dynamics is an in silico method that simulates the physical motion of atoms and molecules over a period of time. This dynamic time course is known as a trajectory because the trajectories of a many-body ensemble are determined by solving Newton's equations of motion for the system. The simulation can be viewed like a movie and because the motion of molecules is generated in a time-dependent manner, non-equilibrium thermodynamics and processes (e.g. ligand binding, transport phenomena, etc) can be studied [95,103,104]. In conjunction with statistical mechanics, information about thermodynamic quantities such as free energies, entropies, enthalpies, pressure, volume, temperature, etc can be obtained. Statistical mechanics is a branch of probability theory that allows macroscopic thermodynamic quantities to be generated via averaging microscopic, instantaneous values of a population. When simulating large systems, amassing a large enough ensemble that accurately represents the population distribution of various microstates of the system is a common problem and statistical accuracy has to be balanced against computational cost [21,103,105 106].

Simulation of a typical biochemical system pertinent to drug design typically contains a protein receptor and a ligand surrounded by a box of solvent such as water. The initial structural data is usually obtained by techniques such as NMR spectroscopy or X-ray crystallography. The atoms that constitute the solvent, protein and ligand are restrained by rules that define the range and degree of permissible motion. The calculation and simulation of water incurs the greatest source of computational expense.

One of the more rigorous methods of achieving this is via *ab initio* MD. Here, interactions between electrons are calculated "on the fly" using knowledge about quantum mechanics. The chief advantage is that many of the approximations and presumptions seen with the other methods are avoided. Thus, this method is useful to describe systems whose behaviour cannot be predicted from first principles, and to represent chemical moieties yet uncharacterised by force field methods (*vide infra*). However, it does not accurately deal with dispersion forces without incurring further computational overhead and also uses some approximations such as Density Functional Theory (DFT) to deal with the many-electron problem and the Born-Oppenheimer approximation to simplify calculation of the Schrödinger equation. In practical terms, the length of time taken for these simulations often means that only short timescales are investigated [105–109].

"Classical MD" obtains an increase in speed through the use of mathematical functions known as force fields to approximate the potential energy of the system. A "fixed-charge" model is used to describe the electrostatic polarisation of a molecule, thus avoiding the computational penalty associated with calculating how the molecule's charge distribution evolves with time. This approximation means that the molecule's polarity is not modulated in response to its environment. Despite this proviso, force fields are capable of providing results comparable to other more rigorous methods in less time. Parameterisation of a molecule is usually accomplished by quantum mechanical calculations and the associated computational cost is ameliorated by a key tenet of MD: transferability. This means that certain fundamental parameters do not have to be recalculated when modelling a new molecule as their behaviour has already been quantified. For example, most CH bond stretching and angle bend values do not exhibit large differences between molecular species. Additionally, assignment of atom types allows atoms to be further differentiated according to their hybridisation state and chemical properties. e.g. Carbon atoms located in aromatic groups versus those in alkyl groups.

A typical force field serves to package descriptions of bonded and non-bonded terms in a functional form to capture inter and intramolecular interactions (**eqn.1.11**). Bonded terms collate contributions to the potential energy by summing bond stretching, torsional rotation and angle bends. Non-bonded interactions usually encapsulate van der Waals and electrostatic contributions, and are calculated via application of the Lennard-Jones (LJ) potential and Coulomb's law respectively.

$$v(r^N) = \sum_{bonds} \frac{k^i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k^i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\,\omega - \gamma)) \\ + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right) + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

eqn.1.11. example of a classical MD force field taken from reference [103].

Force fields are continually refined by empirically fitting the potential energy function against experimental data (e.g. NMR spectroscopy; X-ray/neutron diffraction; vibrational spectroscopy; etc) and theoretical, quantum mechanical methods [95,103,104,110].

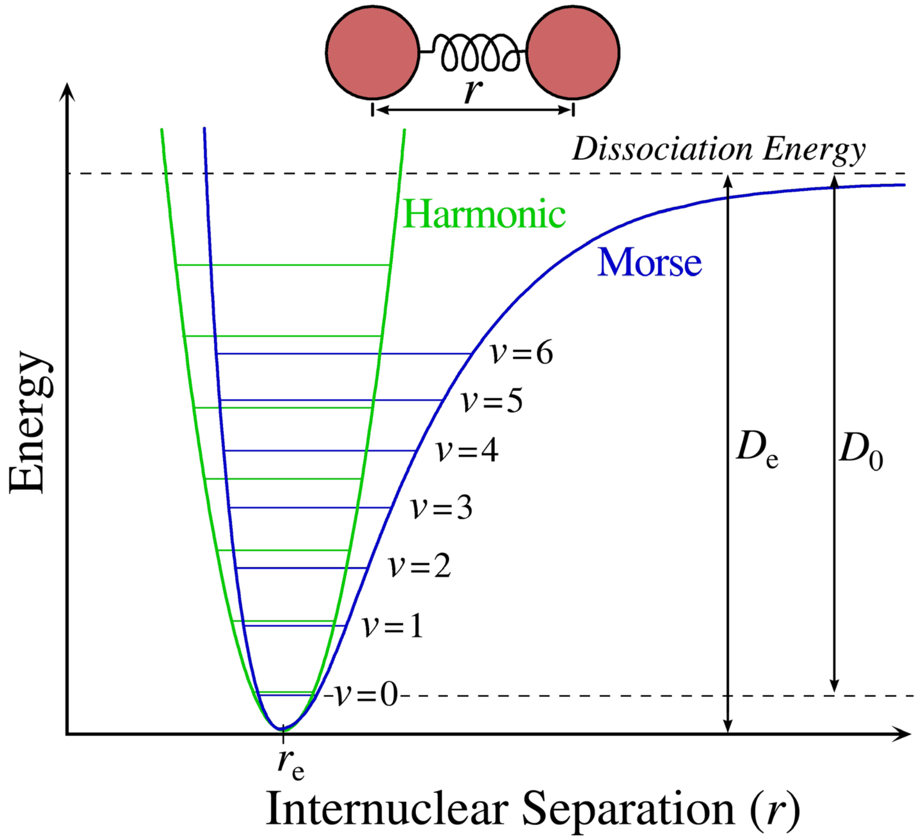


Fig.1.24. Harmonic oscillator potential (green) compared to Morse potential (blue). Image taken from reference [111].

Most currently implemented force fields use a harmonic oscillator to describe the interatomic distance for systems near equilibrium. This method aids computational speed and is a good approximation as bonds generally stay around their equilibrium values. A limitation with this approach is that bond breaking cannot be simulated. A better model that accounts for bond cleavage is known as the Morse potential (**Fig.1.24**). However, this is computationally more intractable and has the additional overhead of each bond requiring three terms to describe it. Thus, it is not generally implemented [105,112,113].

There are various MD software implementations available such as CHARMM (Chemistry at Harvard Macromolecular Mechanics) and AMBER (Assisted Model Building with Energy Refinement). These suites are capable of working with a variety of different force fields [103,105,113].

1.4.1. Modelling water

MD can explicitly model every water molecule, or use implicit models which speed up the calculation significantly. The latter is also known as a continuum model as it treats water as a continuous entity, as opposed to simulating the interactions of discrete water molecules. Early continuum models such as the Poisson-Boltzmann (PB) calculated the molecular electrostatic potential and the change in solvent polarisation on binding. As they neglected to factor in the hydrophobic effect, an additional surface area term was added to create the PBSA model. This operates by adding an unfavourable solvation energy term to solute possessing a greater amount of exposed surface area. This negative term can be minimised via solute clumping or binding (in the case of a protein-ligand interaction), in order to simulate the hydrophobic effect. The Generalised Born (GB) is another notable implicit model that is designed as a faster alternative to the PB model. It also has a (GBSA) version that factors in surface area [103,112,114].

Though implicit models are faster than explicit ones, they neglect to account for water molecules that fulfil unique functional roles within locations such as the binding pocket. These waters may be resident for a greater than average length of time, make large contributions to the free energy of binding, or assist ligand binding by acting as a bridging molecule. Ideally, the descriptive quality of a single explicit water molecule should scale up to accurately simulate the characteristics of bulk water. e.g. density, surface tension, self-diffusion coefficient, the hydrophobic effect, dielectric constant, etc. However, this is not always the case and models of water tend to excel in different areas depending on the training sets and criteria used to parameterise them [42,73,115]. There are several rigid, fixed charge water models that vary in bond geometry, charge distribution and how well they reproduce the properties of water. One of the pertinent differences is the number of interaction sites utilised in different models and the associated impact on computational cost. In the TIPnP family (transferable intermolecular potential n point, where n is 3, 4, 5 or 6) a positive charge lies on the hydrogen atoms and placement of the negative charge is variable depending on model. Some models place this charge on, or slightly offset the oxygen atom, whilst others model it as lone pairs (**Fig.1.25**).

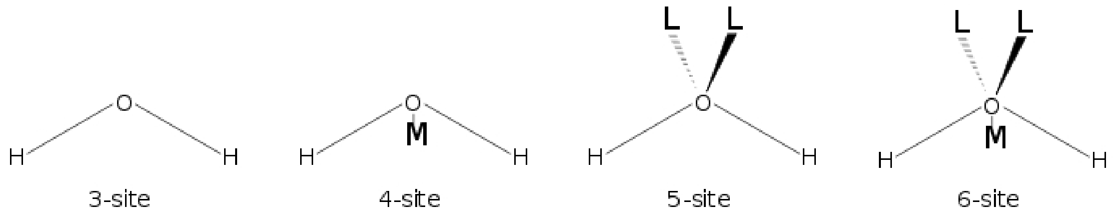


Fig.1.25. From left to right, water models TIP3P, TIP4P, TIP5P and a six site model. The point marked M is the offset location of the negative charge in the 4 and 6 site model. This is purported to improve the electrostatic distribution around the molecule. Image taken from reference [116].

The computational cost increases with the number of sites on the model. For example, 9 interactions (every site on one molecule with every site on another) are calculated for a pair of TIP3P molecules. There should be a cost of 16 interactions for TIP4P, but this can be reduced to 10. i.e. 9 Coulombic relationships and a single estimation of the O-O distance to ascertain the LJ interaction. Using the same approximation, a TIP5P water dimer can be modelled with 17 interactions instead of 25. Recent evaluations of the different TIPnP models based on their ability to reproduce the properties of bulk water, nominated TIP4P/2005 as the best rigid model in all categories tested, bar dielectric constant [73,115]. It is important to reiterate that it is extremely difficult to find any fixed point model that can even begin to reproduce the many properties of water simultaneously. Every model is designed to fit certain experimental criteria and so the experimenter should choose a model best suited for the application at hand.

It is believed that large improvements in the simulations of water can be made by moving away from fixed point charges towards polarisable and flexible force fields. There are a variety of methods of executing this with varying degrees of (large) computational cost. The AMOEBA (Atomic Multipole Optimised Energetics for Biomolecular Applications) force field is one such model. The water model is flexible and replicates the target properties exceedingly well. The force field deals with both inter and intramolecular polarisation and can better deal with the subtleties of hydrogen bonding and other electrostatic interactions such as van der Waals interactions, ionised side chains or ligands, and movement of polar groups from bulk solvent into the binding site, etc [117–119].

1.4.2. Class I vs. Class II Techniques

Computational methods that determine details about protein-ligand binding affinities can be divided into two broad classes. Class I techniques are recognised as being the slower, more rigorous methods. To achieve better resolution of parameters, they generally use explicit solvent and tend not to limit conformational DOF available to the system. Hence, they are more suitable to acquire precise values for the key thermodynamic quantities of interest. Examples include *ab-initio* and Thermodynamic

Integration (TI) methods.

Ideally, computer based screening should allow the *in silico* analysis of multiple drug candidates in a manner that is quicker and cheaper than carrying out the same tests in the lab. Thus, class II techniques (e.g. molecular docking or posing) are optimised for high throughput analysis and concentrate on the speed of calculation to generate "hits" from vast libraries of drug-like compounds. In order to do this, there exists a trade-off in terms of increasing the speed of the calculation, at the expense of accuracy. Rigid body approximations freeze solute (protein and ligand) DOF, whilst solvent complexity is minimised via the use of continuum models. Due to the loss of entropic contributions, a true estimate of values such as the free energy is not possible. Rather, consensus schemes use a variety of different scoring systems to rank ligands in order of relative binding affinity. To redress inaccuracies created by these approximations, methods have been developed that generate a variety of "rigid body snapshots" from the dynamic trajectory of a target protein from MD simulations and screen potential ligands against these alternate conformations [114,120].

1.4.3. The evolution and revolution of simulation hardware

Biologically relevant simulations usually deal with a range of timescales depending on the system being studied. Apart from the solvent, protein dynamics are the second largest contributor to the size and tractability of the calculation.

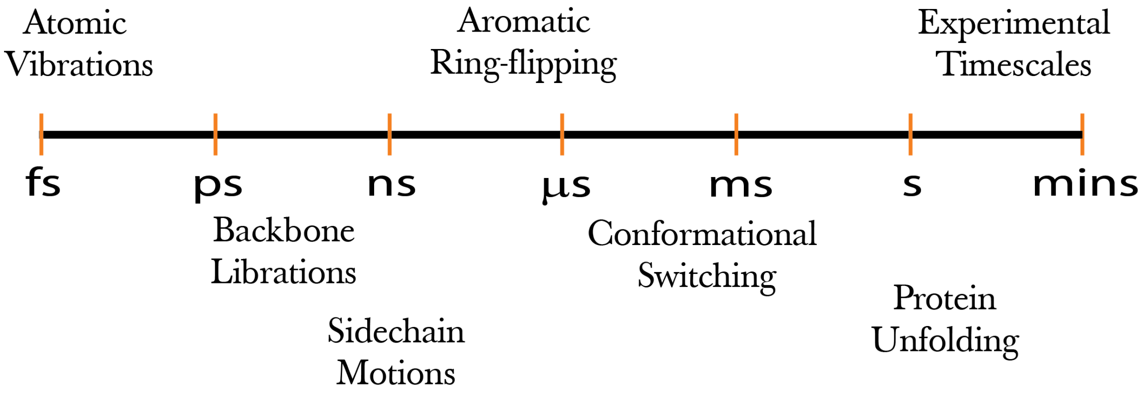


Fig.1.26. Timescale ranges of protein dynamics

The time scales in which various biologically relevant motions are observed are depicted in **Fig.1.26.** A limiting factor in simulations is the time step. This is the basic unit of simulation time at which the forces between atoms and molecules are calculated. Through necessity, this is limited to a period less than the fastest atomic vibrations (~1 fs) in the system to avoid simulation instability. The SHAKE Algorithm removes DOF by freezing rapidly moving bonds involving hydrogen. This allows the time step to be doubled and the calculation is thus accelerated [121,122]. Despite this, CPU-based calculations typically yield simulations in the hundreds of nanoseconds, whilst research covering the microsecond to millisecond range is rarer. For example, simulations of MUP consist of ~27,000 atoms including explicit water molecules. A single 2 fs time step takes ~72.7 s to calculate and thus a single 100 ns simulation takes ~8 days of real world time to complete. When multiple protein-ligand complexes and the requisite repeats are factored with queuing time for shared computational resources, the time taken for these calculations can extend into months.

The availability of parallel processors allows rapid calculation and access to longer timescales. However, in practice there are an optimal number of processors for any given system. This is because scalability is limited by the MD code, CPU interconnect speed and latency. Thus, benchmarks must be conducted to establish the best settings for a given computer configuration. The hardware used to run MD simulations is constantly evolving and, until recently, researchers would have had to pay substantial amounts of money to purchase the necessary equipment or have access to large computer clusters whose resources are shared. Roughly 40% of the equilibrium simulations run in this thesis were run on the Leeds arc1 and N8 polaris clusters. The former uses 4,512 processors, 8.5 Tb RAM and an 115 Tb parallel file system, whilst the latter consists of 5,504 cores, 25 Tb RAM and has a 175 Tb parallel file system. Theoretical peak performance is measured at 50 and 110 teraFLOPS (TFLOPS) respectively. A teraflop is a measure of computer performance and stands for FLoating-point Operations Per Second. Thus a TFLOP equates to 10^{12} FLOPS and equates to a human (e.g. a hirstute, yet erudite professor) doing a calculation per second for 31,688.77 years to match what such a cluster can do in a single second. As a further reference point, a 2.5 GHz desktop processor typically delivers performance at a rate of 0.01 TFLOPS. Simulations of the dihydrofolate reductase (DHFR) benchmark on such clusters typically achieve a speed of ~12 ns/day.

There have been two significant game changers in the field of MD. The first is the creation of the Anton supercomputer by DE Shaw research, which specifically runs MD simulations. The machine is composed of 512 nodes, each formed by a single Application-Specific Integrated Circuit (ASIC) with instructions hard-coded onto the chip. Performance is in the multi-petaFLOP region and ~17,000 ns/day can be calculated against the DHFR benchmark [123–125]. However, access to such powerful resources are limited by a run time of ~8 days and competing applications are only accepted from principal investigators affiliated with US academic or non-profit organisations [125]. Whilst the terms controlling access to these finite resources are generous, they inevitably widen the gap between researchers fortunate enough to qualify and those that do not. The second game changer redresses this balance by reducing the cost of running microsecond length simulations to a level affordable by the *hoi polloi*. This is the advent of GPGPU (General-purpose Computing on Graphics Processing Units).

Initially, the job of rendering computer graphics on the screen was handled by the CPU. However, computers designed from approximately the 1990's onwards, offloaded this job to a specialised, dedicated device known as the Graphics Processing Unit (GPU) [126]. The hardware evolution of these specialised devices was partly driven by the consumer gaming industry and its quest to simulate photorealistic effects, whilst simultaneously avoiding the depths of the "uncanny valley" (**Fig.1.27**). This concept was conceived in the field of robotics by Masahiro Mori and describes the feelings of aversion and discomfort humans have in response to depictions or forms recognised as being different from themselves. e.g. A computer generated person walking with an incorrectly programmed gait. The term valley is used because, after a certain point, the feeling of the "uncanny" shifts back to familiarity as the object becomes less recognisable as human. e.g. a cuddly toy [127,128].

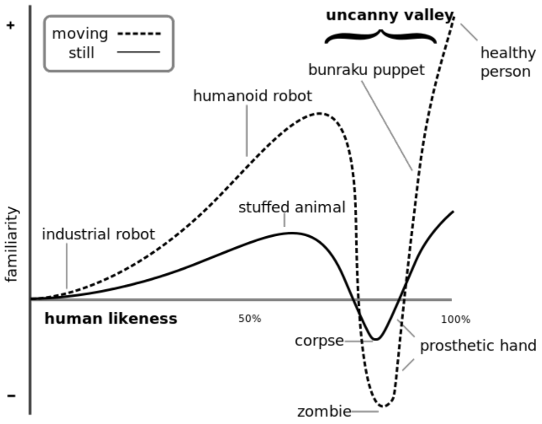


Fig.1.27. Depiction of uncanny valley taken from reference [129].

In order to simulate the levels of realism required, GPUs were engineered with multiple specialised processors that worked in parallel to process pixels and vertices. To increase performance and cope with innovations created by rapid development cycles, these units were gradually replaced by more general purpose programmable units that could fulfil a variety of roles. The company NVIDIA brought GPGPU to the scientific developers through the development of CUDA (Compute Unified Device Architecture), a platform that supports commonplace programming languages (e.g. C, C++, FORTRAN, etc) and thus made knowledge of specialised graphics programming languages unnecessary.

Traditionally, compute for scientific and engineering applications relies on expensive 64 bit double precision (DP) floating point arithmetic that can represent numbers to 15-17 decimal places, whilst 32 bit single precision (SP) can only manage 6-9 decimal points [130,131]. In the context of MD simulations, the rounding errors caused by SP arithmetic can cause significant deformations of protein structure and dynamics. Hence, NVIDIA marketed a line of expensive Tesla graphics cards (Fermi architecture) capable of rapid DP calculations to the scientific community labelled with a price tag in the thousands of pounds. Whilst consumer grade GeForce cards had negligible DP performance, professional GPUs could calculate DP arithmetic at ~1/2 the rate of SP. The AMBER suite of programs accelerated speed by developing a stable, mixed single and double precision (SPDP) model, wherein nonbonded forces are calculated with SP, whilst more critical bonded terms and force accumulation are dealt with by DP arithmetic. When the Kepler architecture superseded Fermi, SP performance increased dramatically at the expense of DP. Consequently, an even faster single precision, 64 bit floating point integer (SPFP) model was developed. It used 5 extra SP data operations to combine a 32 bit integer and a SP float, to form a 48 bit pseudo-DP structure at the expense of a few significant figures. This model has been extensively benchmarked and found to have the same accuracy as the SPDP model [132,133]. Synchronously with these software developments, the hardware technology filtered down to consumer grade devices. As NVIDIA's 28nm chip fabrication process suffered from poor yields, chips that did not meet quality control standards were repackaged as cut-down models with an affordable price tag. Typically, these sub-performing chips have some of their processing power or other features disabled (**Table.1.4**).

	Stream Processors	**SP (TFLOPS)**	**DP (TFLOPS)**	**DHFR (ns/day)**	**Cost**
Tesla K40	2,880	4.29	1.43	81.4	£3,500
Tesla K20	2,496	3.52	1.17	65.7	£2,500
GeForce TITAN	2,688	4.50	1.30	84.0	£950
GeForce 780	2,304	4.00	low	77.1	£500
GeForce 680	1,536	3.00	low	59.9	£400

Table.1.4. Performance statistics of various Tesla and GeForce cards.

As can be seen from **Table.1.4,** the price to performance ratio of purchasing a GeForce 780, coupled with the AMBER SPFP model, makes this an attractive card for long MD simulations. With a quad GPU setup, a researcher has access to ~16 TFLOPS of unshared processing power. This equates to 32% and 15% of the power of the arc1 and polaris clusters respectively for a cost of around £3,000 or less. While a specialised supercomputer like Anton can achieve a far greater rate of calculation compared to GPUs, there are three important factors that make GPU computing more significant. Firstly, microsecond long calculations are available to the majority of scientists on a ubiquitous hardware platform at a fraction of the cost of a typical cluster. Secondly, GPU computation coupled with enhanced sampling methods such as accelerated Molecular Dynamics (aMD), achieve greater conformational sampling than millisecond long conventional MD simulations in less time [134,135]. Finally, researchers can rapidly prototype and test hypotheses, which would previously have been agonised over due to the cost of computational resources and time. This forms the basis of promising new avenues of research in both computational and experimental terms that might not have otherwise been explored.

1.5.0. Aims and scope

MUP is an ideal model system in which protein-ligand interactions can be interrogated, partly due to its ostensible simplicity and also due to its affinity for a large number of hydrophobic ligands. Some key points that will be examined by this thesis are:

1. Experimental thermodynamic decomposition on the binding of primary alcohols to MUP has indicated that the enthalpic contribution to binding is favourable, whilst the entropic portion is unfavourable. These terms change in a linear fashion upon increasing the length of the ligand by a methylene group. The key question considered is whether the intricacies of the thermodynamics of binding can be captured by *in silico* techniques, so as to allow the better rational design of drugs. This question is rather broad and can be further subdivided into queries that are pertinent for MUP.

2. MUP has an atypical enthalpically driven binding signature. Do molecular dynamics simulations have sufficient depth of detail to capture this subtlety?

3. Do the limited number of binding site waters observed in MUP's suboptimally hydrated calyx act to optimise van der Waals contacts?

4. As the entropic penalty is thought to be mainly contributed by the ligand, can this value be computationally recreated and does it match the experimental values?

5. Does pre-organisation of the ligand mitigate entropic losses typically seen upon binding?

6. A key assumption in many of the thermodynamic decompositions of MUP is that the ligand loses a significant amount of translational and rotational entropy when bound. However, other works have noticed significant translocation of the ligand in the pocket. So this work will seek to ascertain whether simulation data can replicate this, and accurately quantify the nature and amount of ligand translocation?

7. How does the ligand gain access to the protein binding cavity and what is the response of bound waters to ligand internalisation?

8. Previous research has suggested that the protein does not greatly contribute to the global thermodynamics of binding. Can this be verified via *in silico* analysis?

An overview of the chapters that follow are presented below:

Chapter 2.0: Thermodynamic Integration: In this chapter an attempt is made to obtain all three thermodynamic values of interest (ΔH, ΔG and TΔS), and capture the enthalpically driven thermodynamic signature of MUP.

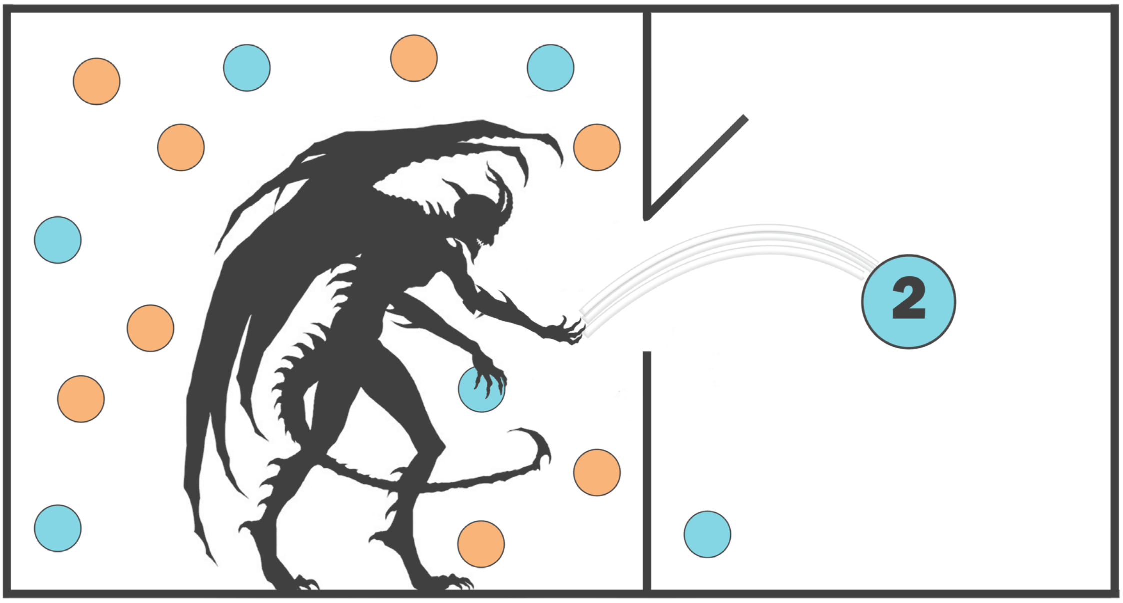
Chapter 3.0: Ligand Conformational Entropy: A method to quantify the amount of conformational entropy lost on ligand binding is employed. This is applied to the question of whether the entropy decreases in a linear manner with the addition of another CH_2 rotor. Furthermore, the impact of improving the conformational entropy on binding is examined via analysis of additional panels of ligands that have restricted rotors.

Chapter 4.0: Ligand Translational & Rotational Entropy: The translational and rotational contribution of the ligand to the global entropy of binding is assessed via the development of two new methods.

Chapter 5.0: Ligand Internalisation & Desolvation: A MD protocol is developed whereby ligand's movement from bulk solvent to the binding cavity can be charted and studied to reveal mechanistic details about the internalisation process. This method also allows an assessment of the contribution made by ligand desolvation and the expulsion of bound waters to the global entropy of binding.

Chapter 6.0: Ligand Mediated Modulation of Protein Conformations: The change in the protein entropy on binding different ligands is assessed on a per-residue basis. This allows a granular analysis of the protein's response to ligand binding and allows further mechanistic conclusions to be reached.

Chapter 7.0: Conclusion: Summary and conclusions.



Maxwell's Daemon

Chapter 2.0: Thermodynamic Integration

2.1.0. Recreating experimental data via an in silico method

MUP is a particularly interesting protein to study because of its atypical binding signature, promiscuity in choice of binding partners, ostensible simplicity and additive affinity upon binding ligands of increasing size. The key hypothesis postulated by Malham et al. (2005), was that binding interactions of MUP were dominated by weak solute-solute dispersive interactions [36]. This is because the magnitude of the entropic term commonly expected in hydrophobic association was greatly diminished due to suboptimal hydration of the binding site. An experimental decomposition method performed on a panel of primary alcohols suggested a structural rationale behind these observations. The favourable intrinsic enthalpic contribution (-8.4 kJ/mol) to binding affinity was linear on extension of the carbon chain length by a single methylene group. Thus, this was primarily responsible for an increase in ligand-protein dispersive interactions, which is commensurate with increases in ligand surface area. At the same time, an intrinsic entropic penalty partially (-5.5 kJ/mol) offsets the enthalpic gain due to the cost of immobilising the additional C-C rotor. This term is chiefly responsible for the unfavourable global entropy, because the large favourable contribution gained from solvent reorganisation on displacing binding site waters is non-existent (§1.3.2 to 1.3.5 & **Fig.1.23)** [36,90,136].

Whilst, the experimental evidence is cogent, it is of value to establish whether these observations can be replicated *in silico* for three main reasons. Firstly, the experimental conclusions are the result of a "top down" decomposition of global thermodynamic values and have a number of associated caveats. Though computational approaches have their own shortcomings, they are of a different nature and thus offer an independent source of validation. The evisceration and dissection of global thermodynamic observations is of particular value to computational methods, because the system necessarily has to be built from the "bottom up". The "decomposition values" obtained possess a finer level of granularity and are a natural by-product of this process. Secondly, most experimental work is expensive compared with the cost of running simulations. The drug discovery process would benefit greatly from computational methods that better assist at all stages within the pipeline, from rapid HTS through to precise rational design based upon knowledge of the fundamental thermodynamic interactions that regulate bi-molecular binding. Knowledge of these interactions and how they vary from system-to-system form the foundation of designing better force fields and improving methods to quantify the binding affinity of putative drug-like molecules. Finally, techniques that offer atomistic level detail provide an understanding of the mechanics and parameters that govern protein-ligand binding. This chapter investigates the ability of the Class I technique known as Thermodynamic Integration (TI) to capture the essential thermodynamic trends behind the binding of congeneric panel of ligands to MUP with respect to the free energy, enthalpy and entropy. TI can be used to calculate the change in thermodynamic parameters resulting from small changes in structure of the ligand or the protein. As a rigorous method that eschews many of the approximations that other Class II techniques accept, it is hoped that trends very close to the experimental values will be captured.

2.1.1. Objectives

A number of key aspects of the TI method were examined to ascertain its suitability to discern the key thermodynamic quantities of interest. The first question analysed was whether the AMBER force field, coupled with the TI methodology, possessed sufficient accuracy to capture the free energy difference between 2 ligands differing incrementally in structure. e.g. One methylene group between hexan-1-ol and heptan-1-ol. Whilst obtaining the free energy difference between two ligands may be difficult, replicating a congruent pattern across a congeneric panel is even more challenging, with accuracy having to be counterbalanced by the associated computational "cost" of the method. Thus, the technique was applied to more than one transformation, in order to check its ability to capture experimental trends. The following transmutations were considered:

1. hexan-1-ol to heptan-1-ol

2. heptan-1-ol to octan-1-ol

According to Malham et al. (2005), binding site waters are postulated to enhance dispersion forces via the optimisation of interfacial solute-solute interactions between protein and ligand [36]. This hypothesis was tested by examining the following transmutations:

1. hexan-1-ol to heptan-1-ol (with two crystal binding site waters)

2. hexan-1-ol to heptan-1-ol (without binding site waters)

As the value for final free energy difference is "built up" by calculating fundamental interactions between atoms, the accuracy of the method was not only scrutinised in terms of its ability to reproduce the relative $\Delta\Delta G_{TI}$ values, but also with regards to how well it was able to capture the experimental thermodynamic signature. The latter describes the domination of van der Waals interactions over that of electrostatics upon addition of a methylene group.

In the second analysis, the entropy difference was calculated using a finite-difference method (FDM). This entailed running TI calculations at three different temperatures (278, 300 & 322 K) for every transmutation. The enthalpy was then calculated using the relationship $\Delta G° = \Delta H° - T\Delta S°$. As TI performance is usually measured only in terms of the veracity of the double free energy difference ($\Delta\Delta G_{TI}$), the additional requirement to successfully predict component entropic and enthalpic trends required a further level of rigour.

2.1.2. The principle behind Thermodynamic Integration

The TI calculation is performed on data obtained from a rigorous MD simulation that makes use of explicit solvent and does not sacrifice solute degrees of freedom to save on computational time. As a consequence, the free energies computed are more likely to accurately represent the entropic and enthalpic contributions [114]. The core concept behind TI is the calculation of the free energy difference between two closely related states, A and B. In practical terms, this can be a comparison of ligands with small disparities in structure binding to a certain protein. e.g. the ligands hexan-1-ol and heptan-1-ol differ by a single methylene group. The free energy difference can be calculated by applying Hess's law of constant heat summation which states that 'the energy change for any reaction is independent of the path taken or number of steps required' [21]. As energy changes are state functions, an "unphysical" reaction between A and B states can be computationally simulated to create a reaction cycle similar to that seen in Hess and Born-Haber cycles (**Fig.2.1**).

The two states can be differentiated from each other on the basis of the identity of the ligand involved. i.e. State A contains ligand A, whilst state B involves ligand B. **Fig.2.1** can be visualised more easily as two discrete simulations. The lower half (Y) measures the change in free energy as ligand A transmutes into ligand B whilst free in solution. *Per contra*, the upper half simulates the same ligand transformation process whilst bound to the protein. It is for this reason that this process has been termed computational alchemy. This can be formalised to yield the following equation:

$$\Delta G(A)_{Bind} - \Delta G(B)_{Bind} = \Delta G(Y)_{Tmut} - \Delta G(Z)_{Tmut} \qquad \textbf{(eqn.2.1)}$$

As the difference between two known free energies of binding (A and B) yields the

$\Delta\Delta G$ between the two different ligands, the calculated free energy difference between two alchemical transmutations (Y and Z) should yield the same result [21,112,137–139]. This method is better suited to a computational approach, due to the ease of simulating ligand transmutation within the binding site compared to simulating the binding process of two discrete bodies. It has also been reported to have an accuracy of around 1 kcal/mol when comparing simulation results to experimental [103,112]. As a result of using this approach with respect to a protein-ligand binding interaction, the interactions involving common portions of the two ligands cancel out. Therefore, the $\Delta\Delta G_{TI}$ contribution comes mainly from the interactions created or disrupted by the difference in ligand structure, be they protein-ligand, protein-protein or solvent-solute.

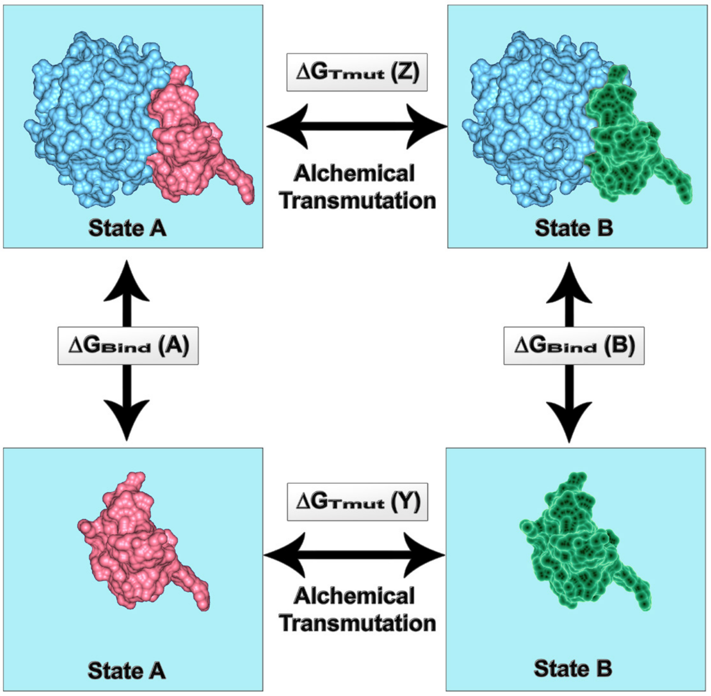


Fig.2.1. An alchemical cycle showing free energy paths for the process of two ligands: A and B binding to the same protein. $\Delta G(A)_{Bind}$ and $\Delta G(B)_{Bind}$, correspond to the free energy of binding of the two different ligands. On the other hand, $\Delta G(Y)_{Tmut}$ and $\Delta G(Z)_{Tmut}$ represent the free energy obtained by transforming one ligand into the other via an unphysical, alchemical transmutation process. Figure adapted from reference [137].

2.1.3. Phase space and ergodicity - the sampling problem

Describing how the position of a molecule evolves with time within a classical computer simulation requires $6N$ descriptors. Each of the N atoms that constitute the molecule, possess three positional coordinates and three descriptors relating to momenta. Every one of these parameters acts as an axis, in what is known as phase space. As a multi-atom system evolves with time, every microstate (snapshot containing $3N$ momenta and $3N$ positions) recorded by an MD simulation can be projected as a point in 6-dimensional space. The latter concept is easier to visualise by considering a subset of phase space known as configurational space. This can be described using only the $3N$ positional coordinates, and is therefore easier to represent in more familiar 3-dimensional space. The positions for every atom in the molecule are merged so that a single point representing one configurational snapshot is created (**Fig.2.2**). The configurational probability density can be evaluated from a large number of these snapshots and information such as the free energy landscape of the molecule can be generated after further calculations [103,140].

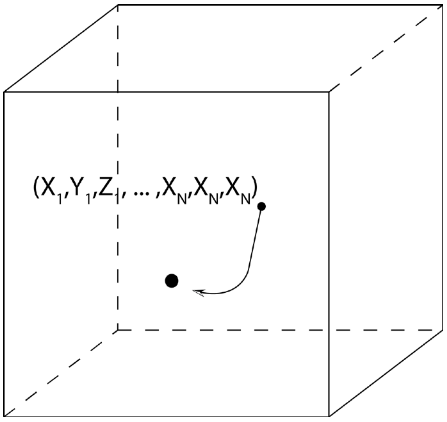


Fig.2.2. Depiction of how 3*N* positional coordinates for a molecule are mapped onto a representation of configurational space. The spatial geometry of the atoms in the molecule, captured from a single static snapshot, is reduced to a point. Figure adapted from reference [140].

In statistical mechanics, an ensemble is a collection of microstates brought together on the basis that they satisfy certain predefined constants that describe the equilibrium state of the system (macrostate). e.g. The temperature, number of particles and the volume of the system make up the Canonical (NVT) ensemble. In order to capture macroscopic properties of the system such as energy or pressure, information about the momentum of particles is required in addition to their positions. At any given moment in time, examining interactions within component microstates allows the instantaneous measurement of any property of the system. As this can vary substantially from measurement to measurement, statistical mechanics obtains what is termed an ensemble average, by averaging measurements from all available microstates of the system. The veracity of this term is dependent upon how thoroughly 6*N*-dimensional phase space is sampled and any trajectory that fully explores phase space is termed ergodic. Simplistically defined, the ergodic hypothesis states that if phase space is populated with sufficient microstates, the ensemble average converges to a limit that is an accurate estimate of the macroscopic property being measured. In practice, simulating a sufficient density of microstates to fully populate the phase space of even small systems (< 50 atoms) presents a considerable challenge. This is partly due to the computational effort in generating sufficient snapshots, and also because MD methods preferentially sample states of lower energy whilst underrepresenting those of higher energy. There are a variety of methods that ameliorate insufficient sampling of phase space [103].

2.1.4. The lambda (λ) parameter

A naive approach to calculating the free energy difference between two simple molecules (such as ethanol and ethane thiol) can be made from a single simulation enclosed within a box of water. At every time step, the energetic contribution is first calculated for the ethanol molecule. Next, the ethanol force field is temporarily modified so that the oxygen atom is described by the parameters relevant to that of the sulphur in ethane thiol. The energy is then recalculated for the same snapshot. In this manner, the ensemble energy of the two states can be averaged and then subtracted from one another to yield the free energy difference. However, if the energetic difference is much larger than k_BT (Boltzmann constant multiplied by temperature), the two molecules will not have sufficient phase space overlap, and the calculation will not be accurate. A separation within phase space results in the ensemble averages inadequately representing the common microstates of both molecules simultaneously [103].

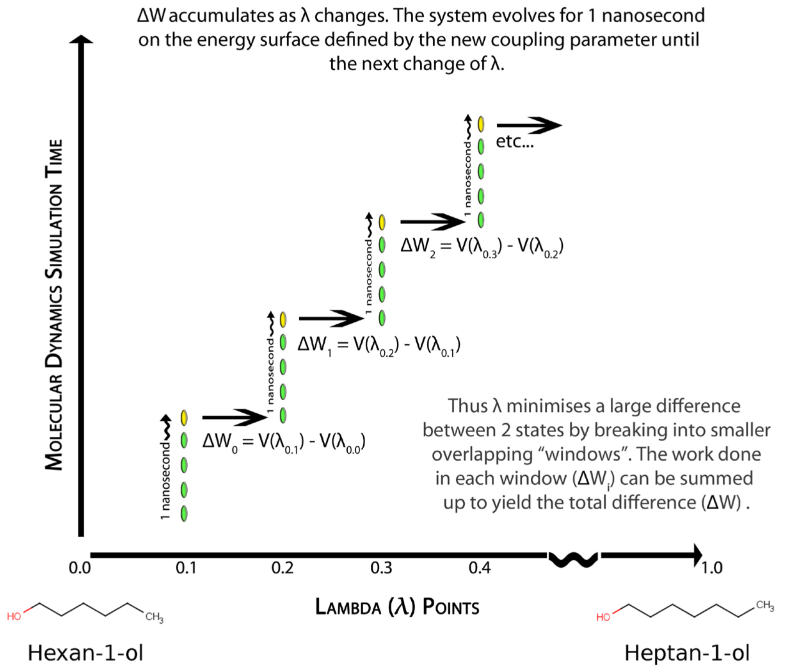

ΔW accumulates as λ changes. The system evolves for 1 nanosecond on the energy surface defined by the new coupling parameter until the next change of λ.
Molecular Dynamics Simulation Time
1 nanosecond
$\Delta W_0 = V(\lambda_{0.1}) - V(\lambda_{0.0})$
1 nanosecond
$\Delta W_1 = V(\lambda_{0.2}) - V(\lambda_{0.1})$
1 nanosecond
$\Delta W_2 = V(\lambda_{0.3}) - V(\lambda_{0.2})$
1 nanosecond
etc...
Thus λ minimises a large difference between 2 states by breaking into smaller overlapping "windows". The work done in each window (ΔW_i) can be summed up to yield the total difference (ΔW) .
0.0
0.1
0.2
0.3
0.4
1.0
Lambda (λ) Points
HO
CH3
HO
CH3
Hexan-1-ol
Heptan-1-ol

Fig.2.3. Illustration of how the λ parameter couples the alchemical transformation within a TI calculation. Discrete simulations at each λ instance accumulate work (ΔW_i) which can be summed to yield the total difference (ΔW) between two states. Figure adapted from reference [142].

Methods such as TI seek to reduce any phase space disparity via the simulation of intermediate states. In order to effect an alchemical transformation, a path needs to be created between state A and B. This is done by introducing the nonspatial parameter λ which acts to couple both states. e.g. In order to simulate the transmutation of hexan-1-ol into heptan-1-ol, the force field description of bond lengths, angles, dihedrals, electrostatics and van der Waals have to be modulated. When λ = 0.0, the force field describes pure hexan-1-ol. When it is 1.0, it is representative of pure heptan-1-ol. λ values between 0.0 and 1.0 (e.g. 0.1, 0.2, 0.3) have hybrid force fields that are a mixture of pure hexan-1-ol and heptan-1-ol. For example, at λ 0.3, the force field has 30% of the character of that used to describe hexan-1-ol and 70% of the character of heptan-1-ol.

Provided the gaps between successive λ steps are small enough, this method ensures that phase space for each state is adequately sampled. As previously mentioned, use of force fields that merely describe the endpoint states do not yield accurate results unless they are extremely similar [103,137–139]. Thus, λ acts to minimise a large difference between two states by breaking it into smaller "windows", each of which explores a part of phase space that overlaps with both the previous and the successive windows (**Fig.2.3**). These work done within these "windows" can be later summed up to yield the total difference [141,142].

In practical terms, a separate simulation has to be run for every instance of λ. The λ derivative of the potential (*V*) is averaged across the range of selected λ values. This is then integrated numerically to give the final free energy as shown in **eqn.2.2** [137,143].

$$\Delta G^{\circ}{}_{TI} = \int_{0}^{1} \left\langle \frac{\delta v(\lambda)}{\delta \lambda} \right\rangle \delta\lambda \qquad \textbf{(eqn.2.2)}$$

2.1.5. Mixing schemes

There are a number of different mixing schemes for force fields when using TI, falling into the category of either linear or non-linear. A commonly seen example of the former is given by **eqn.2.3**, where V_0 represents the potential energy calculated by the force field describing state A and V_1, state B:

$$v(\lambda) = (1-\lambda)v_0 + \lambda v_1 \qquad \textbf{(eqn.2.3)}$$

This is sufficient for most examples of states differing by only atom types or partial charges. However, when the number of atoms increases or decreases, convergence is difficult at high and low λ values. This is because atoms to be deleted are turned into "dummy" atoms that need to have their van der Waals interactions with the surroundings gradually removed. At V_0, the potential energy for disappearing atoms are calculated with the LJ interactions fully switched on, whilst at V_1, these atoms have their LJ interactions completely turned off and thus become dummy atoms. The linear mixing scheme described above creates difficulties due to issues that occur when the LJ potential is linearly scaled across intermediate instances of λ. This often results in large forces and numerical instabilities that manifest as an end-point singularity as λ values approach 1.0. This is because the scheme effectively results in the van der Waals radii of deleted atoms decreasing, whilst concurrently developing a "hard-core" potential that is strongly repulsive. When λ is not exactly 0.0 or 1.0, these "hard-core" atoms can generate massive energy fluctuations upon contact with other atoms in the system.

A variety of other scaling schemes have been suggested to overcome the difficulties with scaling the LJ potential, with varying degrees of success. e.g. a nonlinear scheme with a multiplicative pre-factor k in the form of **eqn.2.4** has been found to be effective in some scenarios:

$$\nu(\lambda) = (1-\lambda)^k \nu_0 + [1-(1-\lambda)^k]\nu_1 \qquad \textbf{(eqn.2.4)}$$

Assigning values of k in-between 4 to 6 removes the singularity and the integrand remains finite. However, this does not resolve the problem, as the final result can still be numerically unstable. Beutler et al. (1994) proposed the best solution of 'soft-core potentials' (**eqn.2.5**) [144]. Instead of multiplication by a pre-factor, the repulsive portion of the LJ term was slowly modulated so that an offset based as a function of λ was added to the inter-atomic distance. The technique allows atoms from different states to gradually overlap, as the full LJ interaction of atoms to be deleted is reduced to zero on λ moving from 0.0 to 1.0 (**Fig.2.4**).

$$\nu_{'Soficore'vdW} = 4\varepsilon(1-\lambda)\left[\frac{1}{[\alpha\lambda + (r/\sigma^{-6})]^2} - \frac{1}{\alpha\lambda + (r/\sigma)^6}\right] \qquad \textbf{(eqn.2.5)}$$

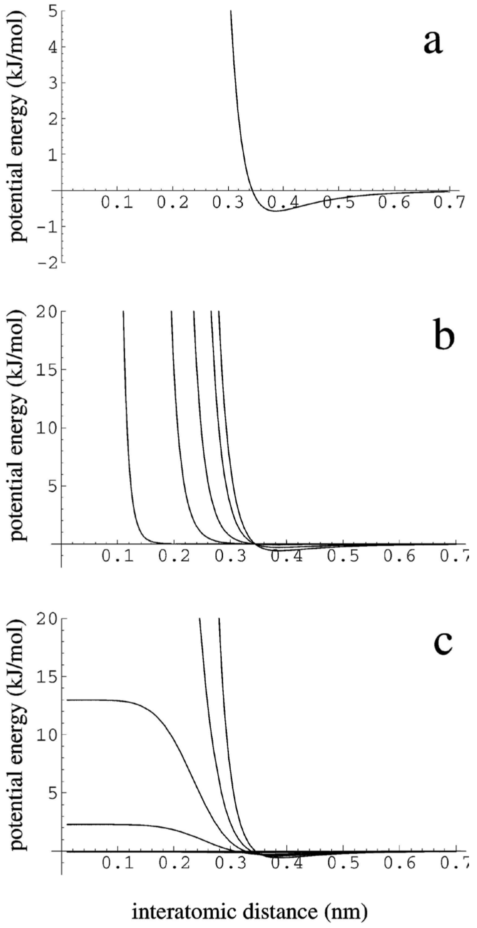


Fig.2.4. Comparing the Lennard-Jones van der Waals potential energy term from the GROMOS96 force field for the interaction of a water atom with an aromatic united carbon (CH). **(a)** The normal (unscaled) interaction. **(b)** The effect of linearly scaling LJ interaction to zero as a function of the coupling parameter λ is displayed for λ = 0, 0.5, 0.9, 0.99 and 0.99999. **(c)** Linear soft-core scaling of λ = 0, 0.25, 0.375, 0.5, and 0.75. Figure and text taken from [145].

In the AMBER 10 implementation (**eqn.2.5**), the atoms scheduled for deletion have their non-bonded interactions with the surroundings smoothly set to zero while still retaining their intramolecular LJ interactions. It would be more accurate to describe them as decoupled from the surroundings as opposed to being annihilated. When soft-core potentials are used, dummy atoms are not required and an even distribution of lambda points assigned by linear mixing ($0.01 < \lambda < 0.99$) is recommended [121]. Additionally, partial atomic charges should be removed prior to the calculation in order to avoid the problem of the van der Waals interactions being scaled down at a faster rate than Coulombic interactions: this can lead to instabilities due to charged solvent interacting with the disappearing atom, despite the soft-core offset [121,138,143–145].

The choice of λ values is important and can strongly influence the final free energy value. The slow growth method relies on λ being altered so that it slowly moves from 0.0 to 1.0. The underlying assumption being that small λ windows ensure that the system is always in equilibrium. Problems with this method include hysteresis and Hamiltonian lag, with the latter occurring due to insufficient equilibration time within a given instance of λ [103]. Additionally, this calculation method is inefficient as convergence failure results in the entire simulation having to be rerun at an even slower rate. The method used in this chapter is much more efficient as the potential energy is evaluated at discrete λ instances and this set-up is amenable to simulations being executed in parallel. If better resolution is required, more λ instances can be added to the simulated data to increase integration quality. Thus, TI is directionless and simulating the reverse transformation to generate the lower error estimate boundary (i.e. hysteresis) is unnecessary [103,137]. A (minor) disadvantage is that every instance of λ must be minimised and equilibrated [138].

2.1.6. Calculating the Entropy and Enthalpy

The alchemical approach can be extended to calculating the enthalpy, but the margin of error is an order of magnitude larger than the values obtained for free energies. This is because the total energy of the system (water molecules plus solute) must be included in the enthalpy calculation. This tends to be dominated by solvent-solvent interactions that are independent of λ and the greater energies generated by thermal fluctuations drown out the signal emitted by the lambda scaled potential linking states A and B. On the other hand, the free energy difference is more easily calculated because solvent-solvent contributions cancel when comparing the ensemble averages of A and B. The difference in solute-solvent energy between states is usually much smaller and thus easier to compute [139,146]. The entropy of a simulation is exceedingly difficult to calculate directly, but could be derived if the enthalpy and free energies were known from **eqn.2.6**.

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ \qquad \textbf{(eqn.2.6)}$$

Alternatively, calculations could be simulated at three different temperatures and the entropy could be calculated by a finite-difference method (FDM) (§2.2.7) and the enthalpy obtained via the use of **eqn.2.6** [147]. There are several other methods of calculating the entropy such as the Shannon entropy equation, the quasi–harmonic approximation, the ad hoc quantum mechanical approximation of Schlitter, hypothetical scanning, NMR-derived order parameters and the mining minima (M2) method [148]. Space limitations prohibit a thorough discussion of these methods here, but some will be discussed in greater depth in subsequent chapters.

2.2.0. Methods

The methods used are based on protocols from the AMBER website and are explained using the example of hexan-1-ol or heptan-1-ol [137].

2.2.1. TI Overview

In order to run the TI calculation, two sets of simulations corresponding to State A and State B are required (**Fig.2.1**). Each member within a set corresponded to the structure of one of the two endpoint structures in the alchemical mutation. The set termed “complex” consisted of hexan-1-ol bound to MUP (State A) and heptan-1-ol bound to MUP (State B), whilst the simulation set termed “FreeLig” corresponded to the endpoint structures of un-complexed ligands.

Throughout this chapter, the three letter acronyms tabulated in **Table.2.1** will be used for various ligands. PDB identifiers are listed where appropriate. The programs and steps necessary to prepare the necessary input files, and the simulation parameters required to run TI calculations, are detailed below.

Ligand	Abbreviation	PDB ID
Hexan-1-ol	hex	1ZNE
Heptan-1-ol	hep	na
Octan-1-ol	oct	1ZNH

Table.2.1. Ligand abbreviations with relevant PDB identifiers.

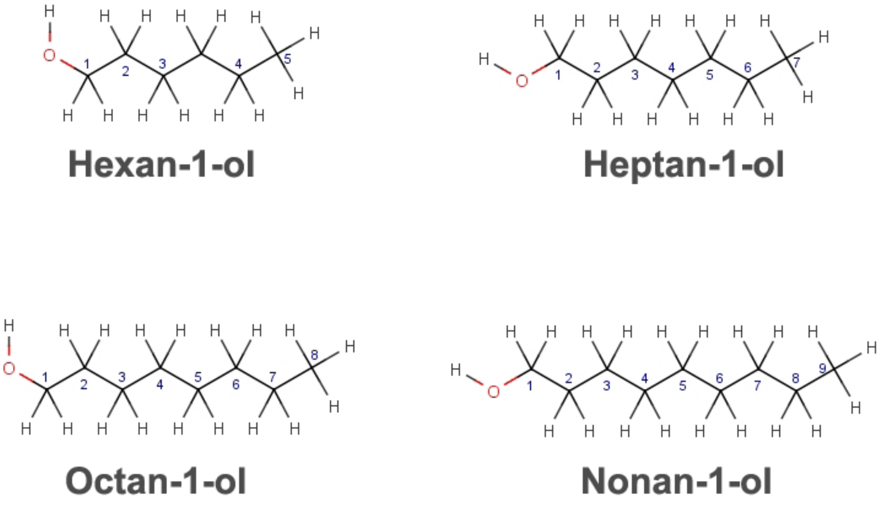
2.2.2. Creation of Ligand Libraries

The program xleap and tleap from the Amber 10 suite were used to create PDB structures of hex and hep with hydrogen atoms, taking care to ensure that atom names were identical for both structures where appropriate. AMBER library files were then created for the ligands so that generation of AMBER topology and initial velocities required to run the simulations could be facilitated. A complication in creating the parameter files involved ensuring the atom starting positions and names between members within a set were identical. The only exceptions to this rule were the atoms being transmutated.

This is particularly important as the TI calculation can only be carried out in-between ligands with small differences such as a single methylene group. Numbering and positional differences result in the sander module crashing catastrophically during the TI portion of the calculation due to the logistics of the procedure.

Ligand parameterisation required calculation of an energy minimised structure with the quantum mechanical (QM) methods used in Gaussian 98 with the 6-31G* basis set. Subsequently, a three dimensional molecular electrostatic potential (MEP) grid was created from this structure and passed to RESP, a program from the AMBER suite that fits the partial atomic charges to the QM derived grid. This process was facilitated by RESP ESP charge Derive II (R.E.D. II) program. R.E.D. II acts as a link between Gaussian 98 and RESP by handling format conversions and automatically ensures file input names are generated correctly. It also allows more accurate partial charge derivation through the use of multiple conformations and orientations.

The PDB files were edited in the manner prescribed by the R.E.D. II manual and the instructions followed therein to generate a Mol2 file [149]. At this juncture, the antechamber module was used to assign atom types from the General Amber Force Field (GAFF) [121]. GAFF is better suited to describe small organic molecules than the normal AMBER force field. As GAFF atom types are in lower case and the normal AMBER force field atom types are in upper case, both force fields can be used simultaneously in one file. This allows analysis of protein and (bound) ligand to occur simultaneously. GAFF parameters were checked with parmchk, to validate bond angles, lengths and dihedrals. Force field modifications were generated for missing or otherwise erroneous assignments. Finally, AMBER library files were created for both hex and hep [150]. The same procedure was used for all the ligands in the primary alcohol panel. i.e. hex, hep and oct (**Fig.2.5**).



Hexan-1-ol

Heptan-1-ol

Octan-1-ol

Nonan-1-ol

Fig.2.5. Primary alcohol panel. Only hex, hep and oct were tested using TI.

2.2.3. Creation of structures fit for Thermodynamic Integration

After creating the library files, the structures for the complexed ligand were created in the following manner.

The PDB structure of heptan-1-ol bound to MUP (1ZNG) was edited so that the ligand had the same three letter identifier (hep) defined in the previously created AMBER library file [36]. The PDB then had all the waters and ions (such as cadmium) left by the crystallisation process stripped out, apart for binding site waters if that particular transmutation required them. This starting structure was then duplicated and used as the starting structure for hexan-1-ol bound to MUP. To maintain identical atom names and starting positions between the two files, the duplicated PDB file was edited so that the carbon atom that formed the terminal methyl group of heptan-1-ol was deleted and the three letter GAFF identifier of the ligand changed from hep to hex.

Each PDB starting structure was then loaded into leap subsequent to loading the Duan et al. (2003) optimised force field [151], appropriate GAFF parameterised library and the all_aminoct02.lib library. The latter enabled the C-terminal CYS157 to be bonded by a disulphide bond to CYS64. Missing hydrogen atoms were added to protein residues according to the Duan et al. (2003) force field parameters. Ligand hydrogen atoms and names were assigned with reference to the AMBER library file created earlier. In the case of loading the hexan-1-ol structure (containing the edited/deleted terminal methyl), leap matched the edited PDB ligand identifier of hex with the library file of the same name and added the appropriate hydrogen atom instead of the terminal methyl carbon (in heptan-1-ol). A solvent box of TIP3P waters with a box to solute distance of 12.0 angstroms was created around the MUP complex. The overall charge of the system was neutralised by the addition of K+ ions.

Ligands that compose the FreeLig set were also generated from the parameterised library file and enclosed in a box of water using the method used to create the complexed set. **Table.2.2** lists the three sets of TI simulation parameter files created for alchemical transmutation:

	Alchemical Transformations
1.	hex to hep (without water)
2.	hex to hep (with 3 waters)
3.	hep to oct (with 2 waters)

Table.2.2. List of alchemical transformations.

2.2.4. Simulation settings used for Thermodynamic Integration

Both complex and FreeLig sets were comprised of three calculation steps, each consisting of 9 λ points. In the first step, the partial atomic charges on (unique) atoms designated for transmutation were switched off to avoid simulation instability in the subsequent step. Secondly, soft-core potentials were switched on and the transmutation of hex to hep was accomplished by replacing the hydrogen attached to C6 with a methyl group. In step-3, partial atomic charges were switched back on.

The calculation for each λ point in any given step consisted of minimisation, equilibration and production stages. Each 500 cycle minimisation utilised a nonbonded cut-off of 9 Å. 10 cycles of steepest descent were followed by 490 steps of the conjugate gradient method for all λ instances, bar those that used soft-core potentials. The latter exclusively used the steepest descent method. Constant volume simulations were used for the initial equilibration step and were followed by another 50 ps equilibration at constant pressure (1 bar). This was carried out to attain the correct density and temperature, and also to avoid the problem of "vacuum bubbles" forming in the solvent. As a result, pressure values vary widely during the simulation as the unit cell volume is adjusted. Energy, volume, density and temperature parameters were monitored to assess whether equilibration had been achieved. The final production stage utilises the equilibration restart file and was run in 200 ps segments for a total of 5 ns. The simulation parameters are essentially the same as that used in equilibration under constant pressure, apart from TI specific settings and switches.

During steps 1 and 3, the SHAKE algorithm was used to avoid calculations of bond interactions involving hydrogen and thus increased the achievable time step to 2 fs. As SHAKE could not be used with soft-core potentials, a smaller time step of 1fs was utilised to capture the fastest motions in the system. The simulations were run at constant temperature at 278, 300 and 322 K using a Langevin thermostat. This particular thermostat uses a pseudo-random number stream to model arbitrary collisions which alter the velocities of molecules within the simulation. As the AMBER restart file does not store the state of the random number generator in the output stream, this leads to the initial seed being reused and the same set of initial random velocities being used in each segment. These repeated sequences have the effect of reducing the conformational space available to the system and can affect dynamics in undesirable ways, such as driving the protein into periodic trajectories. The shorter the segment the greater the problem, and in extreme cases the protein can lose stability and denature [152]. Therefore, in order to avoid introducing the correlation errors that accumulate when using a static random number seed, a script was developed to generate a different random number seed for each equilibration and production segment.

To provide an overview of the computational intensity and costs of these calculations, a 1ns simulation (compromising both FreeLig and complex) utilised 108 parallel processors, with each λ point making use of 2 processors. Thus, a 1 ns TI calculation run at a single temperature took ~3 days on the arc1 cluster at Leeds in 2010. Addition of two further TI calculations at alternate temperatures only added another day due to the parallel nature of job submission. However, finite resources and logistics associated with the cluster queue dramatically reduced the parallel processing advantage if longer simulations, additional repeats and/or additional TI transmutations were carried out. The compressed size of simulations run at three different temperatures was ~550 GB per repeat. In order to achieve convergence (§2.4.4), 5 repeats were performed and this yielded an aggregate data size of ~2.5 Tb per transmutation.

2.2.5. Analysis of simulation data

The ptraj program (from the AMBER suite) was used to calculate the root mean square deviations (RMSD) of the complex. The RMSD of the MUP carbon alpha chain was monitored over the equilibration and production phases so as to gauge the stability of the protein (**eqn.2.7**). The RMSD can be calculated between a reference structure (typically the first or last production frame, or alternatively as an average of all production frames) at time 0 (t_0) to the protein structure at another time point (t_1). The term r_i corresponds to the co-ordinates of *N* protein atoms and is used to measure the average distance between all atoms in the reference structure to that of any other given snapshot in time. This can be undertaken after removing translational and rotational rigid body motions (minT,R) and superimposing the two structures. This work used the first frame as a reference.

$$RMSD(t_0.t) = \min_{T.R} \left\{ \sqrt{\frac{1}{N} \sum_{i=1}^{N} | r_i(t_0) - r_i(t) |^2} \right\} \qquad (\mathbf{eqn.2.7})$$

Further analysis of simulation conditions from output files were accomplished using a suite of in-house developed tools that were built using standard Linux programs such as grep, awk, sed, gnuplot and python. For each step, the free energy change was calculated by numerical integration (using the trapezoid rule) of pathway described by ∂V/∂λ values.

2.2.6. Analysis of statistical error

The statistical error inherent in the TI calculation is complicated to calculate. The principal sources of error are insufficient sampling of phase space, finite-size effects, interaction cut-off and other inaccuracies in the Hamiltonian. There is also a degree of error depending upon the quality of the final numerical integration. As each instance of λ simulated is independent from adjacent λ instances constituting the integration path, calculation of hysteresis is unnecessary [103,138,153,154].

Analysis of statistical error is reliant on independent measurements being made. However, measurements made during an MD simulation are limited by finite sample sizes and the computational time it takes to adequately describe phase space can be prohibitive. In order to avoid inaccuracies it is important to establish whether the dataset obtained is correlated or not. Correlation can be assessed through block averaging the data or via the use of an autocorrelation function [153,154]. However, a method known as "Independent-Trajectories Thermodynamic-Integration (IT-TI)" proposed by Lawrenz et al. (2009) appears to demonstrate a more promising alternative [155]. IT-TI improves the sampling of phase space by averaging *N* independent trajectories. Due to differences in the way phase space was explored in independent trajectories, free energy differences were typically under or overestimated by more than 9%. This was due to inconsistent sampling of the conformations of flexible loops. When the trajectories were averaged, the researchers calculated a value within 1.1 kJ/mol (< 1%) of results obtained from ITC measurements of the binding of Amprenavir to the H5N1 neuraminidase N1 receptor [155]. Consequently, five independent repeats were simulated at three different temperatures. The standard deviation from the mean of *N* (independently repeated) free energy results ($\sigma_{\Delta G}$) can be calculated to give a measure of the statistical error. Larger sample sizes therefore give more accurate estimates [155]. The resulting errors were propagated by taking the square root of the sum of the squares.

2.2.7. Calculation of entropy and enthalpy via finite difference analysis

As has already been discussed, the TI method cannot be extended to a calculation of the enthalpy or entropy as the statistical uncertainty is roughly an order of a magnitude greater than that for free energy calculations. Consequently, a finite-difference method (FDM) that determines the entropy by evaluating the temperature derivative was utilised. To calculate the entropy difference between states A and B with the FDM, the relationship described by **eqn.2.8** is utilised.

$$\Delta S^{\circ}{}_{ab}(\mathrm{T}) = -\left(\frac{\delta \Delta G^{\circ}{}_{ab}}{\delta T}\right)_{P,N} \qquad \textbf{(eqn.2.8)}$$

The derivative of the three simulations run at 278, 300 and 322 K was calculated using **eqn.2.9**.

$$\Delta\Delta S^{\circ}{}_{(300)} = -\left(\frac{\Delta\Delta G^{\circ}{}_{(322)} - \Delta\Delta G^{\circ}{}_{(278)}}{2\Delta T}\right) \qquad \textbf{(eqn.2.9)}$$

T equals the target temperature, whilst ΔT is the difference between the target temperature and the flanking temperatures. A value in-between 30 to 50 K is recommended for ΔT because the method assumes that heat capacity is independent of temperature over this range. At a ΔT of 30 K, the statistical error obtained with this calculation is approximately ten times that obtained for the ΔΔG [156,157].

To compare the results to experimental data, the TI calculation was run at the selected temperatures and the finite difference method was used to calculate the entropic and enthalpic differences between state A and B at 300 K. The lower bound for the temperature was selected to avoid running the simulation at a temperature of 0°C where water exists as ice. The upper limit was constrained by the fact that ΔT is a constant, and thus has to be 322 K. This results in a temperature difference ($2\Delta T$) of 44 K [156,157]. The enthalpy was obtained by using **eqn.2.10** and the corresponding error obtained by calculating the square root of the sum of the squares.

$$\Delta\Delta H^{\circ}_{(300)} = \Delta\Delta G^{\circ}_{(300)} + T\Delta\Delta S^{\circ}_{(300)} \qquad \textbf{(eqn.2.10)}$$

2.3.0. Results

2.3.1. Minimisation and equilibration phase

Simulation stability was assessed by measuring key metrics during the equilibration and production stages for all λ points. A representative sample from the equilibration phase of a single complex (hex to hep) transmutation at 300 K for step-1 is shown in **Fig.2.6.** The system was minimised to a sufficient level to ensure stability instead of undergoing exhaustive minimisation to obtain the lowest energy minima. Initial starting structures generated by tleap add water molecules in an ordered lattice around the solute. The equilibration phase serves to “melt” this unphysical configuration to one akin to that observed in liquid water. This is achieved by assigning initial velocities from a Maxwell-Boltzmann distribution and heating the system to attain the desired temperature, volume and density. The density is slightly higher than the density of water due to the presence of solute (ions, protein and ligand). After approximately 10-15 ps the density and volume plateaued, fluctuating around stable averages and 50 ps equilibration was deemed sufficient (**Fig.2.6**).

2.3.2. Production phase

The production stage was run for 200 ps initially. A sample depicting the first production stage segment from a complex (hex to hep) transmutation at 300 K for step-1 is depicted in **Fig.2.7.** Ptraj analysis of production RMSD shows that the MUP carbon alpha backbone is stable. After a period of time that is sufficient for loss of correlation with its starting co-ordinates, the RMSD of the protein initially increases with respect to the initial crystal structure as the configurations sampled move further away from the starting structure. The RMSD fluctuations indicate that the system fully equilibrates after ~50 ps. The simulations were then extended by 200 ps segments for a total length of 5 ns.

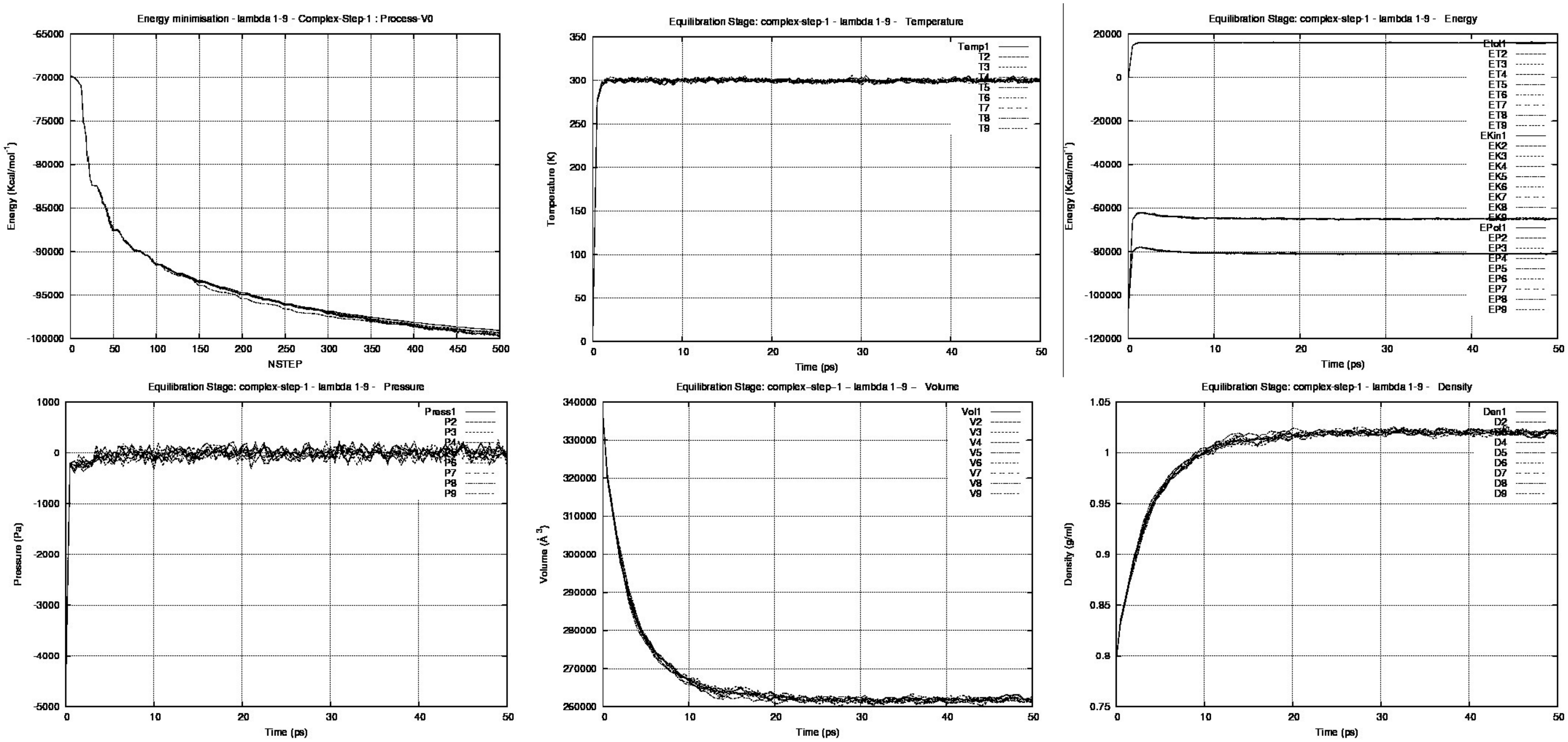


Fig.2.6. Energy minimisation and metrics from 50 ps of complex, equilibration step 1 at 300 K, for the transformation of hex to hep. Graphs depicted (clockwise around panel are initial system energy minimisation; temperature; energy (potential, kinetic and total); pressure; volume and density. Values from all 9 lambda instances are plotted on every graph.

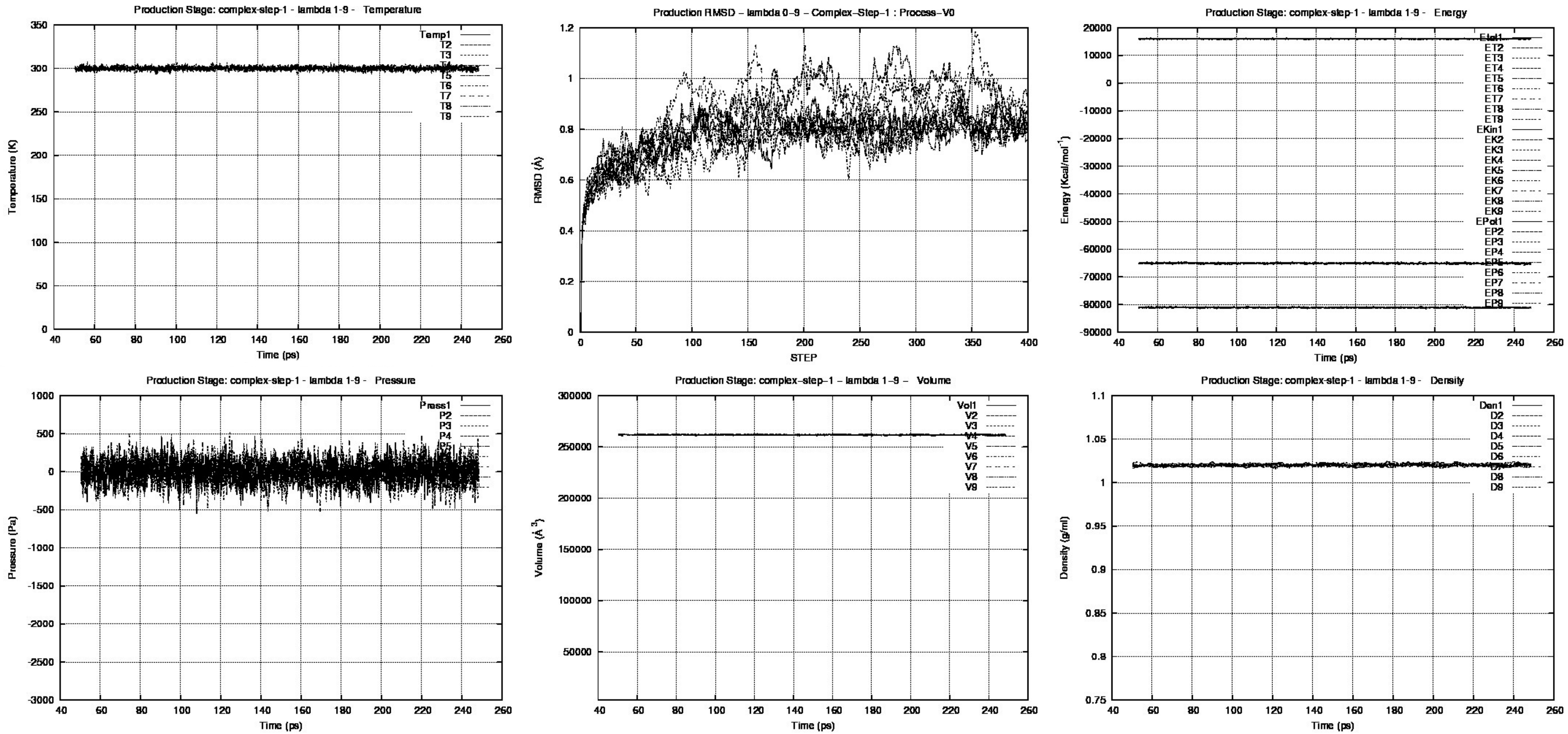


Fig.2.7. First 250 ps of complex, production step 1 at 300 K, for the transformation of hex to hep. Graphs depicted (clockwise around panel are temperature, RMSD, energy (potential, kinetic and total), pressure, volume and density. Values from all 9 lambda instances are plotted on every graph.

2.3.3. Decomposing the double free energy difference

Due to space considerations, the procedure for calculating the free energy difference is illustrated by the transmutation of hex to hep with binding site waters. At the end of 5 ns, each of the nine instances of λ simulated generated a value for ∂V/∂λ. The resulting ΔG_{TI} curves were numerically integrated for both complex and FreeLig for steps 1-3. The values generated for FreeLig were then subtracted from complex to generate the $\Delta\Delta G_{TI}$ for each step. The results for each integration step are located in **Table.2.3**. This was done for 5 repeats across 3 different temperatures. The error is represented by the standard deviation as discussed in the methods. The final double free energy difference ($\Delta\Delta G_{TI}$) is obtained by summing steps 1-3 for "complex minus FreeLig". **Fig.2.8** shows the averaged λ integration path (alongside individual repeats), with associated errors for FreeLig, complex and "complex minus FreeLig".

Results Summary for Transmutation of hex to hep (with water) at 278K (kJ/mol)							
	Repeat 1	**Repeat 2**	**Repeat 3**	**Repeat 4**	**Repeat 5**	**Mean**	**Std Dev**
Step-1	-0.16	-0.05	-0.04	-0.07	0.02	-0.06	0.06
Step-2	-6.00	-4.77	-4.39	-4.50	-4.75	-4.88	0.65
Step-3	-0.75	-1.02	-0.62	-0.72	-0.73	-0.77	0.15
					Total	**-5.71**	**0.73**
Results Summary for Transmutation of hex to hep (with water) at 300K (kJ/mol)							
	Repeat 1	**Repeat 2**	**Repeat 3**	**Repeat 4**	**Repeat 5**	**Mean**	**Std Dev**
Step-1	-0.12	0.01	-0.02	-0.08	-0.02	-0.04	0.05
Step-2	-3.96	-4.94	-5.55	-4.23	-3.64	-4.46	0.77
Step-3	-0.92	-0.84	-0.71	-0.68	-0.74	-0.78	0.10
					Total	**-5.29**	**0.74**
Results Summary for Transmutation of hex to hep(with water) at 322K (kJ/mol)							
	Repeat 1	**Repeat 2**	**Repeat 3**	**Repeat 4**	**Repeat 5**	**Mean**	**Std Dev**
Step-1	0.04	0.01	-0.06	0.02	0.05	0.01	0.04
Step-2	-3.93	-4.65	-6.36	-4.54	-4.15	-4.73	0.96
Step-3	-0.58	-0.67	-0.45	-0.47	-0.68	-0.57	0.11
					Total	**-5.29**	**0.94**

Table.2.3. ∂V/∂λ values for transformation of hex to hep with water at three different temperatures. Rows depict the energetic value obtained by integrating under the averaged (across 5 repeats) "complex minus FreeLig" curve (**Fig.2.8** bottom row) at 278, 300 and 322 K.The averaged values for all 3 steps are then summed to generate the total free energy difference between and hep. Errors are represented by the standard deviation.

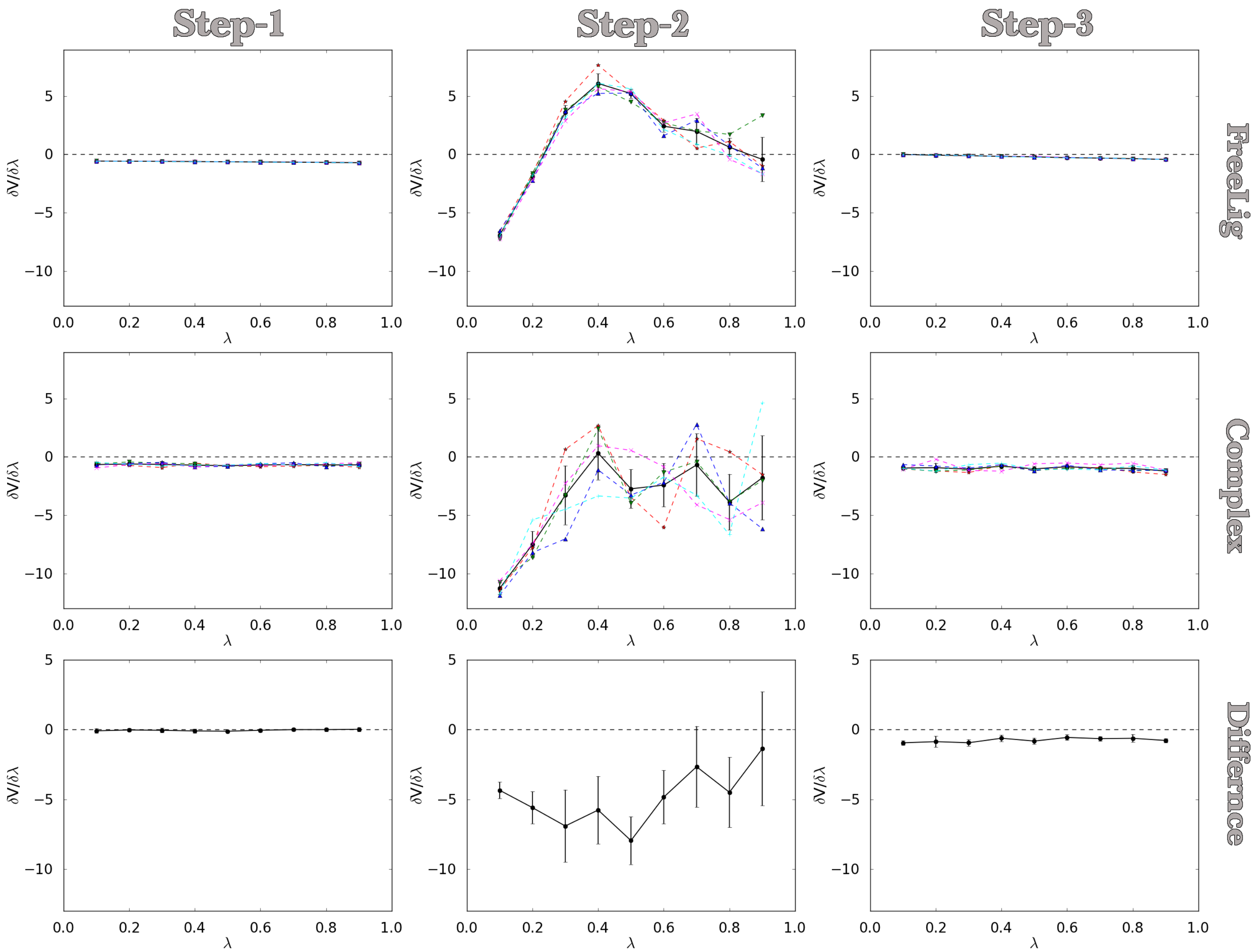


Fig.2.8. A representative image showing ∂V/∂λ values for the three steps in the TI transformation of hex-hep (with water) at 300 K. Columns 1-3, correspond to steps 1-3, whilst row 1 represents the FreeLig simulations; row 2 the complex simulations; and row 3 depicts complex minus FreeLig. The 5 repeats are displayed as coloured, dotted lines, whilst the average is a solid black line with errors shown as standard deviations.

The largest change occurs in step-2, when a single hydrogen atom located on the terminal methyl of hex is deleted and transmutated into the CH_3 group possessed by hep (**Fig.2.8** & **Table 2.3**). This is indicative of van der Waals interactions being the primary contributor to the $\Delta\Delta G_{TI}$. If the contribution were due to electrostatics, the largest energetic differences would have manifested themselves during step-1 and step-3. Although the integration path taken during step-2 for both complex and FreeLig describes a change of much large magnitude than that for step-1 and step-3, the FreeLig contribution to $\Delta\Delta G_{TI}$ is insignificant. The averaged step-2 "complex minus FreeLig" value equates to -4.46 kJ/mol at 300K, and is obtained by subtracting the component FreeLig (0.40 kJ/mol) from the complex (-4.07 kJ/mol) value. The latter is a consequence of hex being transmutated into hep whilst surrounded by water, and thus the ΔG_{TI} value (which is very small and unfavourable) reflects the difference in solute-solvent dispersive interactions between the two ligands. This is expected, as any favourable solute-solvent van der Waals interactions formed with the additional methyl group are offset by the energetic cost of breaking favourable solvent-solvent bonds upon creating the required cavitation space in water. On the other hand, the same transformation within the confines of the protein binding pocket engenders a favourable ΔG_{TI} contribution that dominates the entire process. This is due to the additional methyl group increasing protein-ligand, solute-solute interactions.

The smallest energetic change occurs within step-1 and step-3 where partial charges on transforming atoms are turned "off" and "on" respectively. The values obtained are logical, as the energetic cost (-0.04 kJ/mol) associated with turning "off" the partial charge on a single hydrogen atom is not considerable. The price of turning "on" the partial charges of an entire methyl group is also minimal, with a slightly favourable electrostatic interaction of -0.78 kJ/mol.

2.3.4. Convergence of $\Delta\Delta G_{TI}$

Global convergence of the free energy curves with respect to simulation length was checked by monitoring how the value varied with time. Representative graphs are shown for the three temperatures monitored in **Fig.2.9.** Despite the fluctuations in the $\Delta\Delta G_{TI}$ convergence plot being small near the end of 5 ns, the free energy curves for the complex is not very smooth and further improvements can be made by increasing the number of λ points and the use of a different integration scheme. Disparities between individual repeats are smaller for FreeLig compared to complex. This is because the free energy landscape explored possesses less complexity along the integration pathway between $\lambda(0.0)$ and $\lambda(1.0)$ (**Fig.2.8**). As the largest magnitude of change occurs in step-2, this is expected to contribute the largest source of error.

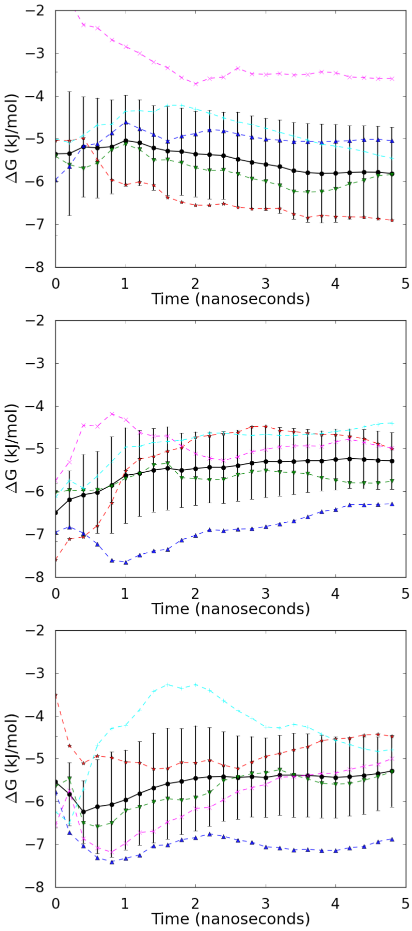


Fig.2.9. Global ΔG convergence for hex to hep (with water) transmutation. Figure panels a, b and c show ΔΔG calculated at 278, 300 and 322K respectively. The 5 repeats are displayed as coloured, dotted lines, whilst the average is a solid black line with errors shown as standard deviations.

2.3.5. Finite difference analysis and experimental comparison

In order to calculate the entropies and enthalpies with the FDM, the key requirement is that the $\Delta\Delta G_{TI}$ must change linearly with temperature. This condition is best met for the transformation of hex to hep without water. The other two transformations are linear within error and a weighted least squares fitting procedure was used to generate entropy and enthalpy values (**Fig.2.10**) [158]. A summary of the final results for all three transformations are listed in **Table 2.4**.

Summary of TI Free Energy Differences (kJ/mol)			
	hex to hep (No Water)	**hex to hep (with Water)**	**hep to oct (with Water)**
278K	-5.87 ± 0.42	-5.71 ± 0.73	-5.01 ± 0.67
300K	-5.51 ± 0.75	-5.29 ± 0.74	-5.03 ± 0.79
322K	-5.15 ± 0.46	-5.29 ± 0.94	-4.47 ± 0.67

Table.2.4. Summary of the final $\Delta\Delta G_{TI}$ for each of the three alchemical transmutations, carried out over three different temperatures.

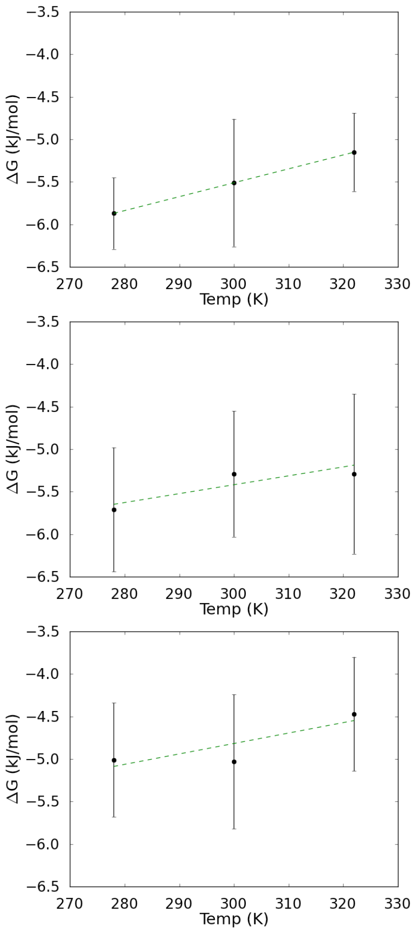


Fig.2.10. Linear dependence of $\Delta\Delta G_{TI}$ on temperature. Figure panels a, b and c represent the transmutations for hex to hep (without water); hex to hep (with water); hep to oct (with water). Errors shown as standard deviations.

The mean of the free energies obtained at 278 and 322 K were averaged and the fitted values were used to calculate the $\Delta T\Delta S_{TI}$ via FDM (**eqn.2.8**), as illustrated by the transformation of hex to hep (without water) The error was calculated by taking the square root of the sum of the squares of the $\Delta\Delta G_{TI}$ values (at 278 and 300 K). This was then divided by Δ44 K and multiplied by 300 K to yield a final error of 4.25 kJ/mol.

$\Delta\Delta S°_{TI}$ = -(-5.15 - (-5.87))/44 = 0.0164 kJ/mol/K

$\Delta T\Delta S°_{TI}$ = -4.91 ± 4.25 kJ/mol

The enthalpy at 300 K was calculated with **eqn.2.10** to yield:

$\Delta\Delta H°_{TI\ (300)}$ = -5.51 kJ/mol + (-4.91 kJ/mol)

$\Delta\Delta H°_{TI\ (300)}$ = -10.42 ± 4.31 kJ/mol

Global experimental values are used to compare the results from the TI calculation, as the method utilised does not separate the thermodynamic values into intrinsic and solvent terms (**Table 2.5**).

Comparison TI results to ITC at 300K (kJ/mol)			
		TI ΔΔG	**TI minus Expt (ΔΔΔG)**
Hex to Hep (No water)	-4.20 ± 0.08	-5.51 ± 0.75	-1.31
Hex to Hep (With water)	-4.20 ± 0.08	-5.29 ± 0.74	-1.09
Hep to Oct (With water)	-3.10 ± 0.11	-5.03 ± 0.79	-1.93
	Expt ΔΔHobs	**TI ΔΔH**	**TI minus Expt (ΔΔΔH)**
Hex to Hep (No water)	-5.80 ± 0.72	-10.42 ± 4.31	-4.62
Hex to Hep (With water)	-5.80 ± 0.72	-8.56 ± 8.14	-2.76
Hep to Oct (With water)	-4.60 ± 0.72	-8.49 ± 6.45	-3.89
	Expt ΔTΔSobs	**TI ΔTΔS**	**TI minus Expt (ΔΔTΔS)**
Hex to Hep (No water)	-1.60 ± 0.72	-4.91 ± 4.25	-3.31
Hex to Hep (With water)	-1.60 ± 0.72	-3.14 ± 8.11	-1.54
Hep to Oct (With water)	-1.50 ± 0.72	-3.68 ± 6.46	-2.18

Table.2.5. Comparing TI to experimental. As the free energy dependence on temperature was not linear for two of the transmutations, a weighted least squares fit was utilised to obtain enthalpies and entropies.

2.4.0. Discussion

The main aims of this chapter were twofold. The first objective was to capture the experimental trends and underlying thermodynamic binding signature observed on binding primary alcohols to MUP via an *in silico* method. Pinpoint accuracy was not a primary concern and it was considered more important to minimise computational cost whilst maintaining an acceptable level of accuracy. The second objective was to see whether the precision that TI is known for could be leveraged to allow calculation of the entropies and enthalpies that constitute the free energy.

2.4.1. Quality of results and free energy convergence

The results for $\Delta\Delta G_{TI}$ at 300K show good agreement with ITC data, which is surprising as the calculation was relatively coarse. The difference between experiment and simulation is under 2 kJ/mol for $\Delta\Delta G_{obs}$, for all transmutations (**Table 2.5**). This is within the kilocalorie window of accuracy expected for TI calculations. The mean $\Delta\Delta G_{TI}$ has been systematically overestimated for all the transformations, whilst the values for individual repeats span a gamut of values, with one even coming within 0.2 kJ/mol of $\Delta\Delta G_{obs}$ (**Table 2.3**). When coupled with the fact that unique random number streams are used to generate every trajectory representing a $\partial V/\partial\lambda$ value, this variance is indicative of enhanced sampling of configurational space. Using the IT-TI method in addition to extending simulation time is an excellent method to decrease the execution time required to achieve adequate sampling by taking advantage of the many parallel processors available in High Performance Computing (HPC) environments.

Of further note is the fact that $\Delta\Delta G_{TI}$ is "built up" from basic interactions using the six steps that compose FreeLig and complex simulations. Even so, the methodology accurately managed to capture the dominant enthalpic signature of the binding reaction as shown by the largest energetic changes occurring in step-2 of the complex and FreeLig simulations. This reflects the van der Waals contribution of growing the ligand by a methyl group. If the chief contributions to $\Delta\Delta G_{TI}$ were due to electrostatic interactions, the largest contributions to the energy changes would have arisen in step-1 and step-3 upon turning the partial atomic charges "off" and "on" respectively. Thus, the method is not a "black-box" and the mechanics behind the calculation are congruent with the known physical data on ligands binding to MUP. As the structural difference scrutinised by the TI calculations is the extension of the ligand by a methyl group, these *in silico* experiments provide additional support to the hypothesis that an increase in solute-solute interactions are principally responsible for the observed increase in affinity, not solvent reorganisation. This conclusion can be drawn because the ligand in the FreeLig transformation obtains a small unfavourable enthalpic term (0.40 kJ/mol), chiefly due to the cavitation cost of elongating the ligand by a methyl group. However, the same transformation within the bound state is accompanied by a favourable enthalpic term that is indicative of the increased solute-solute dispersion interactions that dictate binding affinity. The source of the favourable solute-solute term could potentially be due to increased ligand-protein and/or protein-protein interactions.

Despite $\Delta\Delta G_{TI}$ values remaining within ~2.0 kJ/mol of experimental results, calculation of the entropies and enthalpies via the finite difference method (§2.4.2) necessarily requires a greater amount of precision. The mean $\Delta\Delta G_{TI}$ convergence at all temperatures (**Fig.2.9)** lies within a narrow energetic window and possesses minimal fluctuations compared to its component repeats. However, a caveat that must be considered is whether the variance in calculated $\Delta\Delta G_{TI}$ values are the result of the ligand visiting metastable states unrepresentative of the principal binding mode. Longer simulation times and/or additional repeats would have to be employed to allow the ligand to adequately sample all possible orientations within the binding pocket. This problem would be aggravated by using longer primary alcohols, as the extra methylene units in the flexible molecule could potentially enable more exploration of the pocket due to additional DOF. Conversely, if the ligand has a limited number of binding poses, sampling and thus convergence may also be negatively impacted. For example, catechol binds to a T4 lysozyme double-mutant in two stable orientations. Even in simulations that approached 5 ns, only a single orientation was sampled due to large kinetic barriers separating the two states [159]. As MUP is a promiscuous protein that can accommodate a variety of hydrophobic binding partners of disparate size, some plasticity of ligand orientation within the binding site is to be expected [160–162]. This is in contradiction to a key assumption made in several theoretical thermodynamic decompositions that assume that the ligand loses significant translational, rotational and conformational entropy [89,163,164]. Thus, validating the binding orientation sampled *in silico* is an important check in assessing the precision of any thermodynamic predictions. This will be determined in a later chapter by running long MD simulations and identifying the most populated ligand binding orientations.

Another consideration as to the accuracy of the TI method is the appropriate selection of λ points. The free energy curves are not as smooth as they could be and possess a "jaggedness" that is exacerbated by the use of the linear, trapezoid integration method. The problem is more pronounced for the complex than for the FreeLig simulations due to the greater complexity of the free energy landscape navigated (**Fig.2.8**). The number and distribution of λ instances simulated is intrinsically linked to the numerical integration scheme chosen. Precise integration is aided through the generation of smooth ∂V/∂λ curves. In an evaluation of the precision of schemes such as Simpson's and the trapezoidal rule, compared to more complex fitting functions based on polynomials, Jorge et al. (2010) employed their own physically based function that was purported to better describe the behaviour of their data [165]. In their analysis of the electrostatic contribution to hydration energy of methanol, all schemes were precise to 0.05 kJ/mol of a reference (composed of 129 equidistant λ points), whilst using 17 λ windows. Reduction of the number of windows caused the trapezoidal rule to systematically underestimate the free energy. As the other schemes were not based on linear interpolation between adjacent λ instances, they could still capture the curvature of the data in as little as 5 λ instances. The estimation of the LJ contribution was more difficult and even the physically based function required a minimum of 11 λ windows. Another study by Steinbrecher et al. (2007) demonstrated that a Gaussian integration scheme with 12 λ instances was capable of a precision that differed by 0.04 kJ/mol from a reference containing 99 instances [143]. Using 9 instances decreased the precision tenfold [143,165]. However, changing the method used in this chapter to use a Gaussian integration scheme would involve repositioning λ instances and running all the simulations again.

The improvement in $\Delta\Delta G_{TI}$ prediction is expected to be marginal by merely replacing the trapezoid method with a non-linear integration scheme, as examination of **Fig.2.8** indicates that an increased density of λ points is required to best smooth the curves. The data was tested using the non-linear Simpson's rule with no substantial change in the final $\Delta\Delta G_{TI}$ value. The trapezoid scheme used in this chapter is popular in TI calculations due to its flexibility. Any number of λ values with variable spacing can be used, and thus more instances of λ can be added after an initial simulation to smooth the ∂V/∂λ curve. As instances of λ at 0.0 and 1.0 cannot be simulated, the integration script linearly extrapolates these values from the closest adjacent values. i.e. 0.1 and 0.9 respectively. Whilst additional λ points could have been strategically generated on a per-simulation basis for areas in the curve that required additional definition, difficulties in generating linear free energy plots across three temperatures (**Fig.2.10**) raise serious questions regarding the ability of TI to consistently generate accurate entropies and enthalpies.

2.4.2. Entropy and enthalpy estimates

From the three transmutations tested, only hex to hep (without water) was able to yield entropy and enthalpy values suitable for FDM without using a fitting procedure. The other two transformations did not exhibit linearity across $\Delta\Delta G_{TI}$ values for the temperatures simulated and fitting the values via linear regression produced large associated errors (**Fig.2.10** & **Table.2.5**). The problem is exacerbated by the fact that the differences between $\Delta\Delta G_{TI}$ values at the temperatures flanking 300 K are very small (**Table.2.4**). As they range from 0.35 at worst to 0.72 at best, this means that the method is very unlikely to reliably capture the entropy change and, by extension, the enthalpy change for this particular system. It is unlikely that a 2ΔT value greater than 44 K could be used, as obtaining results via the FDM for a target temperature of 300 K necessitates that the lower temperature boundary for ΔT would approach, or be below the freezing point.

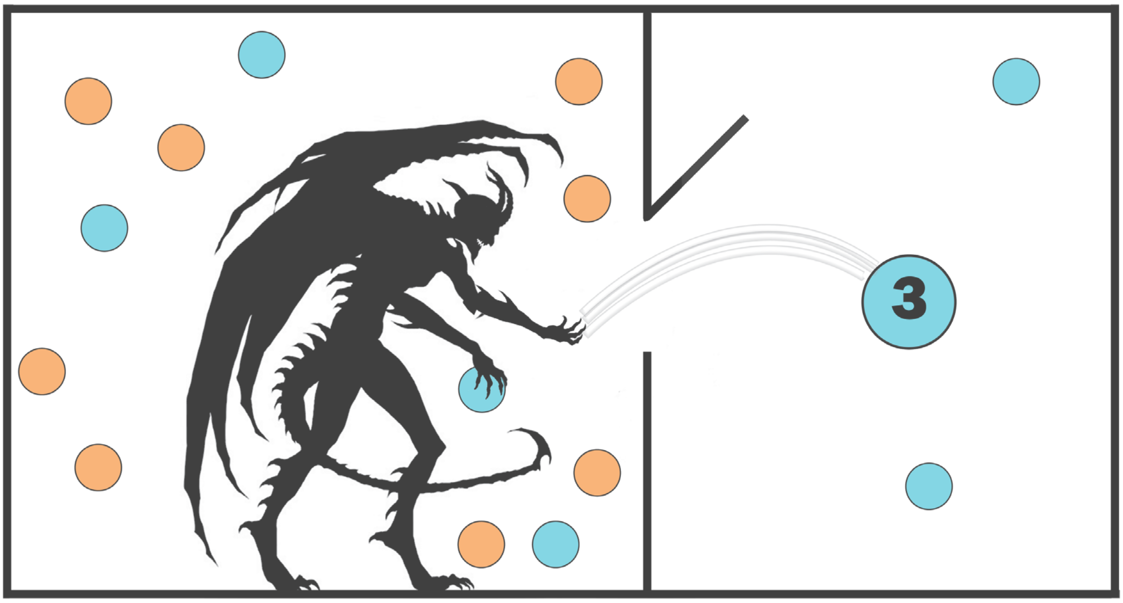
The disparity between calculated $\Delta\Delta H_{TI}$ and $\Delta T\Delta S_{TI}$ to experimental results and their errors are both within a kilocalorie for the transformation of hex to hep (without water), whilst the other two transformations possess larger errors in the region of 2 kCal. The values for $\Delta\Delta H_{obs}$ and the $\Delta T\Delta S_{obs}$ are also approximately within 2 and 1 kCal/mol respectively. Thus, this particular computational approach will not replicate these experimental values with sufficient accuracy, unless the margin of error is substantially reduced. As this is represented by the standard deviation, the best method of doing this is to run additional simulations to increase the number of repeats. As detailed in the previous section, the simulations could be run for a longer period of time with additional λ points. However, the margin of improvement is expected to be small, whilst the computational cost in terms of processing time and hard disk space would be too high, to be considered feasible.

2.4.3. Inclusion of binding site waters

The number of waters bound to MUP is quite variable and Malham et al. (2005) postulated that the limited number of waters in the calyx served to optimise binding at the solute-solute interface [36]. Published crystal structures indicate between 0-4 waters can be present depending on which ligand is bound. The effects on the inclusion of binding site waters were roughly tested by comparing the transmutation of hex to hep, with and without crystallographic waters. The results (**Table.2.5**) indicated that there is a small energetic difference (0.22 kJ/mol) between the two transmutations at 300 K. Further simulations would have to be run to assess whether this is statistically significant and if it varies for other ligands in the primary alcohol panel. Ideally, a dedicated calculation would be set up where the focus of the transmutation is the removal of the key waters. e.g. oct (with water) to oct (without water) at 300 K. As the size of the ligand increases in the pocket, it is possible that these bound waters might play a more prominent role in facilitating protein-ligand interactions than with smaller ligands. However, problems with adequate conformational sampling are still applicable and as further simulations were an expensive proposition, this initial foray was deemed to suffice at this stage.

2.5.0. Conclusion

The Thermodynamic integration methodology was able to generate $\Delta\Delta G_{TI}$ values to within 2 kJ/mol of experimental values, whilst replicating the expected enthalpic signature of the binding interaction. However, it was not precise enough to allow calculation of $\Delta\Delta H_{TI}$ and $\Delta T\Delta S_{TI}$ via the finite difference method as the $\Delta\Delta G_{TI}$ range across the different temperatures available to simulation is too small. Although it is possible to increase the accuracy of the TI calculation, the computational cost is deemed to be too expensive for this method to be feasible when factored against the many ligands that require calculation. Thus, alternative methods of looking at enthalpic and entropic contributions are examined in the following chapters.



Maxwell’s Daemon

Chapter 3.0: Ligand Conformational Entropy

3.1.0. MUP binding entropy and ligand pre-organisation

A key finding from ITC data generated on a panel of n-alkanols obtained by Malham et al. (2005), was that $T\Delta\Delta S_i$ decreased in a linear fashion (-5.5 kJ/mol) and was correlated with the addition of a methylene group (§1.3.3) [36]. A natural hypothesis resulting from this observation is that longer ligands have more DOF (in the form of rotatable bonds) and these have the potential to be restrained upon binding to MUP. This represents a loss in conformational entropy (§3.1.1) and estimates obtained from various studies range from ~2 to 6 kJ/mol [166–172]. Panels of ligands with incremental differences in length are very good experimental models through which the thermodynamic impact of rotor addition can be assessed. As the only difference between successive ligands in the series is an additional rotatable bond, any perturbation to the system should be due to the effect of this structural modification alone.

Natural compounds often have highly rigid structures with conjugated ring systems that afford specificity and large binding affinities. Some authors suggest that pre-organisation of the ligand can yield better binding affinities by negating entropic losses [173,174]. This is a controversial concept that seeks to minimise unfavourable entropy by designing ligands that possess bonds with limited DOF. The energetic cost of "freezing" ligand DOF to better match its bound conformation is thus prepaid during its chemical synthesis. The expected entropic penalty is thus avoided and an increase in free energy is observed [175–178]. Theoretically, pre-organisation of the ligand only impacts DOF within the ligand itself and is highly dependent on the designed compound possessing shape complementary to the protein binding site [173]. This is not easy to achieve and there are examples of pre-organisation yielding unexpected results. e.g. Global entropy losses accompanied with increases in binding enthalpy [178,179]. Generating structural analogues that possess double bonds *in lieu* of one or more rotatable bonds is a simple method by which ligand rotor restriction and its effect on binding affinities can be studied. The impact of these systematic modifications can then be compared to a relevant control group. In the case of this chapter, two panels of pre-organised ligands structurally related to the n-alkanol panel are investigated to ascertain the entropic penalties they pay on binding to MUP. The concept behind these panels were originally developed and tested using ITC by

Malham (2012) and the computational work forming the backbone of this chapter was conducted to complement the experimental data [180]. Ligands in the first of these panels contain a double bond at the hydrocarbon terminus of the molecule and are known as terminal olefins. The second panel contains compounds with a cis-3-4 restriction and are known as 3Z-olefins.

Ligands within the terminal olefin panel ascertain whether a terminal restriction is equivalent to the removal of a rotatable bond. If this were the case, the entropic penalty associated with rotor restriction on binding would be avoided whilst the additional atomic mass theoretically favours enthalpic gains through increased protein-ligand van der Waals contacts. A positive entropic contribution from burying the additional surface area of the non-rotatable bond is also expected. Experimental ITC data only compared one 6C terminally restricted ligand to its unsaturated counterpart and obtained a favourable intrinsic entropy difference ($T\Delta\Delta S_i$) of 4.5 kJ/mol (§1.1.5). As the value is very similar to the average $T\Delta\Delta S_i$ loss (5.4 kJ/mol) associated with methylene removal, this gives weight to the hypothesis that the main source of entropy loss on binding is from the restriction of rotor DOF. However, no large affinity gains could be made by adopting this as a strategy in designing ligands as the gain was compensated by an intrinsic $\Delta\Delta H_i$ loss of 4.6 kJ/mol [180]. This chapter extends this work by verifying whether the observed entropic loss is replicable in other ligands possessing various lengths, and whether this is directly attributable to torsional restriction.

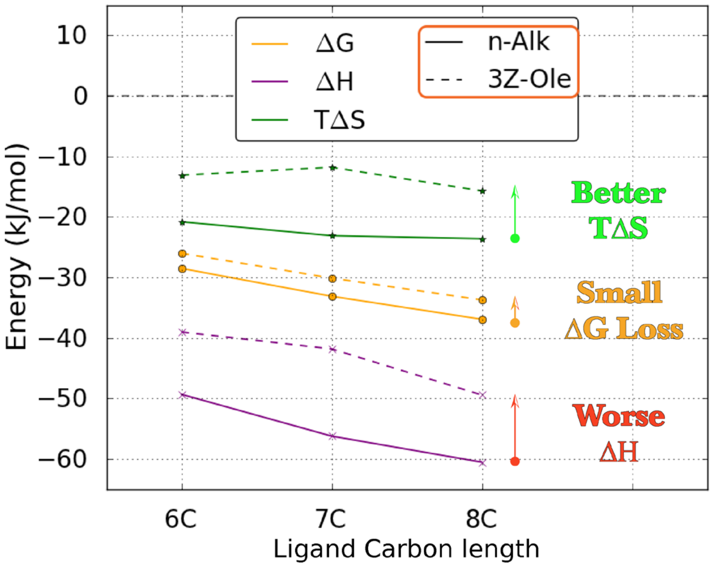


Fig.3.1. Entropy-enthalpy compensation effect between n-alkanol and 3Z-olefin ligand panels. Graph created using global ITC data from reference [180].

Globally measured ITC results obtained by Malham (2012) are reproduced in **Fig.3.1.** The graph illustrates the effect of EEC between 6C to 8C compounds in the 3Z-olefin panel and their n-alkanol counterparts. On binding to MUP, 3Z-olefins possess more favourable entropies (~ +8.9 kJ/mol) compared to their n-alkanol counterparts due to the structural modification of an alkyl to an alkenyl group. However, these entropic gains were again counterbalanced by an unfavourable enthalpic term of greater magnitude (~ +11.9 kJ/mol) that translated into a small net reduction in binding free energy (+2.9 kJ/mol). As the global values encompass ligand and protein desolvation terms, the decomposition technique outlined in §1.1.5 was utilised to isolate the solute-solute contribution. Where experimental data from air-solvent partition equilibria were unavailable, an additive group contribution technique was utilised to acquire summed theoretical values [94]. This could then be used to generate intrinsic thermodynamic values. It should be noted that these intrinsic values are based on the assumption that protein-protein interactions do not change when binding different ligands. Additionally, no account is made for differential expulsion of binding site waters. Whereas global ITC values show linear to almost linear trends, the intrinsic values were more variable and the *average* $T\Delta\Delta S_i$ gain on introducing a cis-3-4 double bond was 13.6 ± 4.2 kJ/mol, whilst the enthalpic penalty was 12.6 ± 4.3 kJ/mol. Contributions from ligand desolvation and other factors not accounted for in the intrinsic thermodynamic values resulted in the net global free energy of binding being unfavourable. The key question posed by this panel of ligands is why the entropic gain is more than double that of terminally restricted compounds. Whilst it is clear that an internal restriction could affect ligand dynamics to a greater extent by virtue of its central position, the exact reasons as to why this might be the case are more opaque.

3.1.1. Partitioning the entropy into its component parts

The root causes for the changes in thermodynamic quantities such as the entropy can be difficult to assess. Entropy is hard to measure accurately and the system's globally measured value is typically composed of disparate components, each making small contributions that may not be strictly additive in nature [173,181–183]. Using statistical mechanics to accurately calculate the absolute entropy or free energy of a system is problematic as it requires evaluation of the total phase space (Ω) that a system has access to. MD simulations generally have issues generating a sufficiently diverse number of microstates (Ω_i) that adequately describe Ω completely and usually only obtain a subset from this multidimensional topography. Long simulations, or enhanced sampling methods, are required to overcome the energetic barriers separating localised energy wells that constitute the potential energy surface of the system. Determination of bulk macroscopic quantities such as the temperature or pressure can be easily accomplished by calculating the ensemble average from instantaneous measurements of individual elements (Ω_i) obtained from a finitely sampled ensemble. However, absolute free energies and entropies require simultaneous assessment of the entire ensemble (i.e. Ω) and this is a computationally intractable problem due to its size. Moreover, if the simulated microstates do not fully represent Ω, the entropy or free energy difference will be inaccurate, as a representative sample does not take into account the contributions made by microstates not considered in the calculation [148,184–186].

As discussed in chapter 2.0, methods such as TI reduce the magnitude of the calculation by focusing on finding the relative ΔΔG between two states with small structural differences. As energetic fluctuations involving interactions between the alchemically mutated portions of the system are small, this method allows good estimates of the energetic disparity between states to be made. However, it is difficult to extend the same treatment to the entropy because the calculation takes into account interactions involving every atom (of solute and solvent) within the ensemble of structures and the magnitude of fluctuations caused by events such as solvent dynamics and amino acid conformational changes can be very high. Evaluating the integral for the 3N-dimensional configurational phase space (§2.1.3 & §3.1.2) is also a formidable, if not impossible, task for even small systems [148,187]. The dimensionality and size of the problem can be reduced by separating the entropy into component parts and it is generally accepted that these can be reconstructed to yield the total. Firstly, the global entropy can be partitioned into solute and solvent portions. In a protein-ligand binding interaction, the solute portion can be further subdivided into contributions from the protein and ligand. Either one of these will have entropy values associated with the principal rotations of the molecule, gross translational motion and internal vibrations (**eqn.3.1**). The latter is analogously known as the conformational or configurational entropy and experimental estimates for this value can be derived via NMR-derived order parameters [172,186,188–190].

$$S = S_{Trans\,+}\,S_{Rot\,+}\,S_{Conf} \qquad \textbf{(eqn3.1)}$$

Contributions to the conformational entropy can also be split into "hard" and "soft" DOF [191,192]. The former is composed of small deviations of bond lengths and angle bends away from their equilibrium positions and these high frequency motions are thus labelled "hard". These motions are not expected to contribute much, and the difference between bound and free states are often assumed to cancel. Moreover, the narrow probability distribution ranges typically obtained for hard DOF make them difficult to treat classically and quantum methods must be employed. However, soft DOF are composed of larger torsional oscillations, and are expected to dominate conformational entropy losses on binding [186,192].

Loss of internal DOF can result in the conformational entropy being a major contributor to global entropy loss on binding. For instance, NMR-derived order parameters of the binding entropy of 6 peptides (modelling the calmodulin binding domains of various target proteins) to the calmodulin receptor revealed that $\Delta S_{Conf}$was comparable in size to that of ΔS_{Solv}. Additionally, variation in the magnitude of ΔS_{Conf}between complexes were posited to be the primary modulator of protein binding affinity, whilst differences in ΔS_{Solv} tended to be invariant as the global binding entropy changed [193,194]. It is intuitively understandable that a large macromolecule such as a protein contains many DOF. Thus, small ΔS_{Conf} differences between free and bound DOF can accumulate to significantly affect the global entropy. However, it is less obvious how loss of ligand DOF can greatly impact the conformational entropy. Another study on the noncovalent binding of oligopeptides (netropsin and distamycin) to the minor groove of the DNA double helix, demonstrated that unfavourable ligand ΔS_{Conf} was a major contributor to the binding thermodynamics. Global entropies for these binders range from -39.7 to -54.0 kJ/mol depending on the identity of the DNA binding sequence. Using an *in silico* method, the total ΔS_{Conf} loss on binding was calculated as being -38.1 kJ/mol and -31.2 kJ/mol for netropsin and distamycin respectively [195].

3.1.2. Entropy calculation methods

A computational measure of conformational entropy can be generated from an ensemble of *n* microstates obtained by MD simulations via the Gibbs entropy formula, which calculates the variation in the density of different states occupied by the system (**eqn.3.2**).

$$S_{Conf} = -k_B \sum_{i-1}^{n} \rho_i \ln \rho_i \qquad \textbf{(eqn.3.2)}$$

If E_i is the energy of a given microstate, p_i can be calculated from **eqn.3.3**.

$$\rho_i = \frac{e^{\frac{-E_i}{k_B T}}}{\sum_{i=1}^{n} e^{\frac{-E_i}{k_B T}}} \qquad \textbf{(eqn.3.3)}$$

The total conformational entropy of a molecule such as a ligand is obtained by calculating the integral of the continuous probability density function formed by its trajectory within phase space. This requires both the momenta (p) and generalised coordinates (q) to describe the kinetic and potential energy terms respectively. These two components compose the total energy and can be calculated separately. The kinetic contribution is typically ignored as most biological systems simulated have constant mass and are studied under conditions of constant temperature. Thus, when calculating the relative entropy difference between two states, the kinetic contribution cancels. This is because it is also constant (on average) due to the equipartition theorem and the primary source of entropy change is derived from variation in the molecule's atomic coordinates. The total conformational entropy with only the potential term is defined in **eqn.3.5**, whilst both kinetic and potential components are taken into account in **eqn.3.4** [187,189,196].

$$S_{Conf} = -k_B \iint \rho(p,q) \ln \rho(p,q)\, dp\, dq \qquad \textbf{(eqn.3.4)}$$

$$S_{Conf(q)} = -k_B \int \rho(q) \ln \rho(q)\, dq \qquad \textbf{(eqn.3.5)}$$

There are several methods of calculating the conformational entropy such as the hypothetical scanning (HS) method, the nonparametric mutual information expansion (MIE) method, the quasi harmonic approximation (QHA)and its many variations [182,187,197–202]. The integral in **eqn.3.5** can rarely be assessed directly, because the dimensions of phase space accessible to proteins or even small ligands is too complex and thus intractable to computational calculation. Hence, the initial QHA method proposed by Karplus and Kushik (1981) assumes that atomic fluctuations can be described as a normalised multivariate Gaussian distribution that approximates a (quasi) harmonic distribution. The conformational entropy can then be calculated as a sum of entropies obtained from component quasiharmonic modes. A covariance matrix for two states (e.g. helix to coil transition) can be generated and the determinant (σ) utilised to encapsulate the conformational density - ρ(q). The equation for a single state is shown in **eqn.3.6**, whilst the relative entropy difference between state a and b is described by **eqn.3.7** [201].

$$S_{Config\,(\mathrm{q})} = \frac{1}{2} k_B \left[n + \ln((2\pi)^n \det(\sigma)) \right] \qquad \textbf{(eqn.3.6)}$$

$$\Delta S \simeq \Delta S_q = \frac{k_B}{2} \ln \frac{\sigma(b)}{\sigma(\mathrm{a})} \qquad \textbf{(eqn.3.7)}$$

An issue with this method is the requirement to convert Cartesian coordinates obtained from MD simulations to internal bond-angle-torsion (BAT) coordinates, as errors are introduced by approximations made within the Jacobian. This step is necessary to remove the centre of mass (COM) rotation of the molecule and can result in the σ being singular if neglected. Consequently, Schlitter (1993) advanced an *ad hoc* approximation that allowed the use of Cartesian coordinates via the addition of a mass weighted diagonal matrix. This method additionally allows calculation of the absolute entropy by including translational and rotational (T&R) terms [188,202]. In 1980, Karplus and Andricioaei reformulated the QHA to use Cartesian coordinates. This proved to be twice as efficient and was able to calculate the exact conformational entropy without the approximations present in the Schlitter method. Additionally, values for the heat capacity, vibrational free energy and enthalpy could also be calculated [200].

The Schlitter method models each internal DOF as a quantum harmonic oscillator, whilst the QHA generates a multivariate Gaussian distribution to functionally describe the complex multiminima of the potential energy surface of a macromolecule. The latter approach is meant to better account for the multimodality of the data and potential anharmonicity. However, the QHA has been found to excessively smooth the distribution and can potentially introduce significant errors into the calculation by merging multiple energy wells into a single well. A better approximation can be obtained by calculating the entropy for each DOF separately and then summing the contributions. However, this is a 1st order approximation which does not take into account correlated motions that act to lower the calculated entropy. Methods that do not take into account higher order correlations only provide an estimate of the maximum possible entropy of the system. Covariance based methods take into account linear correlations only and are therefore considered to be 2nd order approximations. There are also other potential sources of error due to the normal distribution being a function that tends to maximise the entropy. The necessity of removing rotational and translational DOF by superimposing structures obtained from MD simulations can also introduce errors up to 80 J/mol/K [188,203].

Histogramming methods such as that used in the mutual information expansion (MIE) put forward by Killian et al. (2007) do not assume an underlying functional form to the probability distribution function (PDF) and are theoretically capable of providing entropy estimates that take into account correlations up to 3rd order or higher [182]. Practically, the method is typically restricted to a 3rd order ceiling, as the construction of 3D or higher dimensional PDFs require increased sampling density. This is necessary to alleviate the issue of a finite number of data points inadequately filling the greater number of bins created within higher dimensional volumes. Furthermore, this incurs a substantial cost in storage capacity which can be limiting when simultaneously studying the trends between several systems of interest. At its heart, MIE calculates the Shannon entropy using **eqn.3.2** and converts this into a physical entropy value by multiplying it by the gas constant or Boltzmann constant. $-\ln \rho(q)$ is known as the self-information of a distribution and the mutual information is a measure of the amount of correlation between marginal probability distribution functions that make up the full PDF of the system. Thus, the total entropy can be calculated by summing tractable 1D and 2D marginal entropies and then applying higher order mutual information corrections [182,183,204].

There are other methods that infer the total PDF via the use of kernel density estimators which can take into account correlations higher than 3rd order. Of note is the *k*th nearest neighbour entropy estimator of Hnizdo et al. (2008) which when combined with MIE was capable of providing up to 6th order correlation corrections (albeit with greatly increased computation time) [183]. The estimator deals with the rarification of data points when considering higher dimensions by using an adaptive bin width, that modulate its size around a sample-point centric hypersphere so that it always contains k (e.g. 1, ..., 5) nearest neighbours. The appropriate value for k is adjudged by its ability to make a smooth estimation whilst keeping the entropy calculation as localised as possible around each hypersphere kernel. This unbiased nonparametric method can better deal with anharmonicities and the multimodality of a given PDF in a better fashion than the QHA, but struggles to achieve convergence when the dimensions (i.e. internal DOF) of the system exceeds 10 to 15 [183,205].

The minimally coupled subspace approach also uses nonparametric kernel density estimation (based on adaptive anisotropic kernels) and can calculate the configurational entropy of systems possessing up to 45 dimensions. Larger systems such as proteins are decomposed by separating highly coupled DOF via a linear orthogonal transformation to Cartesian coordinates using Full Correlation Analysis. These coordinates are then clustered according to their correlation coefficients and the mutual information within each cluster is minimised. Inter-cluster correlations are ignored as it is assumed that the initial assignment of coordinates eliminates the majority of the significant terms. Oversized clusters with dimensions greater than 15 are subclustered so that the entropy calculation is more manageable and then summed whilst taking into account MIE corrections. The authors also proposed that the method could be extended to include a form of the Schlitter formula to account for the narrow distributions of stiffer DOF (e.g. bond lengths) [188,206,207].

3.1.3. Objectives

As ligand specificity in MUP is principally governed by non-specific apolar interactions, it is a simple model system in which the systematic modification of ligands can be studied. The experimental data obtained from the binding of primary alcohols suggest that global thermodynamic values scale in a linear fashion on extending ligand length with additional methylenes. This is indicative of an additive system whose global binding characteristics can be probed via elementary modification of the ligand alone.

In order to open an aperture that captures bound ligand dynamics in a feasible amount of time, three panels of ligands were scrutinised by means of long MD simulations. This chapter specifically examines the entropic aspects of ligand binding to MUP and the key questions to be assessed are listed below.

1. “Top down” experimental decomposition indicates that reductions in $T\Delta S_i$ of 5.5 kJ/mol are attributable to the penalty of restricting a C-C rotor. As estimates for torsional restriction of such a rotor are ~2-6 kJ/mol, it is of interest to validate whether the primary source of entropy loss on binding n-alkanol ligands is traceable to this specific component and whether "rotamer counting" is a valid approach in drug design.

2. Pre-organisation of the ligand can easily be accomplished by introduction of a double bond and a systematic study on the binding of nonpeptidic ligands to μ-opioid receptors demonstrated that insertion of a single stereo-specific restriction has the potential to increase affinity by an order of magnitude [208]. This category of ligand alteration is not expected to excessively perturb the system being studied and offers one of the finest levels of granularity that is practically accessible to synthetic chemistry. Though this is a simple unit modification the choice of restriction position, factored with ligand length, may create additional dynamic effects that impact binding affinity in unexpected ways. Experimental work suggests that these simple alterations can be reduced to group additive effects, but variance in some of the results suggests that further analysis is needed to verify the findings.

The equivocal results obtained from pre-organisation demonstrate that whilst this approach might be theoretically sound, extensive benchmarking of proposed modifications must be carried out in order to reap acceptable affinity gains. An *in silico* approach is an ideal method through which this can be accomplished in a relatively cheap manner whilst offering an atomistic level of detail and explanation. The question this chapter chooses to address, with regards to the 3Z-olefin panel, is why these ligands have better entropies of binding compared to their n-alkanol counterparts. In order to begin answering it, the contribution of the conformational entropy is assessed.

3. Only one Ligand with a terminal restriction was experimentally tested by ITC and the $T\Delta\Delta S_i$ gain for modifying hexan-1-ol to hex-5-en-1-ol was found to be 4.5 kJ/mol. Because of the similarity to the value obtained for rotor removal (5.4 kJ/mol), this result suggested that the principal source of global entropy losses on binding is due to decreased ligand DOF. The computational analysis will attempt to reproduce this result by examining the contribution of internal torsional DOF. Additionally, the binding of longer ligands will also be assayed to assess if the effect is consistently repeated within all members of the panel.

4. Most *in silico* studies on the binding of ligands to MUP have focused on one to four compounds and while the resulting analyses have been revealing, they do not fully meta-analyse the underlying mechanisms of promiscuous binding in this protein. As several panels of ligands are being studied, we are more interested with trends in the data versus pinpoint accuracy and will only be calculating 1st order torsional entropies as the computational cost in terms of storage space and processing time is too prohibitive to analyse 12 ligands using methods such as MIE. A measure of the extent of linear correlations within the ligand will be obtained using principal component analysis (PCA).

5. In order to ascertain the quality of the histogram method used to calculate the entropy, the method will be cross-validated against data obtained from tight ligand binding to HIV protease. A number of studies have been conducted on this system and it will provide an informative counterpoint to the results obtained from MUP ligand binding.

3.2.0. Methods

3.2.1. MUP ligand parameterisation

Ligands depicted in **Table.3.1** and **Fig.3.2** were parameterised using Gaussian 03 and the R.E.D. III suite of tools with the 6-31G* basis set in the manner described in §2.2.2 [121,150,209]. In this thesis, the following naming convention is used. Primary saturated alcohols are referred to as n-alkanols; unsaturated alcohols with terminal restriction are called terminal olefins and unsaturated alcohols with a cis-3-4 double bond are labelled 3Z-olefins.

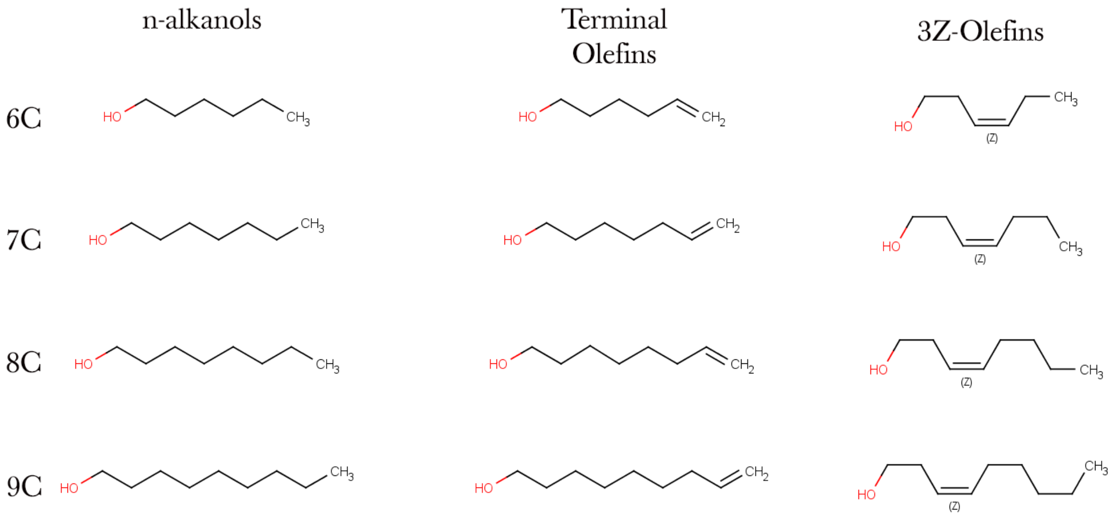


Fig.3.2. Structures of ligands composing the three primary alcohol panels.

Amber parameter files involving ligands complexed with protein were based on crystal structures obtained from the PDB files listed in **Table.3.1**. Where these were unavailable, the ligand from the closest matching PDB file was mutated to the desired ligand. Unpublished alken-1-ol crystal structures were kindly provided by Dr Richard Malham. The Duan et al. (2003) force field parameters were used for all simulations [151].

	Abbrev.	Ligand	PDB ID
Primary Alcohols	**hex**	hexan-1-ol	1ZNE
	hep	heptan-1-ol	1ZNG
	oct	octan-1-ol	1ZNH
	non	nonan-1-ol	1ZNK
3Z-olefins	**3c6**	(Z)-hexa-3-en-1-ol	unpub
	3c7	(Z)-hepta-3-en-1-ol	unpub
	3c8	(Z)-octa-3-en-1-ol	unpub
	3c9	(Z)-nona-3-en-1-ol	(3c8)*
Terminal Olefins	**ThX**	hex-5-en-1-ol	1ZNE*
	ThP	hept-6-en-1-ol	1ZNG*
	ToC	oct-7-en-1-ol	1ZNH*
	TnO	non-8-en-1-ol	1ZNK*

Table.3.1. Breakdown of three ligand panels with three letter abbreviations and PDB identifiers. Ligands marked with an asterisk do not have solved crystal structures and the ligand within the listed PDB ID was mutated to the appropriate compound.

3.2.2. HIV-I protease ligand parameterisation

Darunavir was parameterised alongside other sulfonamide PIs using a fragment based approach to ensure that ligand backbone and common groups had consistent charge derivations. This method is particularly suitable for large molecules that can present conformations that possess intramolecular, non-bonded gas phase interactions which are unsuitable for charge derivation [210]. Thus, the ligand was split into constituent parts centred on a central scaffold onto which X, Y and R groups could be interchanged to produce Amprenavir, Darunavir, GS-8374 or TMC-126. Darunavir was created by combining fragments marked with an asterisk (**Fig.3.3**). A two stage RESP fit was utilised to ensure that chemically equivalent atoms retained the same charge as described in references [211,212]. The same references describe the use of blocking groups which are surrounded by boxes in **Fig.3.3.** Briefly, charges are derived for the different fragments separately and a subsequent step uses Lagrange constraints to force the charge on the blocking groups to become zero upon marrying fragments during the final fit. This is so that the net charge on the fused fragment is an integer. Combining the central ligand scaffold to either one of the R groups uses ACE ($COCH_3$) and NME ($NHCH_3$) blocking groups around the peptide bond. This is a well known, established procedure that is used to build the terminal and central fragments in AMBER force field libraries. The identity of the other (Y and R) blocking groups had to be established via iterative modifications that were assessed in terms of the final quality of fit. The metric to determine this was based on a comparison of charges derived with and without constraints. Multiple orientations were used to improve the fitting procedure as described in reference [209]. Charges obtained with this approach are depicted in **Fig.A1.8**. Standardisation of the common backbone elements would not have been possible if this charge derivation approach had not been used.

The crystal structure (2HS1) of HIV-I protease bound to Darunavir included all crystallographic waters and the catalytic aspartates were monoprotonated (ASP25), as the literature reported that simulation instability could occur if the wrong protonation state was utilised [213–217].

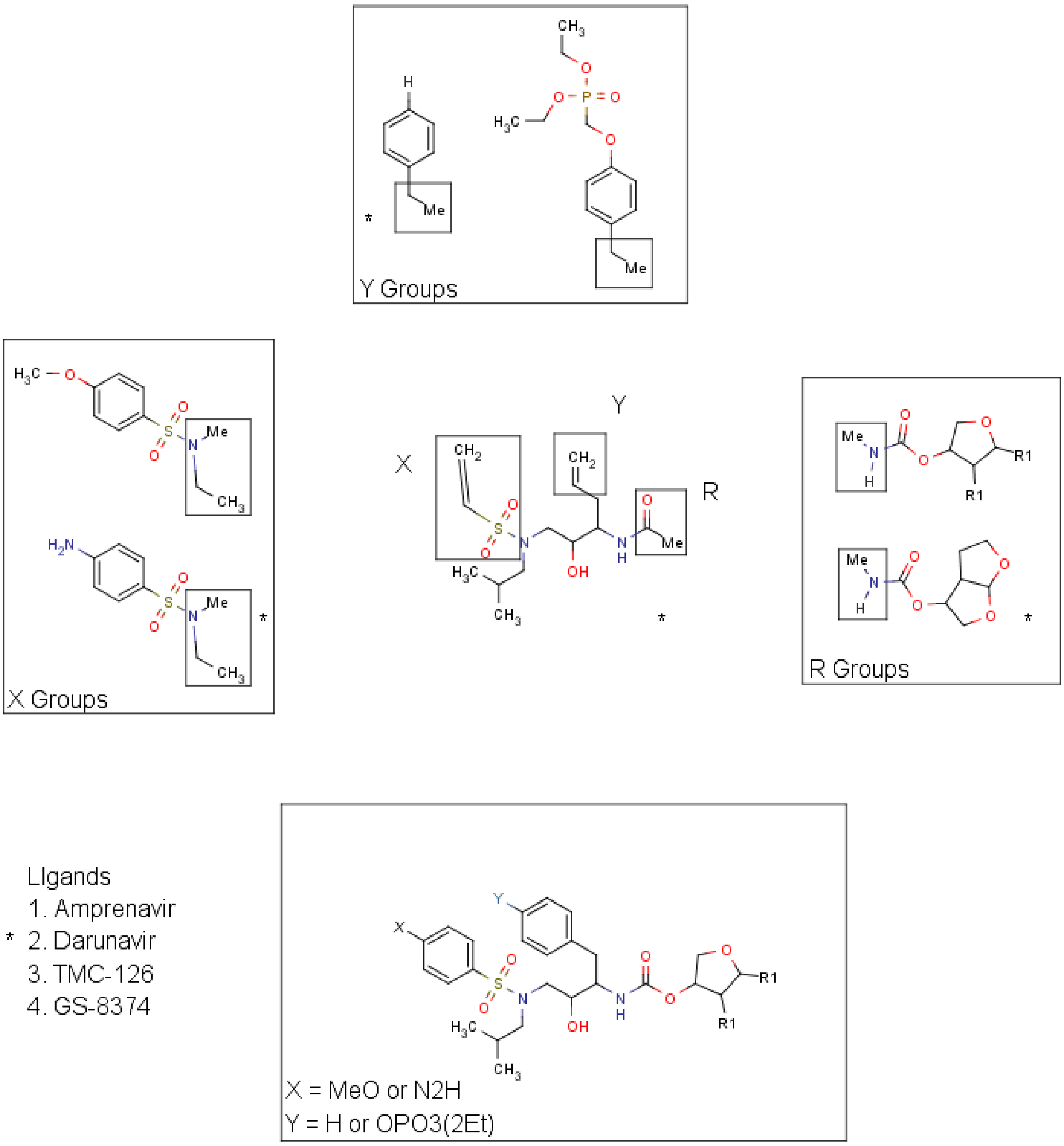


Fig.3.3 - Fragment based reaction scheme to derive partial charges for HIV-I protease sulfonamide based inhibitors. Fragments used to make Darunavir are marked with an asterisk. Me stands for methyl; and Et, ethyl.

3.2.3. Long MD simulation setup

Long molecular dynamic simulations of a 100 ns were carried out in the following manner for free and bound species.

Bound: The equilibration phase was run with a single minimisation stage, followed by a heating stage at constant volume (NVT) and a final step at constant pressure (NPT). However, this created problems in terms of protein stability in the case of 3Z-olefin panel. Therefore, a gentler 2-step approach was taken for minimisation and NVT in all the simulations. The first minimisation step was run for 2,000 steps using a nonbonded cut-off of 12 Å. This initially consisted of 500 steps of steepest descent and the remaining steps utilised the conjugate gradient method. All protein and ligand atoms were restrained with a strong harmonic restraint of 30.0 kCal/mol Å^{-2}, whilst the solvent was allowed to "melt". The second minimisation phase used the same settings as the first, but turned off all restraints. The first step of constant volume simulation modulated the temperature from 100 to 200 K for 20 ps using weak coupling (Berendsen thermostat) while using the restraint of 30.0 kCal/mol Å^{-2}. Subsequently, another 20 ps NVT simulation was run with the restraint switched off and the system was heated from 200 to 300 K. Short NVT runs were used to minimise the possibility of "vacuum bubbles" forming as the ordered crystal lattice of initial waters placed by tleap underwent melting. The density and volume were then allowed to equilibrate in a single NPT stage of 40 ps. All equilibration stages were carried out using the Sander module from AMBER. The production stage used PMEMD and was run in 4 ns segments for a total of 100 ns. SHAKE was turned on in all simulation steps (apart from minimisation) so that a 2 fs time step could be used and care was taken to assign unique random number seeds to all stages.

Free: Stages with restraints used for bound simulations were omitted and the equilibration and production settings were the same as that used for the bound state.

Six repeats were performed for every simulation set (Bound and Free). Half of the simulations used the CPU code and the other half used GPU code. In all analysis steps, differences between the two sets of repeats were scrutinised and no disparities could be detected [132,133]. Ideally, longer simulations of 200 to 300 ns were desired, but because the GPU's were unavailable for full production use till the last half of 2013 this work uses the first 100 ns for analysis. The simulations have been extended by an additional 100 ns and this will be used to confirm all findings at a later date. Special thanks goes to Emanuele Paci, who kindly donated the use of one of his GPU's which did the bulk of the work to make the total number of repeats up to six.

3.2.4. The entropy calculation

The conformational entropy is calculated by measuring the number of microstates in the system. An early method of doing this was through counting rotamer distributions. The more states a rotamer samples, the higher the conformational entropy [181]. Thus, the entropy difference is calculated from the difference in the density of states between the free and bound species. This is comparable to the intrinsic entropy obtained via thermodynamic decomposition as the conformational entropy does not take into account any solvent terms, even though ligand dynamics within a MD simulation is perturbed by collisions with water molecules. The conformational entropy for ligands binding to MUP were calculated as follows: All snapshots from the trajectories were analyzed using the ptraj module of AMBER Tools 13 [150]. Water, ions and, in the case of the bound simulations, protein residues were removed. The ptraj dihedral command was used to output dihedral angles using heavy atoms from the ligand backbone as a point of reference, whilst dihedrals near the termini also included a terminal hydrogen atom.

The resulting angles for each ligand species were binned according to population. The resulting 1D histogram can be viewed as the inverse of a typical potential energy map and allows direct calculation of the entropy without prior assumptions about the underlying functional form of the data. The torsional entropy (TS_{Tor}) was calculated for every dihedral using the statistical mechanical formula (**eqn.3.8**) proposed for torsion angles by Edholm and Berendsen (1984). Physical units of kJ/mol were obtained from the dimensionless logarithm by multiplication by -RT. The contributions of angle bends and bond lengths is reported to be small and therefore assumed to cancel between bound and free states [192].

$$S_{Tor} = -RT\sum_{i=1}^{bins}\rho_i \ln\frac{\rho_i 2\pi}{\Delta} \qquad \textbf{(eqn.3.8)}$$

The probability distribution of each bin (ρ_i) is given by **eqn.3.9**, where n is the total number of samples and n_i represents the number of samples in a given bin. The bin width in radians (Δ) is used to correct systematic errors caused by use of a fixed interval size [187].

$$\rho_i = \frac{n_i}{n}\frac{1}{\Delta} \qquad \textbf{(eqn.3.9)}$$

TS_{Tor} values obtained from all 1D marginal PDFs in a ligand species were summed to yield a 1st order torsional entropy estimate (TS_{Conf}) and then subtracted using **eqn.3.10** to give the relative difference in conformational entropy on binding ($T\Delta S_{conf}$).

$$T\Delta S_{Conf} = S_{Conf\text{-}Bound} - S_{Conf\text{-}Free} \qquad \textbf{(eqn.3.10)}$$

As this method is only suitable to describe classical dynamics, the vibrational entropy occurring in the deep energy wells possessed by double bonds were not calculated as they are best described by a quantum mechanical treatment.

3.2.5. Error analysis

The entropy calculations estimate the population mean using a sample size of 100,000 snapshots per simulation. The value yielded by the calculation can be overestimated or underestimated depending on the population of microstates furnished by the simulations and the minimum number of data points required to generate a reliable ligand entropy estimate is uncertain. The number of independent repeats of ligands bound to MUP is necessarily small (six) due to the computational cost of accumulating and storing data. Scrutiny of dihedral distributions in amino acids reveals that, for the most part, a single simulation samples the total available conformational space adequately. However, conformations populated by certain dihedral angles are separated by larger energy barriers that reduce transitions between available energy minima within 100 ns of simulated time. Combining individual 100 ns simulations produces a distribution that is theoretically a closer approximation to the true ergodic distribution, and thus should provide a better entropy estimate. This effect is illustrated in **Fig.3.4** and **Table.3.2**.

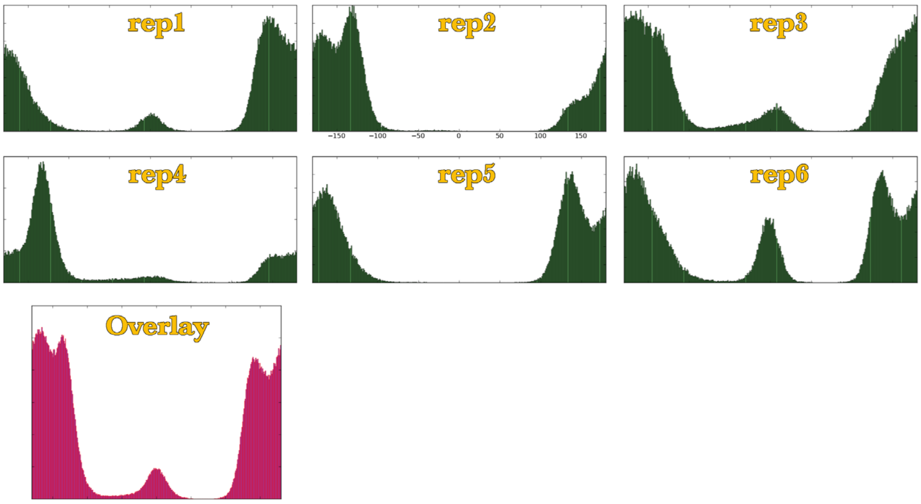


Fig.3.4. Dihedral angle distribution for an amino acid bond in MUP. Six 100 ns repeats are coloured green, whilst an overlay representing the aggregate composing 600 ns is coloured magenta.

The difference between the averaged entropy obtained from six 100 ns simulations and the entropy estimate from a single concatenated 600 ns simulation is within 0.1 kJ/mol for a single protein dihedral possessing a complicated conformational landscape (**Table.3.2**). Though this is a small disparity, the difference systematically accumulates

	Entropy Estimate (kJ/mol)
Rep 1	0.551
Rep 2	0.494
Rep 3	0.696
Rep 4	0.573
Rep 5	0.519
Rep 6	0.721
6x 100ns Avg	**0.59**
1x 600ns Calc	**0.68**

Table.3.2. Entropy estimates from MUP simulations for a protein dihedral angle; an averaged entropy obtained from the 6x 100 ns simulations; and an entropy calculated from a single 600 ns concatenated trajectory.

when summing the contributions from multiple dihedrals. Therefore, the more ergodic value calculated from the combined sample is always used in this analysis. This is considered more accurate as more states of the measured observable are accounted for in a single calculation and the entropy estimate improves because Ω is better approximated. **Fig.3.4** also shows that a single simulation does not always fully represent the population distribution attainable with increased sampling from multiple independent repeats. Thus, to obtain a measure of the spread of mean entropic values generated from sample distributions (that may not fully represent the population distribution), the standard error is used to quantify the error inherent in the method [205,218]. Errors were propagated by taking the square root of the sum of the squares.

3.2.6. Principal Component Analysis

PCA allows reduction of data dimensionality inherent in molecular motions via the construction of a 3N x 3N covariance matrix, where N is the number of atom in the system multiplied by three spatial Cartesian DOF. Before constructing the matrix, T&R DOF are removed by fitting the structures obtained from MD simulations to a reference structure. The covariance for two variables is given by **eqn.3.11**, where X_i represents the instantaneous position of a specific atom, whilst X with the overbar is the mean atom position for that DOF.

$$C_{ij} = \left\langle X_i - \bar{X}_i \right\rangle \left\langle X_j - \bar{X}_j \right\rangle \qquad \textbf{(eqn.3.11)}$$

The matrix is then diagonalised to generate eigenvectors that describe the direction of the linear combinations of modes of motion, whilst the corresponding eigenvalues contain information regarding the magnitude of that motion. In practice, the first few eigenvectors possessing the largest eigenvalues are referred to as the principal components as they are able to capture the majority of motion inherent in the trajectory. Thus, projection of the largest principal component modes back onto the trajectory allows visualisation of the collective, correlated motions of atoms [219–223]. PCA was carried out with programs from PCAsuite and the inbuilt Gaussian RMSD algorithm was used to fit all trajectories to the post-minimised crystal structure prior to diagonalisation of the covariance matrix and subsequent analysis [224].

3.2.7. Additional analysis

The radius of gyration was calculated with ptraj [150]. Creation of graphs was accomplished by using python scripts that made extensive use of NumPy and matplotlib [225,226]. Visualisation of molecular models utilised the program UCSF Chimera [227]. Reduction of 3D structures to 2D representations was accomplished with MarvinSketch and LigPlot+ [228,229].

3.3.0. Results & Discussion

3.3.1. Simulation production stage

Simulation parameters such as energy, temperature, density and pressure were monitored to check for instability during the production stage (**Fig.3.5**). Values oscillate around stable averages as expected.

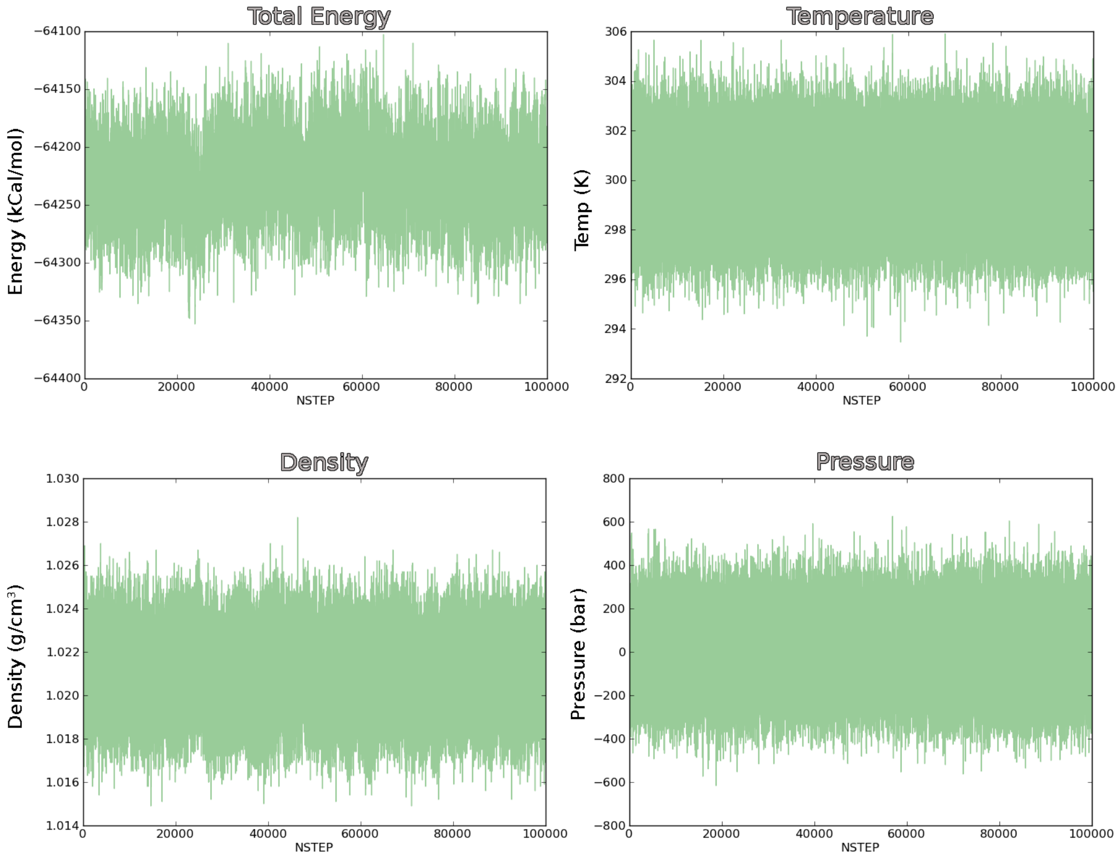


Fig.3.5. Representative simulation production stage plots from a simulation of heptan-1-ol (clockwise), Total energy; Temperature; Pressure; and Density.

The conformational space explored by each 100 ns simulation was visualised using principal component analysis. Projections from the first two principal component modes (PCM) were plotted against each other, so as to give an indication of the degree of conformational sampling the ligand undergoes within each simulation, for free and bound states (**Fig.3.6**).

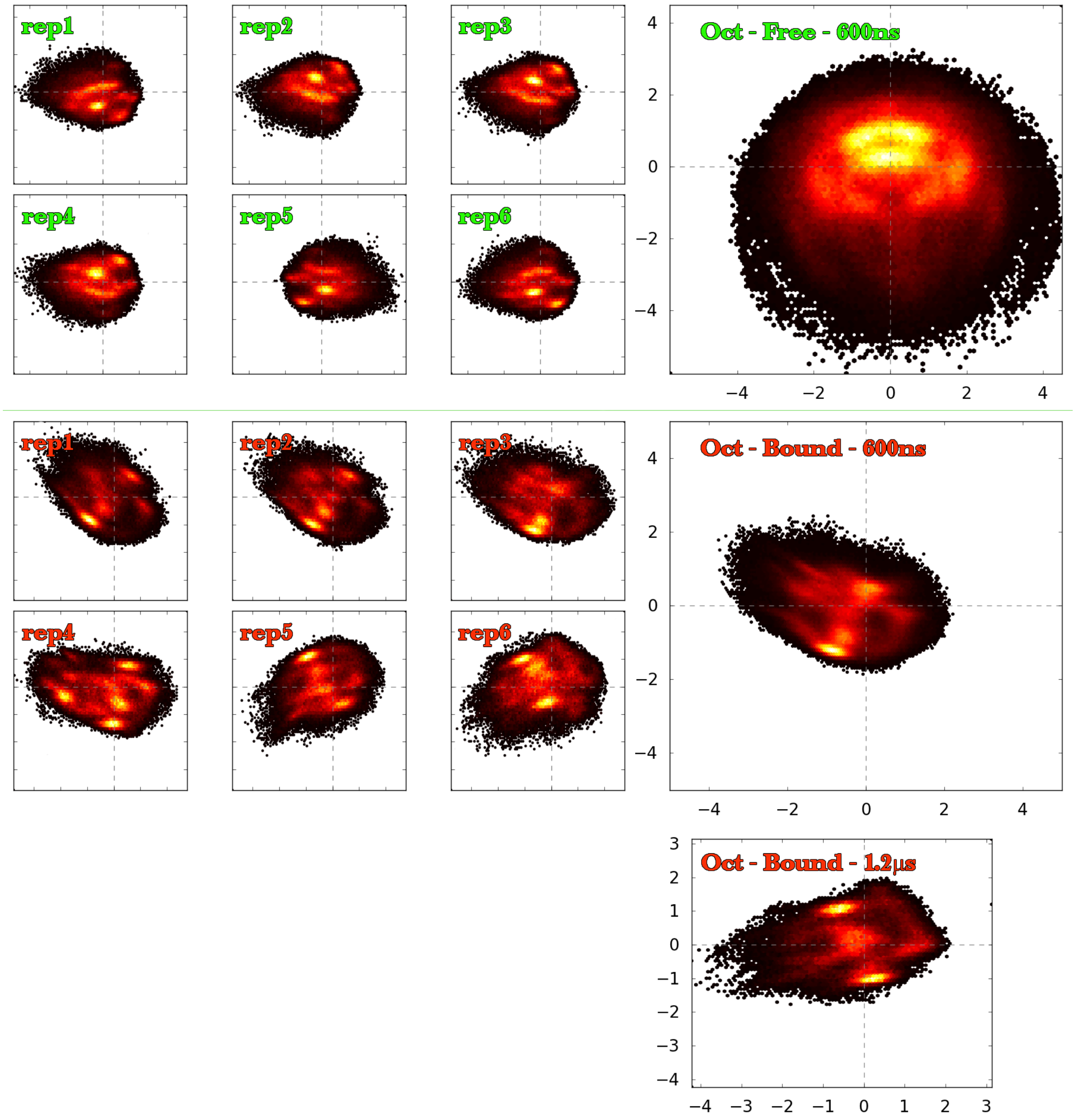


Fig.3.6. Density plots of PCM1 versus PCM2 that give an indication of the modal space explored by oct. Principal components were calculated for the ligand without hydrogen atoms. The 6 smaller graphs show the space explored by individual repeats, whilst the largest graph on the right plots principal components calculated from a trajectory made by concatenating all the repeats for that particular species. Note that some of the individual repeats are rotated with respect to each other. **(a)** shows the free distributions of oct ligand heavy atoms coloured green. **(b)** bound oct coloured red **(c)** a plot of bound oct generated from a single 1.2μs simulation.

Concatenating the PCMs from individual simulations generates a density plot that represents an aggregate sampling time of 600 ns (**Fig.3.6.a large inset**) and produces a distribution that is more ergodic than its parts alone. This demonstrates that a single shorter simulation does not adequately sample all possible ligand conformations. Individual simulations in the bound state have distributions that more closely match the shape of the aggregate simulation as mobility of the ligand is somewhat restricted in the binding pocket. Whilst there is overlap in the space explored within individual repeats of a simulation, it can be seen that the ligand explores conformational space in a different manner to that in the free state. Despite some reduction, there is still a considerable amount of variability in the bound state. This suggests that the ligand experiences considerable residual motion whilst bound and this has implications with regards to conformational entropy loss on binding.

Entropy calculations using the larger concatenated simulations should yield more accurate values than averaging smaller samples as the microstates of the system are better represented. A single long MD simulation of bound oct was run for 1.2 µs to assess the difference in sampling. The same areas are visited as those in the aggregate 600 ns simulations. Whilst the latter has broader, more diffuse coverage, the former contains two areas of higher intensity than the aggregate simulations (**Fig.3.6.b-c**). Despite having a greater density of points, the single 1.2 µs simulation has less variability than multiple independent simulations and could potentially under/over represent conformations as the system tends to dwell in certain areas of phase space. Hence, it is better to run greater numbers of simulations to sample phase space as efficiently as possible with the computational time available. This has also been observed in other studies [221,230]. For principal component density plots of the protein alone, please see **Fig.A1.6-7.**

3.3.2. Ligand radius of gyration

It is well understood that the potential energy of butane moves through three minima and maxima as the dihedral angle is varied as a result of steric hindrance and hyperconjugation. The most stable (-20.9 kJ/mol) trans (or anti) conformation is at an angle of 180°, whilst the two less populated gauche states (-60° and +60°) are less stable (-17.1 kJ/mol). Because the kinetic barriers to rotation at room temperature are not greater than the thermal energy available (84 kJ/mol), the molecule rapidly interconverts between different conformations when it is free in solution and tends to spend more time in wells with greater energetic favourability [231,232]. It should be noted that studies of thermal desorption of alkanes from solid surfaces indicates that short acyclic aliphatic hydrocarbons from hexane to octane do spend a significant amount of time in trans conformations, but dihedrals (particularly near the end) of molecules longer than nonane start spending increased time in gauche conformations [233]. Computer simulations have observed that linear alkanes possessing more than 22 carbons can even fold back to form collapsed structures such as hairpin, double-hairpin and broken paperclip motifs [234].

It is quite difficult to directly characterise the shape of a flexible polymer whose conformation constantly shifts from one frame to the next within an MD simulation. A standard method from polymer physics that indirectly assesses flexible ligand dynamics utilises the radius of gyration squared (Rg^2). This was considered a better method than merely measuring the end-to-end distance, as it takes into account the distances of every heavy atom to the polymer's centre of mass (COM). For linear molecules a large Rg^2 value corresponds to the ligand being extended, whilst small values correspond to more compact conformations.

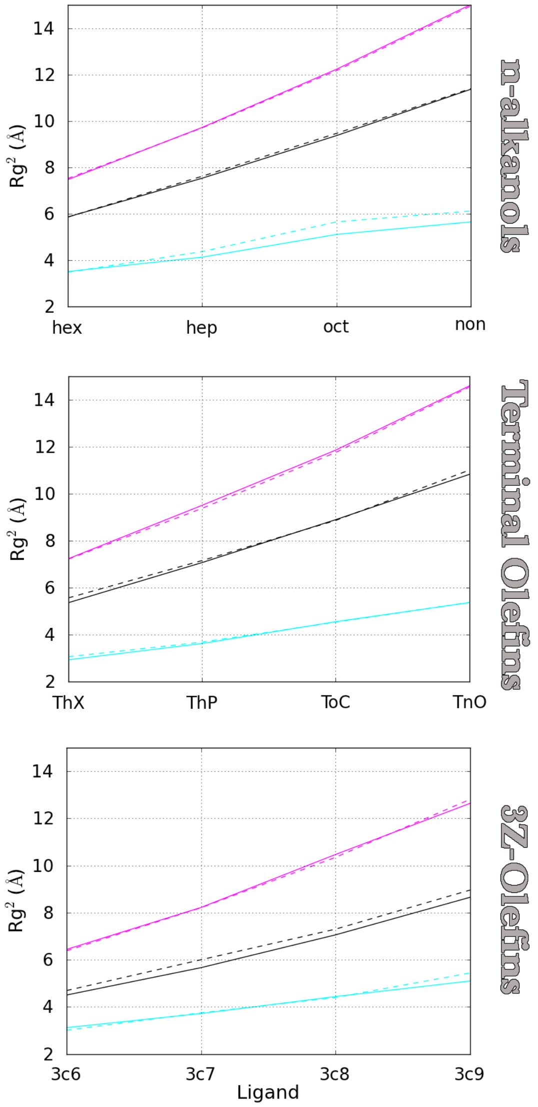


Fig.3.7. Rg^2 values for three panels of ligands (from left to right) corresponding to n-alkanols, Terminal olefins and 3Z-olefins. Minimum values are coloured cyan; means black; and maximum magenta. Solid lines and dashed lines represent the free and bound states respectively.

The mean Rg^2 increases in a linear fashion commensurate with increasing carbon chain length for both free and bound species (**Table.3.3** & **Fig.3.7**). The n-alkanol panel possesses the highest Rg^2 values, whilst ligands within the terminal olefin panel differ from their saturated counterparts by ~ -0.49 Å. As expected, 3Z-olefins have the smallest Rg^2 values due to the geometric restraint imposed by the central cis-3-4 double bond. This effectively shortens the maximum attainable ligand length compared to that observed in terminal olefins.

The disparity in mean Rg^2 between free & bound states is minimal for all three panels and this is the second indicator that the ligand explores a similar conformational space upon binding to MUP to that when unbound. Though all 3Z-olefins exhibit a slightly larger mean Rg^2 value in the bound state, the discrepancy is small. The disparity between the smallest and largest Rg2 values for free and bound species are also minor (**Table.3.4** & **Fig.3.7**).

	n-alkanols mean Rg^2 (Å)			
	hex	**hep**	**oct**	**non**
Mean F	5.86 ± 0.60	7.53 ± 0.81	9.37 ± 1.07	11.37 ± 1.35
Mean B	5.86 ± 0.61	7.62 ± 0.74	9.47 ± 0.81	11.40 ± 1.20
	Terminal Olefins mean Rg^2 (Å)			
	ThX	**ThP**	**ToC**	**TnO**
Mean F	5.37 ± 0.61	7.07 ± 0.84	8.90 ± 1.05	10.83 ± 1.36
Mean B	5.57 ± 0.57	7.16 ± 0.76	8.87 ± 0.90	11.01 ± 1.19
	3Z-olefins mean Rg^2 (Å)			
	3c6	**3c7**	**3c8**	**3c9**
Mean F	4.50 ± 0.45	5.66 ± 0.60	7.06 ± 0.81	8.65 ± 1.11
Mean B	4.69 ± 0.46	6.00 ± 0.63	7.30 ± 0.84	8.96 ± 1.04

Table.3.3. Mean Rg^2 with standard deviations for all 3 ligand panels.

	n-alkanols extrema Rg^2 (Å)								
	hex	**hep**	**oct**	**non**			**hex**	**hep**	**oct**	**non**
Min F	3.51	4.13	5.11	5.65		**Min B**	3.48	4.37	5.65	6.12
Max F	7.47	9.72	12.23	15.02		**Max B**	7.52	9.70	12.17	14.96
	Terminal Olefins extrema Rg^2 (Å)									
	ThX	**ThP**	**ToC**	**TnO**			**ThX**	**ThP**	**ToC**	**TnO**
Min F	2.94	3.63	4.56	5.37		**Min B**	3.07	3.69	4.54	5.38
Max F	7.24	9.51	11.86	14.59		**Max B**	7.22	9.38	11.76	14.54
	3Z-olefins extrema Rg^2 (Å)									
	3c6	**3c7**	**3c8**	**3c9**			**3c6**	**3c7**	**3c8**	**3c9**
Min F	3.12	3.72	4.44	5.10		**Min B**	3.01	3.75	4.40	5.44
Max F	6.44	8.22	10.47	12.64		**Max B**	6.38	8.21	10.35	12.81

Table.3.4. Minima and maxima values for all 3 ligand panels, representing compact and extended conformations respectively.

Further analysis of the differences between free and bound species in all 3 panels was made by comparing the Rg^2 distributions on a per-ligand basis (**Fig.3.8**). From this analysis several conclusions can be drawn.

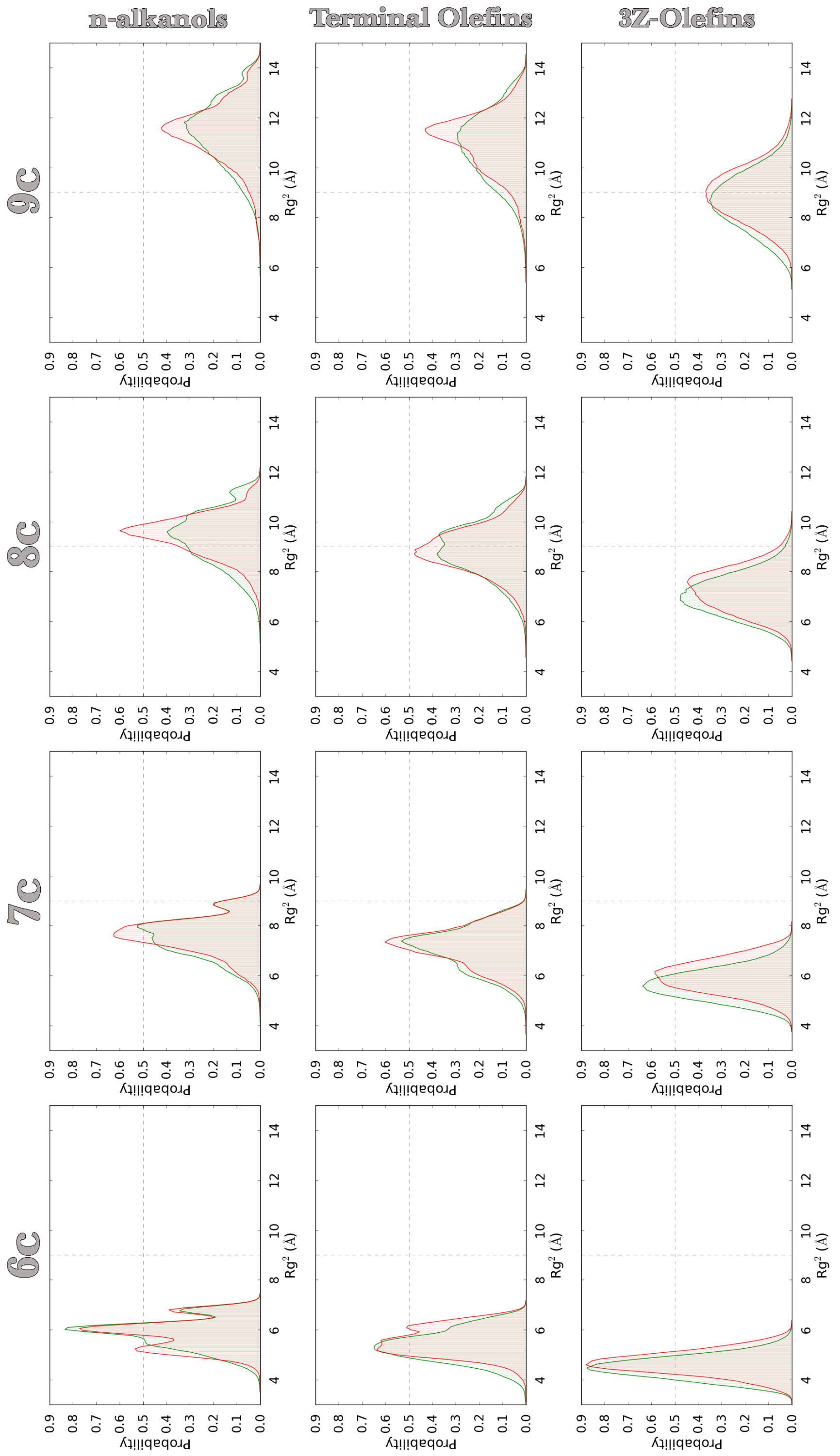


Fig.3.8. Rg^2 distributions for three panels of ligands (from top to bottom) corresponding to n-alkanols, Terminal olefins and 3Z-olefins. The leftmost graph depicts 6C ligands, whilst the rightmost contains the data for 9C ligands. The free and bound states are coloured green and red respectively.

1. Both free and bound states possess a continuous Rg2 distribution that is indicative of the ligand being able to adopt a variety of conformations. As bound Ligands have distributions broadly similar to those in the free state, this strongly suggests that the ligand is not tightly bound in the pocket and is able to adopt the majority of conformations attainable whilst free in solution.

2. The Rg^2 distribution for both states broadens with concomitant peak flattening as carbon chain length increases due to greater ligand conformational exploration. In the case of the n-alkanol and Terminal olefin panels, the general effect is that the bound state has a slightly reduced density of populations in very extended or compact conformations compared to that of the free. On the other hand, the trend observed within the 3Z-olefin panel is a "right-shifting" of the bound populations so that more extended conformations are marginally favoured.

3. The n-alkanol panel has the most complicated Rg^2 distribution because of the absence of restricting double bonds. Terminal olefins are broadly similar, but the distribution is smoother shaped. This is because its members have one less freely rotating unit than their saturated counterparts and thus fewer conformational permutations are available to them. Whilst the number of available torsional units has an effect on the shape of these linear, polymeric ligands, the presence and position of the double bond has a greater impact on ligand dynamics. The smoothing effect is greatest and close to Gaussian within the 3Z-olefin panel. This makes intuitive sense, as the polymer is effectively subdivided into two shorter segments that flank the central restriction. Both segments explore a volume of space closer to the ligands COM and exhibit less conformational complexity as they possess fewer rotational units than the entire molecule. In contrast, saturated n-alkanols possess rotatable central bonds that allow the molecule to access a greater number of internal configurations and permutations.

Examination of bound maximum, mean and minimum Rg^2 representative structures show that whilst ligands from all three panels are capable of adopting fully extended conformations, the most highly populated conformations are roughly 70-80% of the length of the fully extended ligand (**Fig.3.9** & **Table.3.3-4**). In the case of larger n-alkanol and terminal olefin ligands, the population distributions of these partially extended conformations tend to be slightly more favoured in the bound state at the expense of populations representing compact and fully extended conformations. However, the characteristic "right-shift" exhibited by bound 3Z-olefins indicate a preference for extended conformations (**Fig.3.8**). This can be rationalised by taking into account the architecture of the pocket and will be explored further in §3.3.5-6.

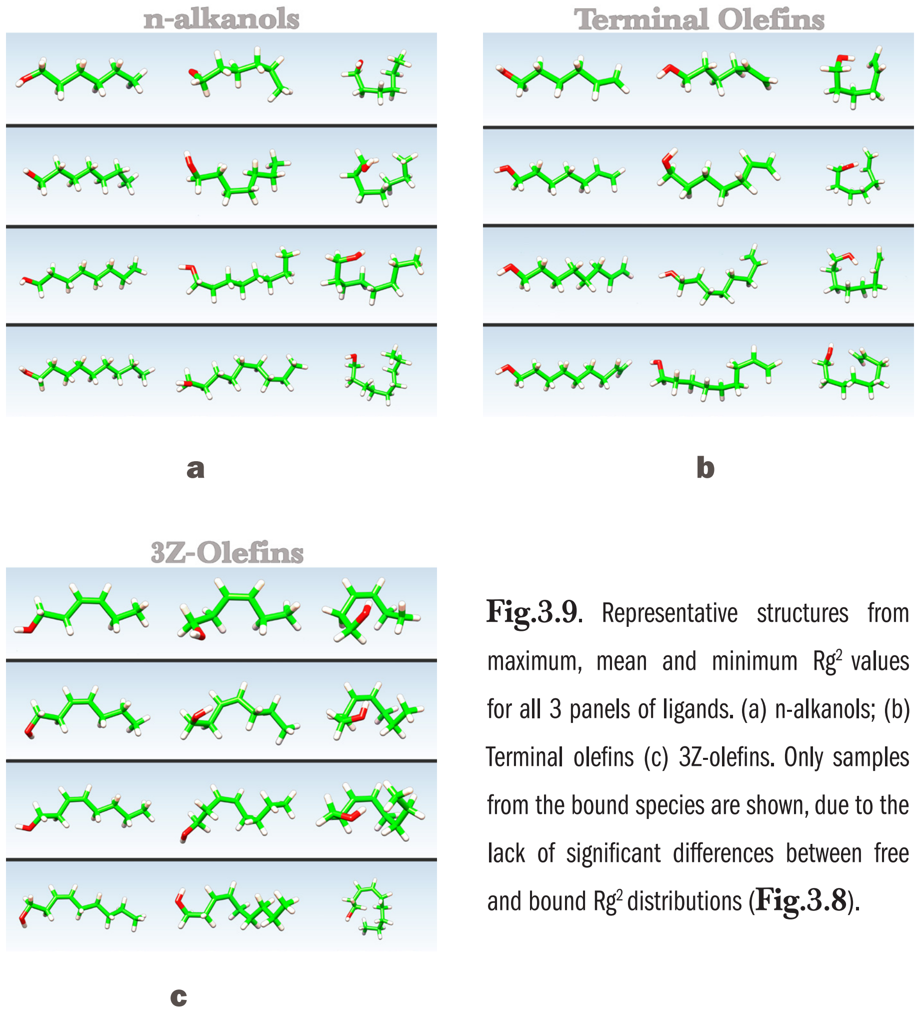


Fig.3.9. Representative structures from maximum, mean and minimum Rg^2 values for all 3 panels of ligands. (a) n-alkanols; (b) Terminal olefins (c) 3Z-olefins. Only samples from the bound species are shown, due to the lack of significant differences between free and bound Rg^2 distributions (**Fig.3.8**).

3.3.3. Selection of appropriate bin widths for calculation of TS_{Tor}

The (frame) spacing of data points obtained from a finite dataset is less important in the histogram method (**eqn.3.8**) than the QHA, as this 1st order method is not *a priori* defined to account for correlations from other DOF. However, the QHA includes linear pairwise correlations which take more time to converge and this results in greater sensitivity to data point density [189]. However, the entropy values generated by histogramming methods are particularly sensitive to the number of bins used to generate the distribution. If the data range is apportioned within too few bins, the data is over-smoothed and the entropic calculation renders a higher value due to loss of data structure. Conversely, too many bins results in greater sharpening of the data. The histogram appears jagged and spiky, and the entropy reported is smaller due to artefacts in the data (**Fig.3.10**). Thus it follows that a finite dataset of limited density (< 2×10^5 to 3×10^5) points will be more sensitive to selection of the number of bins than a larger dataset, and greater care must be taken in selecting the appropriate number of bins [187,204]. As the datasets examined in this chapter possess a higher density of 6×10^6 points, the risk of over-binning is marginal.

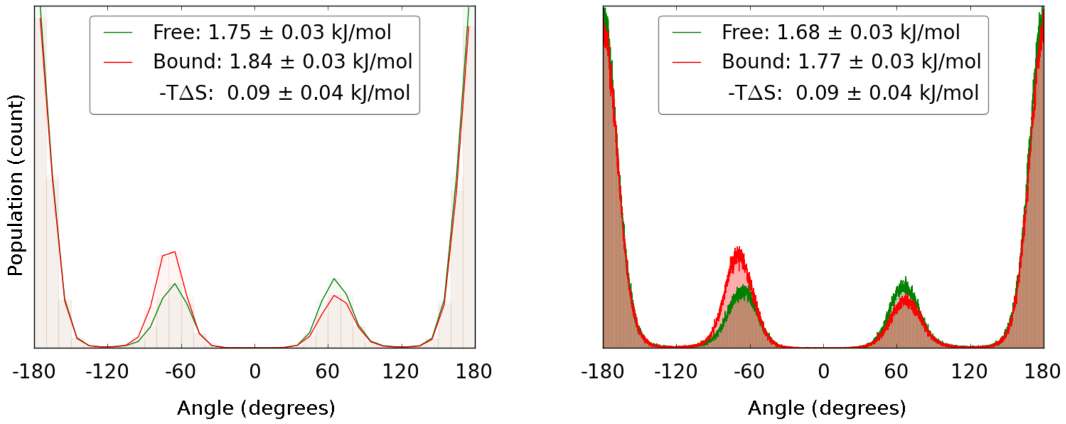


Fig.3.10. Dihedral distributions for the bond between two methylene groups. The graph on the left utilises 36 bins and is over smoothed, whilst the right depicts the jaggedness caused by excessive discretisation of the data when using 3,600 bins.

There are several methods (such as the Sturges and Scott formulae) for deducing optimal bin-widths for histograms, but analysis suggests that they are not best suited for torsional angles due to the data frequently adopting multi-modal distributions [182,204]. Hence, the appropriate bin size for the ligand dihedrals were assessed by generating and plotting entropic values for a variety of bin sizes (**Fig.3.11**).

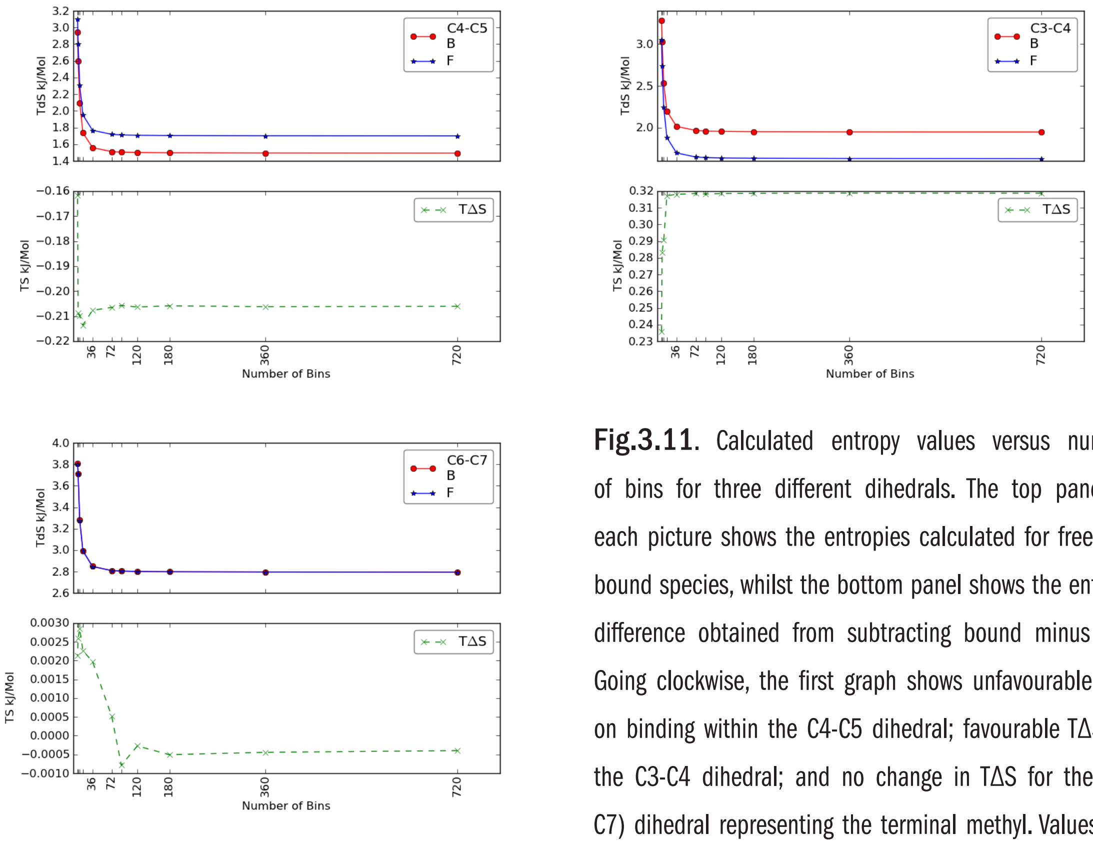


Fig.3.11. Calculated entropy values versus number of bins for three different dihedrals. The top panel in each picture shows the entropies calculated for free and bound species, whilst the bottom panel shows the entropy difference obtained from subtracting bound minus free. Going clockwise, the first graph shows unfavourable TΔS on binding within the C4-C5 dihedral; favourable TΔS for the C3-C4 dihedral; and no change in TΔS for the (C6-C7) dihedral representing the terminal methyl. Values are tabulated in **Table.3.5**.

	Conformational entropies of selected dihedrals (kJ/mol)										
	Dihedral C3-C4				Dihedral C4-C5				Dihedral Methyl		
Bins	**Bound**	**Free**	**Diff**		**Bound**	**Free**	**Diff**		**Bound**	**Free**	**Diff**
8	3.2876	3.0517	0.2359		2.9432	3.1049	-0.1617		3.8053	3.8031	0.0021
9	3.0269	2.7436	0.2832		2.5993	2.8079	-0.2086		3.7136	3.7110	0.0026
12	2.5330	2.2425	0.2905		2.0971	2.3069	-0.2099		3.2797	3.2769	0.0028
18	2.1973	1.8801	0.3172		1.7382	1.9517	-0.2135		2.9937	2.9914	0.0023
36	2.0146	1.6966	0.3180		1.5601	1.7678	-0.2077		2.8483	2.8463	0.0020
72	1.9660	1.6475	0.3185		1.5127	1.7192	-0.2065		2.8090	2.8084	0.0005
90	1.9590	1.6408	0.3182		1.5067	1.7124	-0.2057		2.8041	2.8048	-0.0008
120	1.9547	1.6362	0.3185		1.5023	1.7086	-0.2063		2.8004	2.8006	-0.0003
180	1.9513	1.6327	0.3186		1.4986	1.7045	-0.2059		2.7976	2.7981	-0.0005
360	1.9489	1.6301	0.3188		1.4963	1.7025	-0.2062		2.7957	2.7961	-0.0004
720	1.9477	1.6290	0.3187		1.4951	1.7012	-0.2061		2.7945	2.7949	-0.0004

Table.3.5. Tables of calculated entropies versus bin widths for the 3 dihedrals depicted in **Fig.3.11**. Bound & Free show relatively large differences, difference does not. Numbers are shown to 4 decimal places to demonstrate the limited oscillation of values.

After around 72-120 bins, the entropic difference between bound and free states remains approximately constant and is largely invariant to any further increase in the number of bins. Therefore, in this case we observe that the histogram method is generally more sensitive to under-binning and no over-binning was observed.

Given N energy wells of equal stability, the theoretical entropy can be calculated using the formula: -R ln(N) [173]. Thus, the correct number of bins can be further calibrated against the torsional entropy value obtained for the rotation of a terminal methyl in the free state. As this dihedral possesses three equally stable energy wells, the above formula yields a value of 2.74 kJ/mol. A range of bin widths were applied to data obtained from the terminal methyl group of ligands from the primary alcohol panel and it was observed that a greater number of bins lower the entropy value so that it approaches the expected theoretical value of 2.74 kJ/mol. However, there was little improvement after ~72-90 bins. It should be noted that this calibration is only relevant for the specific dataset size tested. Taking into account the additional computational cost, stability of the entropy basin, smoothness of the data fit and values used in the literature, 360 bins were selected, aesthetically corresponding to a degree per bin. As rigid double bonds possess deep energetic wells and have no discernible torsional rotation, their vibrational entropies were treated as having zero contribution. In the case of ligands binding to MUP, there is no significant difference between the bound and free PDFs generated for the double bonds. Therefore, the entropic difference would always equate to zero, regardless of the method used.

3.3.4. A brief aside on torsional entropy estimates from the literature

There is disagreement in the literature as to the true cost of freezing an internal torsional rotation. Page and Jencks (1971) estimate the value at 2.7 to 4.3 e.u (3.4 to 5.4 kJ/mol at 300 K) per internal rotation for the cyclisation of saturated hydrocarbons of various lengths [171]. Searle and Williams (1992) reported average values of 1.6 and 3.4 kJ/mol (at 300 K) for the restriction of a single rotor. The measurements were derived from entropies of fusion obtained from melting hydrocarbon crystals and the two distinct values represent odd and even numbered linear hydrocarbons respectively. The disparity in values was due to the tendency of the latter to exhibit phase transitions below the melting temperature, whilst the former did not. They made the point that even in a crystalline structure, residual torsional vibrations result in internal bonds being restricted rather than being frozen. Thus, the entropy values obtained from the isomerization reactions of Page and Jencks were potentially an overestimate, as the cyclisation process converts flexible chains into highly restrained products that are unrepresentative of non-covalent, protein-ligand binding interactions. Biological complex disassociation is postulated to be closer in analogy to weakly orientated hydrocarbon crystals, though even this model falls short of an ideal description [170].

Mammen and Whitesides (1998) also obtain a lower value (of approximately 2.0 kJ/mol) for all central torsions in linear alkanes via a computational approach [196]. However, their methodology and assumptions were criticised by Ercolani (1999) as their technique was classically based and did not suitably account for the transformation of the multimodal potential energy surface into a single high frequency torsional well upon total restriction of the bond. Additionally, Ercolani argues that whilst the model could hold true for partially hindered restrictions as are typically seen in drug design, molecular recognition, and self-assembly, calculating the entropy loss on complete restriction of an internal rotation would also have to account for the kinetic component and the requisite quantum corrections [235].

Even higher estimates of around 8.3 kJ/mol per restricted bond have been proposed by Carver (1993) for the binding of oligosaccharides to proteins, whilst more modern computational studies have predicted 1.7 to 4.2 kJ/mol per torsion [112,167,169]. Gilson makes the comment that the greatest losses in conformational entropy are due to narrowing of the energy wells versus loss of stable rotamers and the true value for the latter should be lower than 1.7 kJ/mol [112,166,173]. Considering the evidence, 5-6 kJ/mol is likely an upper bound corresponding to the total freezing of a rotamer. It is unlikely that the majority of torsions in typical binding reactions would be constrained to this extent, as tight binding compounds still exhibit significant residual torsional vibrations after binding. Thus, it is of interest to assess the upper limit of entropy differences calculated by the histogram method. Consequently, tight ligand binding was investigated by studying simulations of Darunavir complexed to wild type HIV-1 protease (HIV PR). As Rg^2 data indicate that ligands bound to MUP do not suffer much conformational restriction, this contrasting system is relevant to the results presented in the following sections.

HIV PR is a symmetric homodimer that can be thought of as a pair of molecular scissors [236]. It cleaves the precursor viral polyproteins Gag and Gag-Pro-Pol to release proteins essential for virus maturation and replication. As benefiting a well-studied drug target, a comparatively large number of competitive active site inhibitors have been developed to abolish its activity. Darunavir mimics the tetrahedral transition state intermediate of the protease's natural substrate and is a sulfonamide based inhibitor. Other structural analogues include PIs such as Amprenavir, TMC-126 and GS-8374. It was developed to overcome resistance caused by HIV-I protease mutants via the engineering of a fused, bicyclic, bis-tetrahydrofuran (bis-THF) moiety onto a ligand scaffold. This created additional hydrogen bond interactions with key residues along the protein backbone that are theoretically less susceptible to mutation [7,8,237,238]. The design successfully optimised a enthalpic signature that earlier first generation PIs lacked, and was assayed with the following thermodynamic values: ΔG -62.8 kJ/mol, ΔH -53.1 kJ/mol and TΔS 9.6 kJ/mol [58,239]. This ligand can make multiple directional hydrogen bonds that ensure it is tightly bound and can be conceptually subdivided into four functional groups which occupy the different subsites that constitute the anatomy of the active site. When free in solution, key dihedrals within the inhibitor backbone separate the bulky functional groups and possess relatively large ranges of rotation. On binding, these dihedrals are expected to become considerably restrained.

A 300 ns aggregate MD simulation of Darunavir bound to HIV PR shows that dihedral angles suffer a range of restriction: from the freely rotating methyl groups and rigid aromatic bonds that show no difference in $T\Delta S_{Tor}$, to large torsional entropy reductions resulting from the narrowing or loss of features within the multimodal torsional landscape. Dramatic tightening of torsional wells is not frequently observed and the largest $T\Delta S_{Tor}$ penalty of 3.1 kJ/mol occurs within dihedral-25 from the selected subset shown in **Fig.3.12.** In this instance, the bound PDF has narrowed substantially and the reported 1st order entropy value is likely to be underestimated by the classical treatment, requiring additional quantum and kinetic corrections as asserted by Ercolani. However, most $T\Delta S_{Tor}$ differences (31 out of 38 rotamers) are well described by the method used in this thesis. For reference, the largest entropy loss obtained in our experiments with this method was 4.0 kJ/mol for the binding of Amprenavir (dihedral-24) to HIV PR (Data not shown).

When Darunavir binds to HIV PR, dihedrals located near the sulfonamide moiety (e.g. dihedral-25 and 30) are highly restricted due to strong protein-ligand interactions. However, inhibitors are typically flexible entities and confinement can be compensated with motion elsewhere, be it large or small distributed increases within multiple ligand

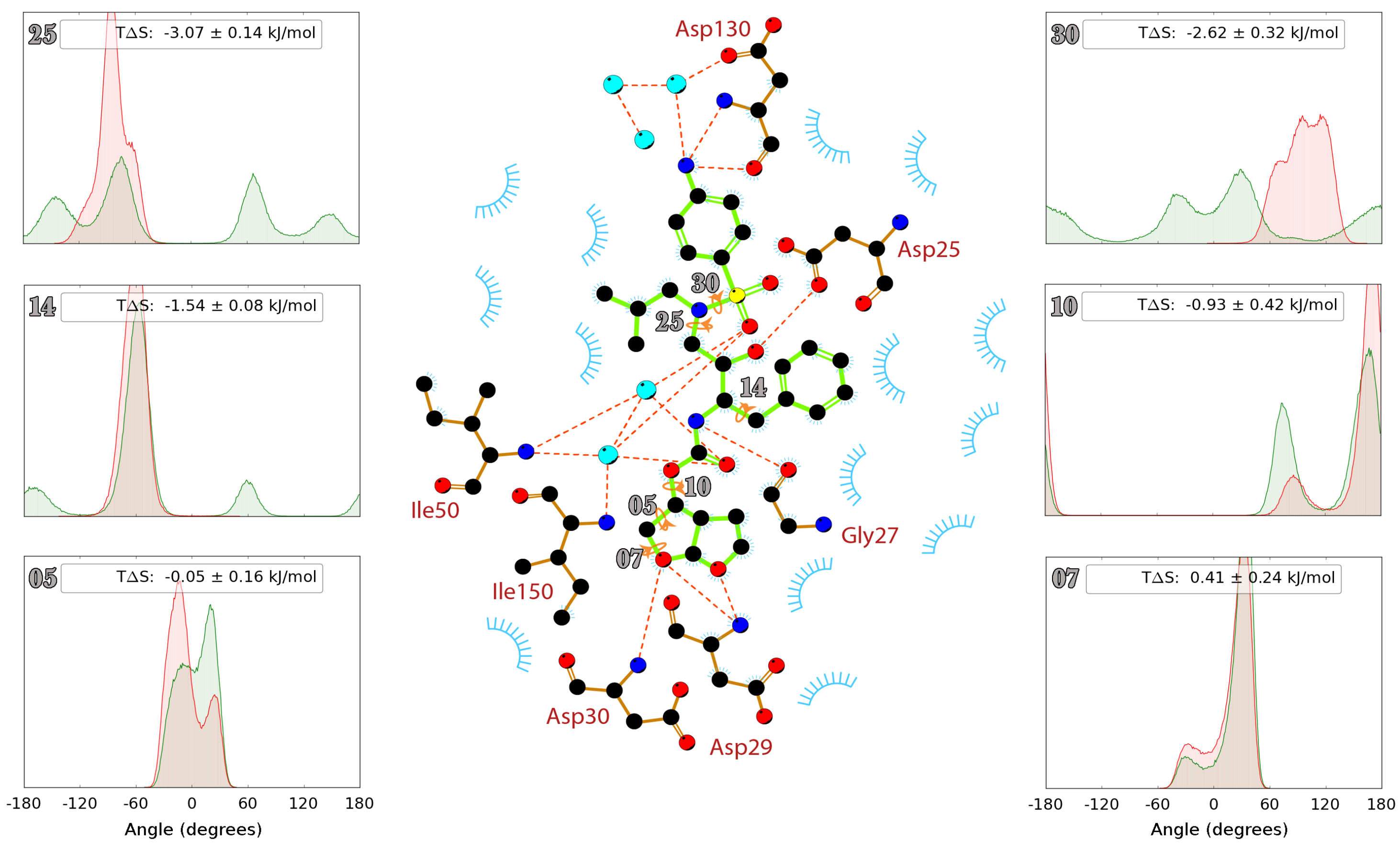


Fig.3.12 Key dihedral distributions for the binding of Darunavir to HIV PR obtained from MD simulations. The ligand backbone is coloured green and protein-ligand hydrophobic interactions are depicted as curved rays. Red dashed lines are hydrogen bonds, whilst cyan spheres represent water. Probability distributions are numbered according to dihedral position. Bound distributions are coloured red and free green.

DOF. For example, the summed torsional entropy of the bis-THF group is actually favourable by 1 kJ/mol. The same group is unfavourable by around 4 kJ/mol in the PI, GS-8374. On binding Darunavir, some dihedral distributions within the constrained double ring undergo small shifts that results in a net $T\Delta S_{Tor}$ close to zero (e.g. dihedral-05), whilst other dihedrals gain in entropy (e.g. dihedral-07). Other dihedrals (e.g. dihedral-14 and 10) within the ligand backbone already have highly occupied torsional wells, and on binding occupancy within these wells only increases moderately and the resultant entropy loss is not as considerable as expected (-1.54 and -0.93 kJ/mol). Thus, it can be seen that entropy losses cannot be easily simplified to simple fixed values on a per-rotor basis without extensive calibration. Even in this case, the rotamer counting method (as used in reference [240]) would be an approximation that yields dubious results, as any calibration is likely to be system dependent due to the subtle dispersion of entropy losses throughout a flexible ligand. Even the ultra-tight binding biotin-avidin complex has been observed to possess some residual ligand motion in MD simulations. The flexible tail rapidly switches between a major and minor conformation in both solution and lattice phase simulations at room temperature. The authors suggested that the minor conformation was not visible in the experimental crystal structure due to cryogenic temperatures disfavouring occupancy of this alternate state [241]. Hence, estimates of 5 to 6 kJ/mol per rotor are most probably too high for the majority of non-covalent interactions due to residual ligand/protein dynamics, and $T\Delta S_{Tor}$ losses of 1.7 kJ/mol and lower were typically observed for the dihedrals tested in HIV PR.

3.3.5. Per-dihedral torsional entropies of ligands that bind to MUP

The torsional entropy of every dihedral, apart from those involving double bonds, were calculated using **eqn.3.8** for bound and free species. 1st order entropy differences can be compared on a per-dihedral basis without worrying about overestimating the entropy due to correlated motions as there is no summation involved. Though the raw dihedral distributions are informative, the amount of information generated for all 12 ligands is overwhelming and the trends are subtle. Therefore, images for all raw dihedral distributions are located in the appendix (**Fig.A1.1-3**).

Generally, there are no large differences between bound and free PDFs for most dihedrals. Typically dihedrals exhibit very small decreases or increases (< 1 kJ/mol) in $T\Delta S_{Tor}$ on binding, whilst others possess no change. This is unsurprising, as the Rg^2 distributions indicated that rotors are slightly hindered rather than being fully restricted. Thus, $T\Delta S_{Tor}$ values are small compared to some of the values obtained for the binding of Darunavir to HIV PR. Extensive tumbling of the IBMP ligand was observed whilst bound to MUP in microsecond long MD simulations performed by Roy et al. (2010) [91]. The authors predicted that conformational entropy losses would be minimal to non-existent and rotational losses modest [91]. Malham (2012) validated the occurrence of significant ligand residual motion by running simulations with the AMBER and CHARMM force fields [180]. **Table.3.6** and **Fig.3.13** depict the per-dihedral entropies for all ligands binding to MUP and the trends observed for the three panels are discussed below.

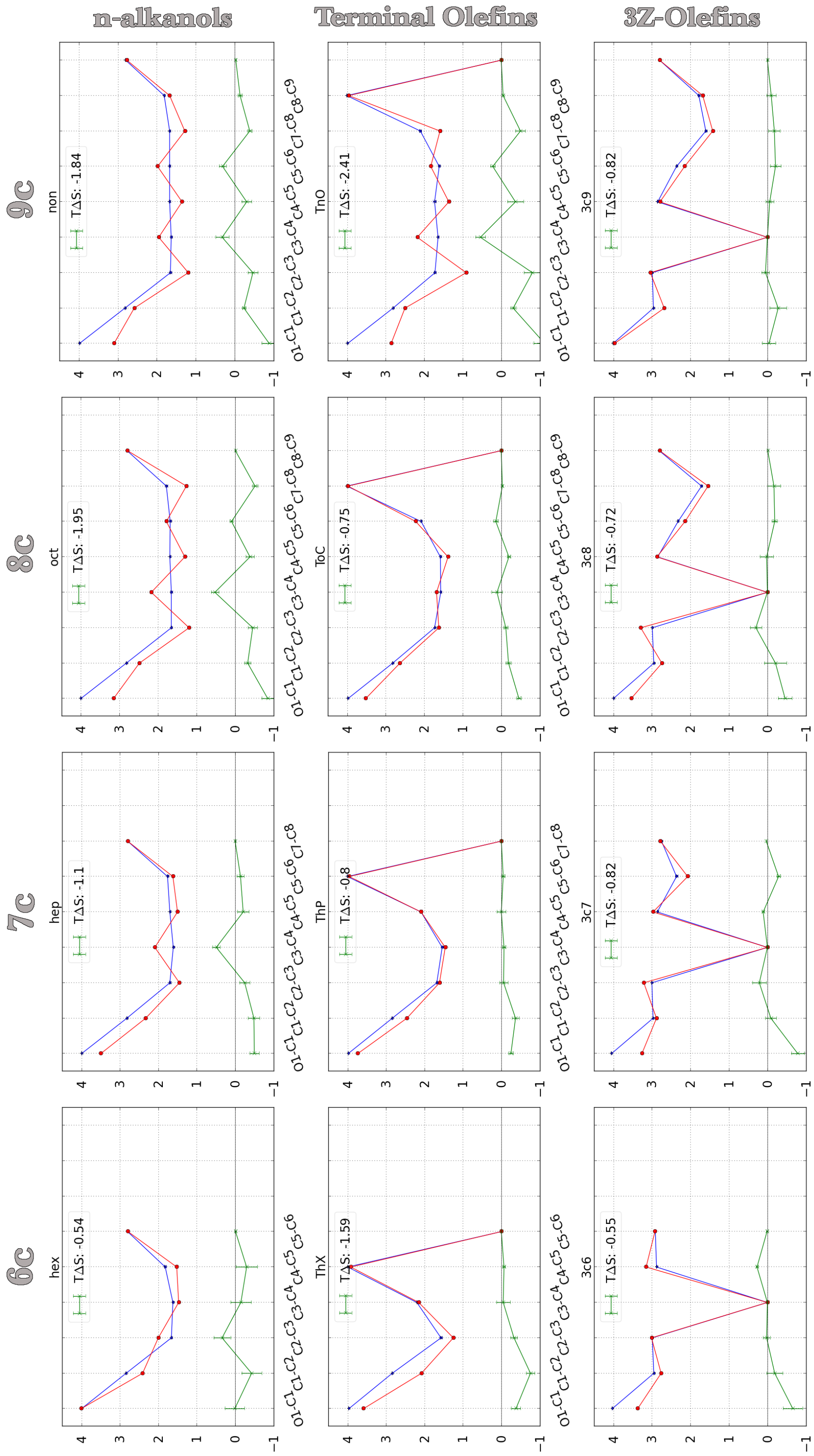


Fig 3.13. Per-dihedral breakdown of torsional entropies for free (blue), bound (red) and "bound minus free" (green). Each data point is calculated from an aggregate 600 ns simulation and the standard error is calculated from entropy values generated for 6 individual 100 ns simulations.

n-alkanols: The largest TS_{Tor} values for the free ligand are obtained for the terminal ends of the molecule, whilst values for dihedrals near the centre of the molecule form a basin that increases in size on adding methylene subunits. TS_{Tor} values are typically around 1.65 to 1.70 kJ/mol and correspond to the lower end of torsional estimates reported in the literature (§3.3.4). NMR spectroscopy studies observed this phenomena as a progressive decrease in correlation times progressing from the centres of n-chain alkanes towards their termini [242,243]. This is because the slowest dynamics take place near the centre of mass of the molecule, as methylene units located at this point have to satisfy the conformational requirements of two neighbours. On the other hand, a characteristic "zigzag" pattern develops in the bound state as the length of the ligand increases, and the O1-C1 dihedral possesses a smaller value of TS_{Tor} than its free counterpart. This is a result of reduced conformational interconversion and increased occupation of the trans state. When coupled with the fact that longer ligands are more likely to abut the boundaries of the calyx, it is likely that the increase in $T\Delta S_{Tor}$ within some central dihedrals are a result of the ligand having less volume to explore within the pocket.

Terminal olefins: Per-dihedral entropies of the free species show a similar, albeit shorter "basin", with a more pronounced central dip than that of their n-alkanol counterparts. The dihedral immediately preceding the terminal restriction has a TS_{Tor} of ~4.0 kJ/mol and is comparable to that of O1-C1. This adds a more pronounced dip in the basin. The effect of freezing the terminal rotor is not akin to removing a rotor as consistent $T\Delta S_{Conf}$ benefits of the expected magnitude (4.5 kJ/mol) are not observed (**Fig.3.13**). The per-dihedral entropies indicate that, subtler effects dependent on ligand length are at play. Whilst bound, 9C TnO has a similar per-dihedral pattern to oct, but it pays the greatest $T\Delta S_{Conf}$ penalty (-2.41 kJ/mol) on binding out of all the simulated ligands and the terminal restriction is a handicap. The shortest ligand, ThX, also does not benefit from the restriction and has an unfavourable $T\Delta S_{Conf}$ (-1.59 kJ/mol) of binding. However, ligands of intermediate size (ThP and ToC) suffer minimal entropy losses and the reasons for this will be assessed in the next section.

3Z-olefins: In the free state, dihedrals immediately flanking the cis-3-4 restriction occupy gauche(+) and gauche(-) conformations and rotate to minimise steric clashes between C2 and C5 methylenes. They have higher TS_{Tor} (~2.8 to 3.0 kJ/mol) than equivalent n-alkanol dihedrals. Other central torsions characterised by three energy minima also have larger TS_{Tor} as a result of reduced torsional barriers. Only dihedrals adjacent to the terminal methyl in 8C and 9C ligands have comparable TS_{Tor} values (1.65 to 1.70 kJ/mol) to their n-alkanol counterparts. This demonstrates that the central restriction reduces the torsional energy barriers in dihedrals near the centre of the molecule and it only increases again after the hydrocarbon chain exceeds a critical length in 3c8. Despite this, this panel performs the best in terms of TΔS$_{Conf}$, with the greatest losses generally occurring near the hydrogen bonded, hydroxyl head of the ligand. Thus, the double bond is well positioned to effectively mitigate torsional entropy losses on binding for all tested ligands.

hex					ThX					3c6			
Dihedral	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**
O1-C1	4.00 ± 0.00	4.01 ± 0.25	0.01 ± 0.25		**O1-C1**	3.97 ± 0.00	3.59 ± 0.12	-0.38 ± 0.12		**O1-C1**	4.03 ± 0.00	3.37 ± 0.25	-0.66 ± 0.25
C1-C2	2.84 ± 0.01	2.41 ± 0.26	-0.43 ± 0.26		**C1-C2**	2.84 ± 0.00	2.08 ± 0.11	-0.76 ± 0.11		**C1-C2**	2.95 ± 0.01	2.76 ± 0.21	-0.19 ± 0.21
C2-C3	1.66 ± 0.02	2.00 ± 0.22	0.34 ± 0.22		**C2-C3**	1.57 ± 0.03	1.25 ± 0.09	-0.33 ± 0.09		**C2-C3**	2.99 ± 0.00	3.01 ± 0.09	0.02 ± 0.09
C3-C4	1.62 ± 0.02	1.47 ± 0.27	-0.15 ± 0.27		**C3-C4**	2.20 ± 0.01	2.15 ± 0.18	-0.05 ± 0.18		**C3-C4**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C4-C5	1.82 ± 0.03	1.52 ± 0.27	-0.30 ± 0.28		**C4-C5**	4.00 ± 0.00	3.92 ± 0.04	-0.07 ± 0.04		**C4-C5**	2.88 ± 0.00	3.15 ± 0.03	0.27 ± 0.03
C5-C6	2.80 ± 0.00	2.79 ± 0.02	-0.01 ± 0.02		**C5-C6**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00		**C5-C6**	2.91 ± 0.00	2.92 ± 0.02	0.01 ± 0.02
hep					**ThP**					**3c7**			
Dihedral	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**
O1-C1	4.00 ± 0.00	3.50 ± 0.12	-0.50 ± 0.12		**O1-C1**	3.99 ± 0.00	3.74 ± 0.06	-0.25 ± 0.06		**O1-C1**	4.05 ± 0.00	3.26 ± 0.17	-0.79 ± 0.17
C1-C2	2.82 ± 0.01	2.33 ± 0.15	-0.49 ± 0.15		**C1-C2**	2.84 ± 0.00	2.46 ± 0.09	-0.37 ± 0.09		**C1-C2**	2.97 ± 0.00	2.88 ± 0.14	-0.09 ± 0.14
C2-C3	1.70 ± 0.02	1.45 ± 0.13	-0.25 ± 0.13		**C2-C3**	1.68 ± 0.02	1.61 ± 0.11	-0.06 ± 0.11		**C2-C3**	3.00 ± 0.00	3.21 ± 0.18	0.21 ± 0.18
C3-C4	1.61 ± 0.02	2.09 ± 0.10	0.49 ± 0.10		**C3-C4**	1.54 ± 0.03	1.46 ± 0.04	-0.07 ± 0.05		**C3-C4**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C4-C5	1.70 ± 0.01	1.50 ± 0.15	-0.21 ± 0.15		**C4-C5**	2.09 ± 0.03	2.09 ± 0.12	-0.00 ± 0.12		**C4-C5**	2.86 ± 0.00	2.97 ± 0.03	0.11 ± 0.03
C5-C6	1.76 ± 0.02	1.62 ± 0.09	-0.14 ± 0.09		**C5-C6**	4.01 ± 0.00	3.96 ± 0.04	-0.05 ± 0.04		**C5-C6**	2.36 ± 0.03	2.07 ± 0.04	-0.29 ± 0.05
C6-C7	2.80 ± 0.00	2.80 ± 0.01	-0.00 ± 0.01		**C6-C7**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00		**C6-C7**	2.75 ± 0.00	2.78 ± 0.00	0.03 ± 0.00
oct					**ToC**					**3c8**			
Dihedral	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**
O1-C1	4.00 ± 0.00	3.15 ± 0.16	-0.85 ± 0.16		**O1-C1**	3.98 ± 0.00	3.52 ± 0.06	-0.46 ± 0.06		**O1-C1**	4.00 ± 0.00	3.54 ± 0.18	-0.46 ± 0.18
C1-C2	2.82 ± 0.01	2.48 ± 0.08	-0.33 ± 0.08		**C1-C2**	2.82 ± 0.00	2.63 ± 0.07	-0.19 ± 0.07		**C1-C2**	2.95 ± 0.00	2.74 ± 0.29	-0.21 ± 0.29
C2-C3	1.65 ± 0.01	1.19 ± 0.12	-0.46 ± 0.12		**C2-C3**	1.73 ± 0.01	1.62 ± 0.06	-0.12 ± 0.06		**C2-C3**	2.99 ± 0.00	3.29 ± 0.15	0.30 ± 0.15
C3-C4	1.65 ± 0.03	2.17 ± 0.10	0.52 ± 0.10		**C3-C4**	1.57 ± 0.02	1.68 ± 0.14	0.11 ± 0.14		**C3-C4**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C4-C5	1.69 ± 0.01	1.30 ± 0.11	-0.39 ± 0.11		**C4-C5**	1.58 ± 0.02	1.38 ± 0.04	-0.20 ± 0.05		**C4-C5**	2.85 ± 0.00	2.87 ± 0.17	0.02 ± 0.17
C5-C6	1.68 ± 0.03	1.78 ± 0.03	0.09 ± 0.04		**C5-C6**	2.08 ± 0.01	2.22 ± 0.06	0.14 ± 0.06		**C5-C6**	2.32 ± 0.01	2.14 ± 0.07	-0.19 ± 0.07
C6-C7	1.78 ± 0.03	1.26 ± 0.06	-0.52 ± 0.07		**C6-C7**	4.01 ± 0.00	3.98 ± 0.02	-0.03 ± 0.02		**C6-C7**	1.71 ± 0.02	1.54 ± 0.17	-0.17 ± 0.17
C7-C8	2.80 ± 0.00	2.79 ± 0.01	-0.01 ± 0.01		**C7-C8**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00		**C7-C8**	2.81 ± 0.00	2.80 ± 0.01	-0.01 ± 0.01
non					**TnO**					**3c9**			
Dihedral	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**		**Dihedral**	**Free**	**Bound**	**Difference**
O1-C1	3.99 ± 0.00	3.10 ± 0.19	-0.89 ± 0.19		**O1-C1**	3.99 ± 0.00	2.86 ± 0.29	-1.13 ± 0.29		**O1-C1**	4.01 ± 0.00	3.97 ± 0.17	-0.04 ± 0.17
C1-C2	2.82 ± 0.00	2.58 ± 0.05	-0.24 ± 0.05		**C1-C2**	2.81 ± 0.00	2.50 ± 0.07	-0.31 ± 0.07		**C1-C2**	2.96 ± 0.00	2.68 ± 0.22	-0.28 ± 0.22
C2-C3	1.65 ± 0.03	1.19 ± 0.12	-0.47 ± 0.13		**C2-C3**	1.72 ± 0.03	0.91 ± 0.22	-0.81 ± 0.22		**C2-C3**	2.99 ± 0.00	3.04 ± 0.10	0.05 ± 0.10
C3-C4	1.63 ± 0.02	1.95 ± 0.17	0.32 ± 0.17		**C3-C4**	1.64 ± 0.02	2.18 ± 0.13	0.54 ± 0.13		**C3-C4**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
C4-C5	1.67 ± 0.02	1.35 ± 0.13	-0.31 ± 0.13		**C4-C5**	1.73 ± 0.04	1.36 ± 0.21	-0.37 ± 0.21		**C4-C5**	2.85 ± 0.00	2.79 ± 0.10	-0.06 ± 0.10
C5-C6	1.67 ± 0.02	1.98 ± 0.09	0.31 ± 0.09		**C5-C6**	1.61 ± 0.02	1.83 ± 0.06	0.22 ± 0.06		**C5-C6**	2.36 ± 0.01	2.15 ± 0.14	-0.21 ± 0.14
C6-C7	1.67 ± 0.02	1.28 ± 0.06	-0.39 ± 0.06		**C6-C7**	2.10 ± 0.03	1.59 ± 0.12	-0.50 ± 0.13		**C6-C7**	1.60 ± 0.02	1.42 ± 0.15	-0.18 ± 0.15
C7-C8	1.81 ± 0.01	1.67 ± 0.06	-0.14 ± 0.06		**C7-C8**	4.01 ± 0.00	3.96 ± 0.02	-0.05 ± 0.02		**C7-C8**	1.78 ± 0.03	1.68 ± 0.11	-0.10 ± 0.12
C8-C9	2.80 ± 0.00	2.77 ± 0.00	-0.03 ± 0.01		**C8-C9**	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00		**C8-C9**	2.80 ± 0.00	2.80 ± 0.01	0.00 ± 0.01

Table.3.6. Calculated TS$_{Conf}$ and TΔS$_{Conf}$ for every dihedral in all 3 panels of ligands. Each data point is calculated from an aggregate 600 ns simulation and the standard error is calculated from entropy values generated from six individual 100 ns simulations.

3.3.6. The dynamics of bound and free ligands

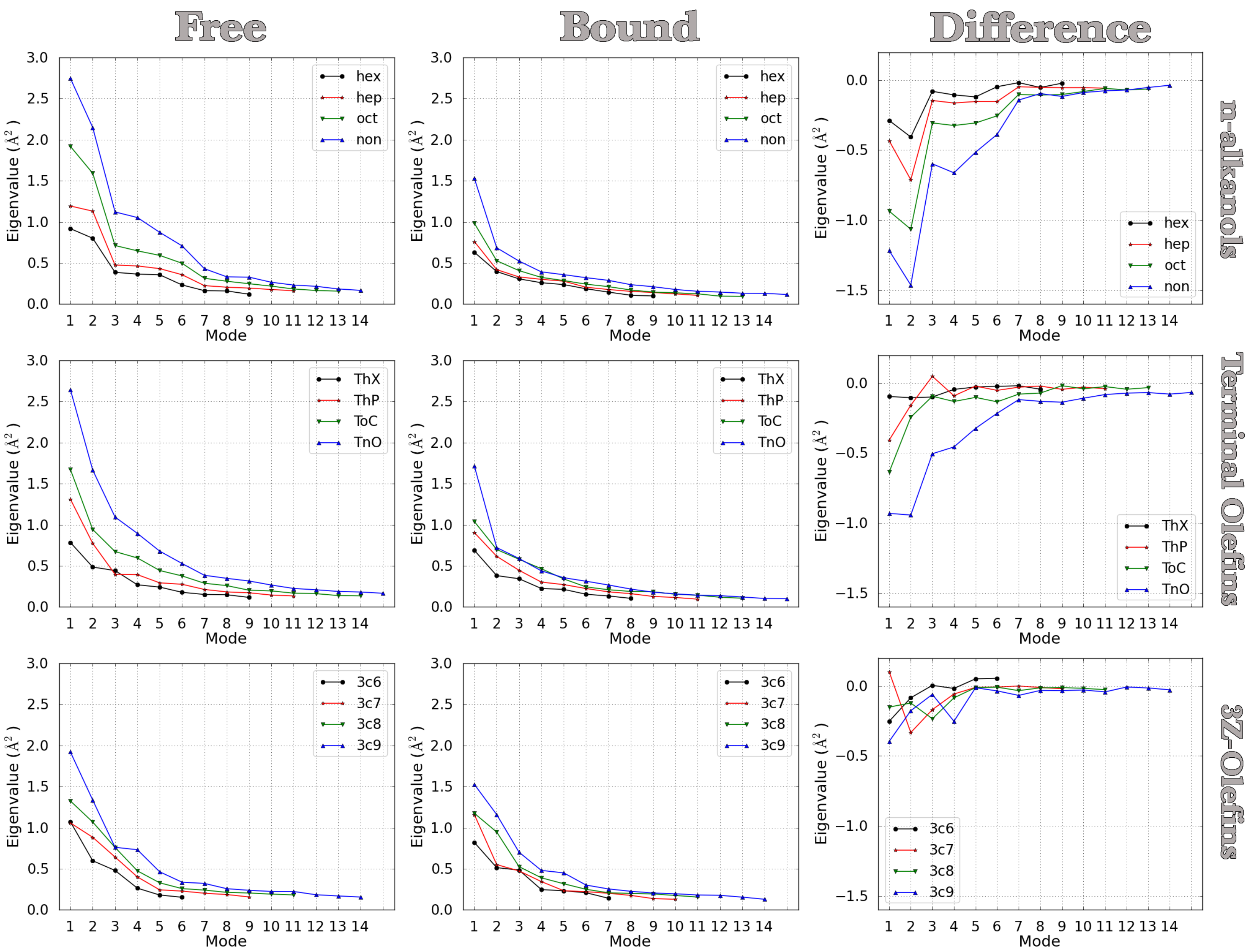
3.3.6.1. Principal component analysis

Eigenvalues from all 3N-6 principal component modes for bound and free ligands were plotted in **Fig.3.14** to visualise the amount of correlated motion present. The lowest PCMs have the greatest variance and describe the largest collective motions of the ligand. The magnitude of the eigenvalue is correlated to the length of the ligand and the first two modes represent the largest displacements. Modes higher than 7 have very small fluctuations and are negligible. On binding, correlated motions are reduced to broadly comparable levels between all three ligand panels. Additionally, the separation between principal component modes of the different ligands constituting a panel is diminished. This suggests that the constraints imposed by the pocket are constant and any differences in co-ordinated motions are purely a consequence of ligand structure. The impact on panels can be ranked in descending order: n-alkanols > terminal olefins > 3Z-olefins. The difference plots were generated by subtracting "free" from "bound". They indicate that 3Z-olefins suffer little reduction in their co-ordinated motions. As they lack the flexible central bond possessed by the other two panels, the degree to which large-scale, correlated motions can take place is limited to the two shorter segments flanking the double bond. Hence, there is less to lose on binding. *Per contra*, n-alkanols can undergo a greater variety of flexing and bending motions as the ligand backbone is the most flexible. Placing a restriction at the terminus reduces the magnitude of correlated motions to a lesser degree. This is because the contiguity of torsional subunits is disrupted in the most minimal manner possible by the position of the double bond (§3.3.6.2).

3.3.6.2. The impact of correlated ligand motions

The first two PCMs for free non are characterised by large wagging motions of the terminal ends of the ligand, whilst the central portion of the molecule flexes in sympathy (**Fig.3.15**). In the bound state, the ability of the termini to move away from the longitudinal axis of the ligand is hampered by the confines of the calyx and modes greater than the first exhibit an increase in smaller amplitude movements of all atoms. This sometimes results in motions reminiscent of the Schatzki crankshaft mechanism [244–246]. In the reversible transition known as the glass-liquid transition, amorphous solids such as polyethylene transform on heating from a stiff, brittle state to a molten, rubbery material with greater viscosity. Different materials have specific glass transition temperatures associated with them and the molecules comprising the polymer may exhibit significant residual motion that can be modulated by modification of its structure. e.g. The addition of rigid aromatic rings can increase the stiffness of thermoplastics [244–247]. Crankshaft motions occur in flexible, linear molecules with 4 or more successive methylenes positioned between two co-linear terminal bonds. They are

Fig.3.14. Eigenvalues plotted against principal component modes for all ligands from their respective 600 ns trajectories.



named as such because the central CH_2 groups rotate in the manner of a crankshaft. The activation energy for this process is only 54 kJ/mol and the free volume required for activation is roughly four times that of a single methylene [245]. The mechanism involves intermediate methylenes possessing a TGTGT sequence, where T represents trans and G, gauche [244].

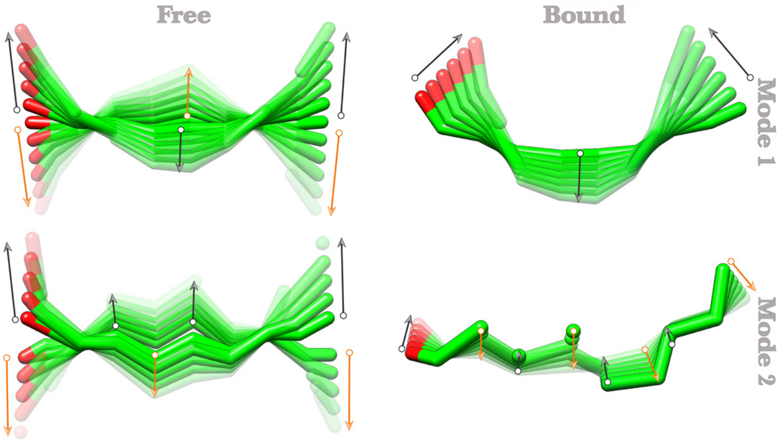


Fig.3.15. The first two modes generated from PCA analysis for nonan-1-ol heavy atoms, in free and bound states. Oxygen is coloured red and carbons green. Only half the motion is shown for clarity. Note that the second mode shows a greater reduction in correlated motion than the first (Also see **Fig.3.14**).

The type of behaviour described above occurs in the condensed phase, where the density of polymer molecules is ostensibly greater than the environment encountered in the MUP binding pocket. However, calyx residues can be said to analogously form a less concentrated amino acid matrix that ensconces the bound ligand. In the case of longer ligands (from hep to non), the shape of the binding site introduces a curvature near the central part of the ligand (**Fig.3.16** & **Fig.A1.4-5**). The large amplitude motions normally observed at the terminal ends of the free ligand are partially hindered within the bound species and excess motional energy (that would otherwise have been dissipated) is spent within the central portion of the molecule. The resulting fast motions are of smaller amplitude; concomitantly, a greater occupation of gauche states is first observed at the C3-C4 dihedral for hep, and then at the C5-C6 dihedral within oct and non. When comparing per-dihedral TS_{Tor} values, it is further observed that dihedrals adjacent to C3-C4 and C5-C6 have reduced entropies in the bound state compared to free (**Fig.3.13**). This is due to further increases in the predominantly occupied trans population, and is detectable in dihedrals C2-C3, C4-C5 and C6-C7 (**Fig.A1.1**). Thus, the TGTGT population distribution signature (in non) results in the “zigzag” pattern seen in the per-dihedral entropies. This in turn results in summed configurational entropy losses being ameliorated by the phenomenon of ligand entropy-entropy compensation.

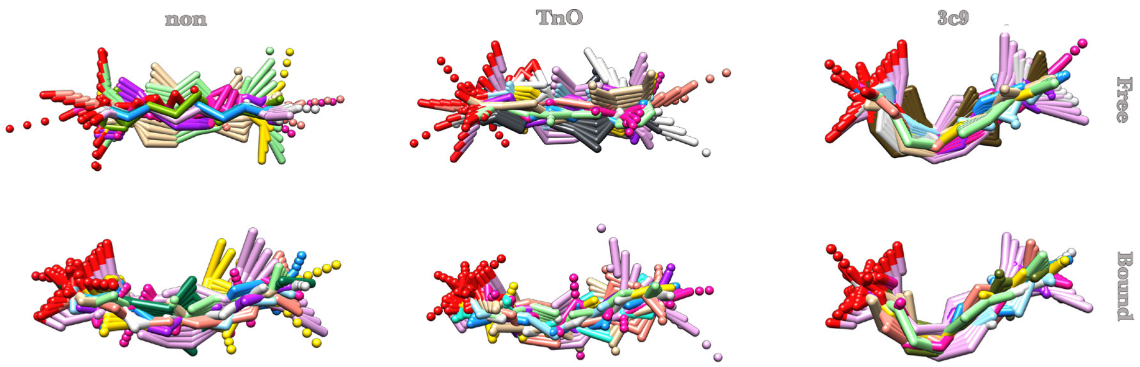


Fig.3.16. Comparison of all the modes obtained from PCA for free and bound states for nine carbon ligands. The top row and bottom rows depict the free and bound species respectively. Columns (left to right) show non, TnO and 3c9.

If structures representing the motion captured from all PCMs are overlaid for 9C ligands, it can once again be seen that the sweeping co-ordinated motions in the free state are replaced by smaller amplitude motions that are a result of fast, local interconversion of dihedrals between different energy minima (**Fig.3.16**). Whilst the permissible dihedral angle range in the bound species is not significantly altered, *specific permutations* of dihedrals allowing large scale motions are curtailed. Thus, the overall dihedral distributions for both states are very similar, whilst disparities in large co-ordinated dihedral motions are *masked* (**Fig.A1.1**). Without explicitly carrying out a 2nd order calculation, it is difficult to assess the impact of correlated motions on values of $T\Delta S_{Conf}$. Correlated motions reduce the magnitude of $T\Delta S_{Conf}$ but the degree of correlation is reduced within the bound state compared to the free. Thus, the free state might have less entropy due to greater amount of correlated motions but the bound state would not be affected to the same degree. Hence, the net $T\Delta S_{Conf}$ would be even smaller. The impact on the calculated entropies is expected to be smallest for 3Z-olefins and highest for n-alkanols, with terminal olefins lying in-between. Any applied 2nd order corrections would additionally be weighted by ligand length (**Fig.3.14**).

3.3.6.3. Ligand pre-organisation and the architecture of the calyx:

It is immediately noticeable that because of the cis-3-4 restriction, free 3c9 is pre-organised into conformations that are close to those that non would be forced to adopt whilst bound (**Fig.3.16**). Additionally, the shorter chain lengths of ligand sub-segments that flank the double bond contribute to smaller losses in $T\Delta S_{Conf}$ than that seen for n-alkanols. The C2-C3 and C4-C5 dihedrals adjacent to the double bond each occupy two gauche energy wells (**Fig.A1.2**). The non-rotatable restriction forms a locus around which the ligand can be conceptually subdivided into two smaller segments. Visual inspection of the trajectories indicates that ligands possess significant motion whilst bound, and an investigation into this phenomenon will form the focus of chapter 4.0. For the purposes of this discussion, it is sufficient to state that space within the calyx can be partitioned into three chambers. The top (*cal1*) and bottom (*cal3*) areas are slightly larger than the centre (*cal2*) chamber, which is bounded by TYR120, LEU105, PHE56 and ALA103. *Cal2* is marginally offset from the others due to the intrusion of TYR120 into the space from behind, PHE56 from the right and the gap created by the small ALA103 residue (beneath LEU105) on the left hand side of **Fig.3.17.** The staggered nature of the three chambers introduces a curvature into the ligand that is observed in many of its bound conformations. The hydrogen bond between TYR120, binding site waters and the ligand's hydroxyl head weakly fix the ligands position and orientation within the cavity. The lack of other directional bonding restraints allows this portion of the ligand to rise and fall and move in a roughly radial manner around TYR120 whilst the linear hydrocarbon tail follows suit.

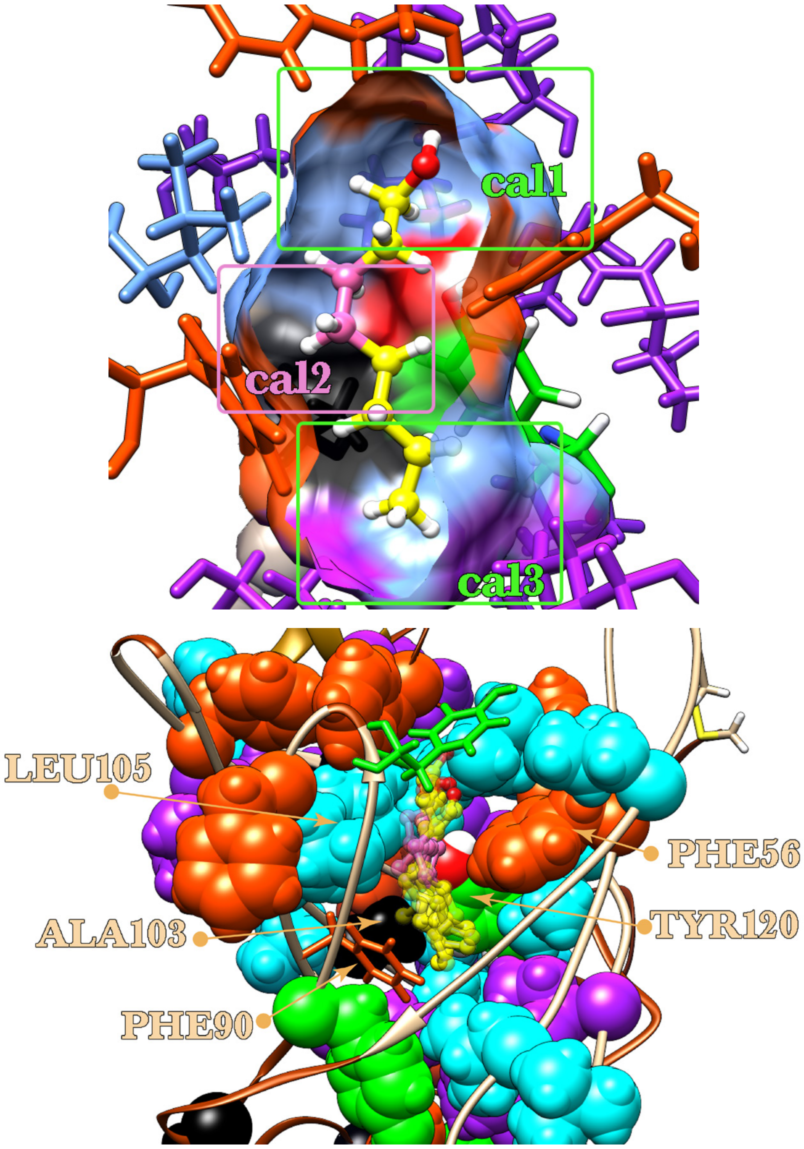


Fig.3.17. Architectural features of the MUP binding cavity and their interactions with bound ligand. The top image depicts octan-1-ol bound to MUP. The interior calyx surface is coloured by residue and has been generated from a static snapshot to show the available space in a moment in time. The C3-C4 ligand dihedral is coloured pink and the ligand yellow. Isoleucines - purple; leucines - cyan; phenylalanines - orange; tyrosines - green; and alanines black. Boxes demarcate the volume of the cavity into *cal1*, *cal2* and *cal3* (see text). The bottom image shows the same

view with the majority of residues depicted as space filling representations. A very small subset of ligand binding poses have been superimposed over one other to illustrate the range of motion these flexible, linear polymers possess whilst bound.

The C3-C4 dihedral is positioned near *cal2* and is generally adjacent to TYR120 most of the time (See Chapter 4.0). Whilst n-alkanol ligands exhibit an increase in conformational interconversion within the C3-C4 dihedral, this is impossible in 3Z-olefins and the central restriction partially decouples the transmission of internal motions from one terminus of the ligand to the other. Note that methylene groups directly flanking the restriction are still affected by steric clashes with each other. It remains to be seen whether the cis-3-4 double bond is favourably pre-organised to match the shape of the pocket and segments to either side of the restriction are better located to fit into *cal1* and *cal3*.

3.3.6.4. Conceptual subdivision of the ligand:

The division of the ligand into segments both prior to and after the C3-C4 dihedral is visualised in **Fig.3.18.** The “OH” segment closest to the hydroxyl terminus comprises summed, per-dihedral entropies for O1-C1, C1-C2, C2-C3 and C3-C4, whilst the rest of the per-dihedral entropies are added to the “CH_3” segment that lies closer to the terminal methyl group. Segmental “strain” is characterised by the redistribution of unbound dihedral torsional populations subsequent to binding and results in certain wells being preferentially occupied along with a concurrent reduction in conformational interconversion. Summing torsional entropies within a segment gives a measure of the conformational restriction suffered by that portion of the ligand. It is intuitively expected that the OH-segment would suffer more strain as the hydroxyl moiety is involved in a directional hydrogen bond compared to the CH_3 segment which is affected by weaker non-directional van der Waals interactions. Despite the growing size of the CH_3-segment upon increasing carbon chain length, ligands within the n-alkanol panel (**Fig.3.18.a**) still exhibit less $T\Delta S_{Conf}$ losses than that observed for the OH-segment because of entropy-entropy compensation. In contrast, 3Z-olefin OH-segment entropies become more favourable with increasing ligand length, whilst CH_3-segment entropies become more unfavourable (**Fig.3.18.c**). This meta-analysis depicts another level of compensatory behaviour that is a consequence of dihedrals composing the CH_3-segment increasingly occupying the trans conformation (**Fig.A1.3**). Thus, a small additive loss of conformational entropy is accumulated for every extra methylene unit that increases ligand length. Note that the unfavourable $T\Delta S_{Conf}$ observed for the OH-segment is also due to increased trans populations. Though the dihedrals adjacent to the double bond (C2-C3 and C4-C5) do exhibit a small asymmetric preference for gauche(+) or gauche(-) conformations, bound distributions broaden so as to increase occupation of states closer to trans. The subtle increase of trans populations in *all* dihedrals account for the "right-shifted" mean Rg^2 population distributions of bound 3Z-olefins, relative to free (**Fig.3.8**).

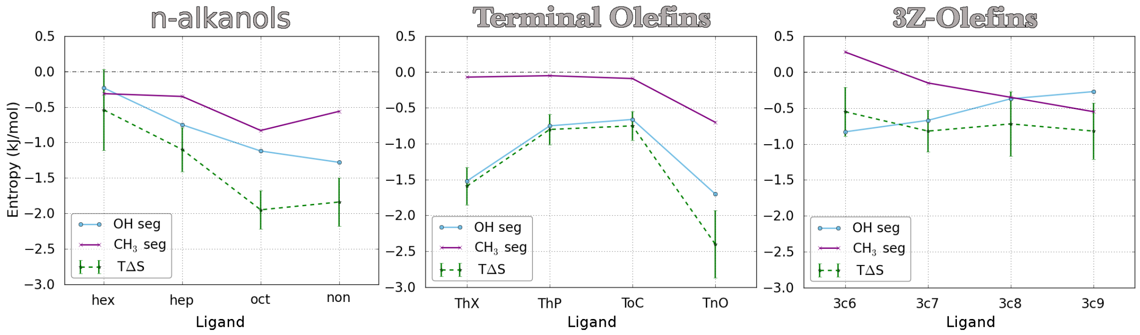


Fig.3.18. Entropy divided into segments before and after dihedral C3-C4. Graphs (left to right) depict n-alkanols, terminal olefins and 3Z-olefins respectively.

Generally, smaller ligands are better able to explore a greater variety of positions and orientations within the binding pocket. The cis-3-4 restriction results in the 3Z-olefins possessing reduced internal flexibility within their shorter sub-segments. In the case of 3c6 and 3c7 ligands, fewer torsional subunits aggravate this issue and the result is greater motional restriction at the hydrogen bonded OH-segment. As the ligand gets longer, it tends to adopt more defined locations in the binding pocket as a consequence of its size. Generally, the hydroxyl segment occupies *cal1*; the middle (around C3-C4) *cal2*; and the hydrocarbon tail *cal3*. Upon doing this, dihedrals flanking the central restriction sympathetically modulate their angles with respect to each other (to avoid steric clashes) and are driven by the dynamics of the adjoining sub-segments whose internal motions are largely decoupled from one another. Thus, $T\Delta S_{Conf}$ losses of dihedrals within these chains are minimal or even favourable (**Fig-3.13**).

CH_3 versus OH-segment compensatory behaviour within the 3Z-olefin panel is length dependent and $T\Delta S_{Conf}$ losses for the two segments perfectly balance for 3c8. Extending the length of the ligand by an additional methylene (i.e. 3c9) tips the balance and causes summed $T\Delta S_{Conf}$ values for the CH_3-segment to become increasingly unfavourable, whilst the OH-segment becomes more favourable (**Fig-3.18.c**). This supports the hypothesis that the root cause for improving OH-segment entropies in longer ligands is related to location within the pocket. In n-alkanols, the increase in C3-C4 and C5-C6 conformational transitions only relieves strain on the OH-segment in a minimal fashion. This is because the rapid fluctuations of these dihedrals often result in other ligand dihedrals reflexively transitioning between favourable and unfavourable conformations in response to the shifting geometry of the pocket. However, the rigid double bond in 3Z-olefins disallows the same internal dihedral interconversions and the central portion of the molecule is probably able to undergo greater rigid-body T&R displacements that better aid location of constituent segments into *cal1* and *cal3*. When ligands are extended by additional methylene units, the "balance" of entropic losses between sub-segments is augmented by greater locational stability (in the pocket), engendered by the increasing bulk of the CH_3-segment. In this manner, strain at the OH-segment is reduced. The last two paragraphs in this subsection form a hypothesis that will be further examined in chapter 4.0.

3.3.6.5. Terminally restricted olefin dynamics

This panel ostensibly test the impact of freezing a single rotatable bond in a more accurate manner than 3Z-olefins, as the latter modification affects torsions in both dihedrals flanking the cis-3-4 restriction. However, summed $T\Delta S_{Conf}$ values (**Table.3.7**) indicate that neither modification has a comparable magnitude to that suggested by Page and Jencks (1971) and entropic reductions are not linearly correlated with ligand length. The conformations adopted by ligands bound to MUP are similar to that when free. The confines of the pocket only hinder conformational transitions enough to introduce subtle dynamic effects that are biochemically interesting but relatively trivial in terms of global entropy contributions.

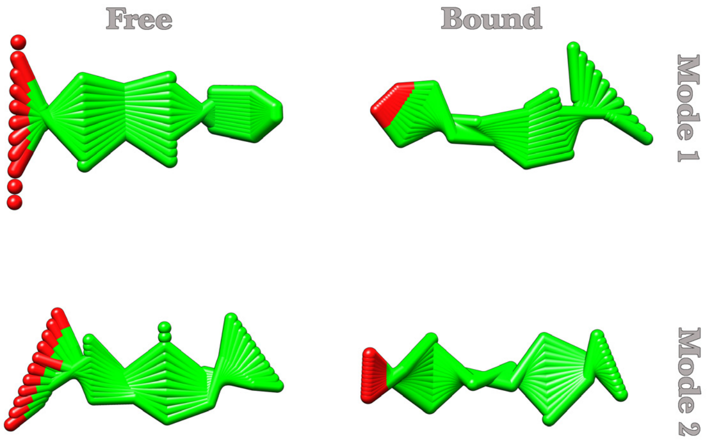


Fig.3.19. The first two modes generated from PCA analysis for hept-6-en-1-ol heavy atoms, in the free and bound state. Oxygen is coloured red and carbons green. Full range of motion is shown.

Large amplitude motions at the termini of free terminal olefins are more pronounced near the alcohol group than the opposite end, as the planar double bond disrupts the full range of correlated motion exhibited in these ligands compared to equivalent compounds in the n-alkanol panel (**Fig.3.19** & **Fig.A1.4-5**). Additionally, these polymers are slightly more predisposed to adopting "curved" conformations, as the double bond in conjunction with the preceding two carbons form a bulky sub-segment that frequently rotates transversal to the main body of the ligand. Even when fully extended, the restriction juts out from the plane formed by ligand's longitudinal axes, and thus its overall length is still reduced (**Table.3.3**). Hence, these factors ensure that the OH-segment loses the most $T\Delta S_{Conf}$, whilst the CH_3-segment loses minimal conformational entropy for all ligands apart from TnO (**Fig.3.18.b**). When the latter exceeds a critical length, an entropy-entropy compensation phenomenon similar to that in oct and non is observed (**Fig 3.13**). However, the dynamics of terminally restricted olefins are fundamentally different from their n-alkanol counterparts due to the loss of a rotational unit and the position of the restriction. Hence, $T\Delta S_{Conf}$ losses are accentuated instead of being attenuated in the case of its 9C member. The majority of entropy loss for ThX occurs at the hydroxyl terminus and is due to the ligand not having the same amount of flexibility and "balance" as ThP and ToC (**Fig.3.18.b**).

3.3.7. Torsional entropy differences: much ado about nothing

The torsional entropies for dihedrals within ligands tested were summed to yield total conformational entropies and plotted to display panel trends (**Table.3.7** & **Fig.3.20**). These are 1st order entropies and summing individual $T\Delta S_{Tor}$ in this manner gives a maximum entropy estimate as it neglects correlations which usually reduce the magnitude of the final sum. Despite the precautions taken in §3.3.3, only relative differences shown by the $T\Delta S_{Conf}$ values can be regarded as being accurate. However, bound and free TS_{Conf} values are displayed to better understand the source of entropic differences and to gauge any additivity on extending ligand length by a methylene. Calculated values only encompass the positional variations in the global entropy, neglecting the kinetic contribution along with bond stretching and angle bends. As the ligand is not tightly bound within the calyx, it is capable of adopting the majority of conformations observed in the free state, and thus these factors are likely to cancel.

Loss of ligand internal DOF were postulated to be the main cause for the linear entropic signature observed on binding n-alkanols to MUP. However, torsional entropy differences are small for all ligands, the largest being -2.41 kJ/mol for TnO. The differences between members within a panel are also small and do not approach the experimental $T\Delta\Delta S_i$ values of -5.4 kJ/mol per methylene [36,171]. The basis for the initial hypothesis was based on estimates from the cyclisation of saturated hydrocarbons by Page and Jencks (1971). As discussed in §3.3.4, this is most likely to be an upper bound for torsional restriction and gains approaching this magnitude are only feasible in very tight protein-ligand interactions. As the simulations indicate that ligands bound to MUP do not undergo much conformational restriction, the global entropic signature must arise from another source, or alternatively be derived from multiple smaller sources.

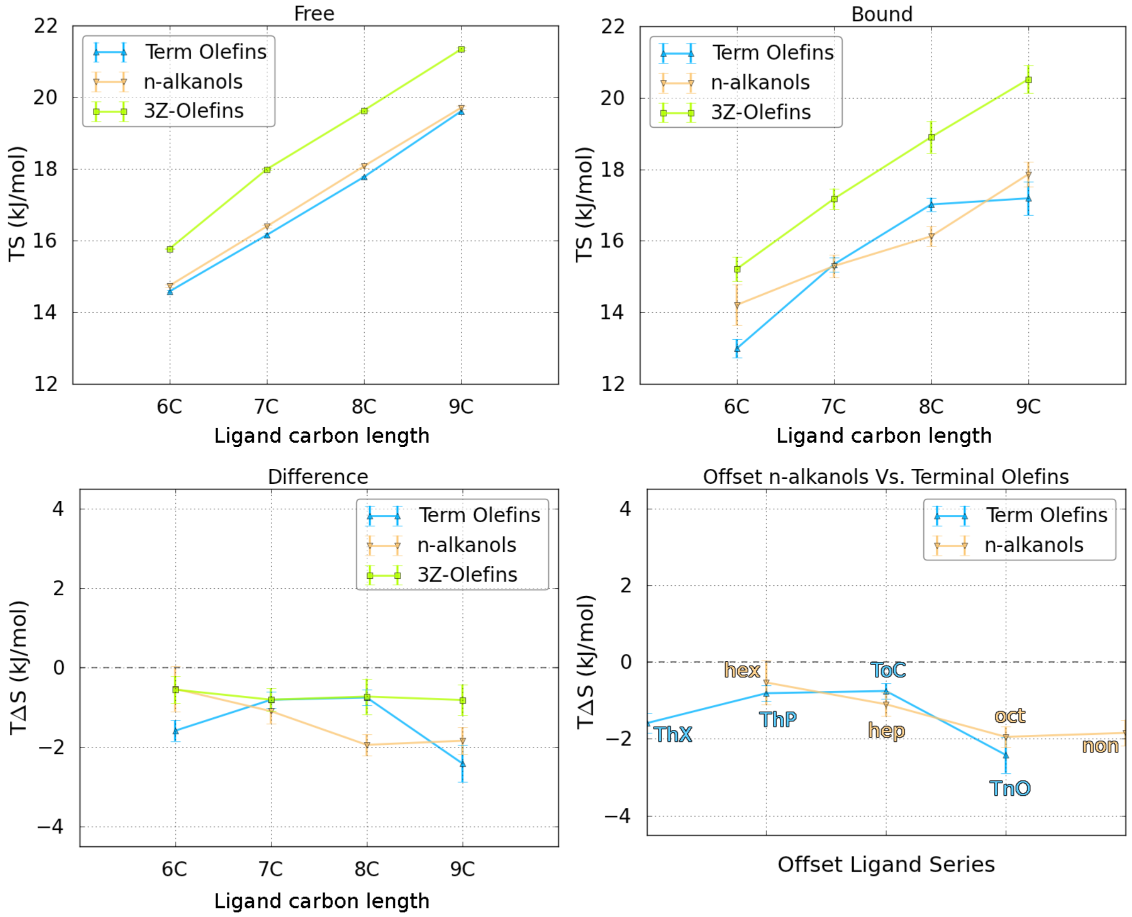


Fig.3.20. Cross panel comparison of entropy totals for all 3 ligand panels. Graphs clockwise depict the free TS_{Conf}; bound TS_{Conf}; Offset $T\Delta S_{Conf}$ values for n-alkanols versus Terminal olefins; and "bound minus free" $T\Delta S_{Conf}$ totals. Each data point is calculated from an aggregate 600ns simulation.

Summed Conformational entropies (kJ/mol)			
	Free	**Bound**	**Difference**
hex	14.74 ± 0.04	14.20 ± 0.57	-0.54 ± 0.57
hep	16.39 ± 0.04	15.29 ± 0.31	-1.10 ± 0.31
oct	18.07 ± 0.05	16.12 ± 0.27	-1.95 ± 0.27
non	19.71 ± 0.05	17.87 ± 0.34	-1.84 ± 0.34
ThX	14.58 ± 0.03	12.99 ± 0.26	-1.59 ± 0.26
ThP	16.15 ± 0.05	15.32 ± 0.20	-0.80 ± 0.21
ToC	17.77 ± 0.03	17.03 ± 0.19	-0.75 ± 0.2
TnO	19.61 ± 0.06	17.19 ± 0.47	-2.41 ± 0.47
3c6	15.76 ± 0.01	15.21 ± 0.34	-0.55 ± 0.34
3c7	17.99 ± 0.03	17.17 ± 0.29	-0.82 ± 0.29
3c8	19.63 ± 0.02	18.92 ± 0.45	-0.72 ± 0.45
3c9	21.35 ± 0.04	20.53 ± 0.39	-0.82 ± 0.39

Table.3.7. Summed 1st order TS_{Conf} and $T\Delta S_{Conf}$ values for all three ligand panels. Each data point is calculated from an aggregate 600 ns simulation and the standard error is propagated and calculated from entropy values generated for 6 individual 100 ns simulations.

Linear additivity is observed in calculated TS_{Conf} values for free n-alkanols and errors are close to zero, indicating convergence. The principal source of error comes from the bound state and is propagated through to the $T\Delta S_{Conf}$ calculation. As the differences are also small, resolution of the resulting trends is near the boundary of accuracy attainable with this method. Bound n-alkanols possess a shallower TS_{Conf} slope, albeit with small non-additive differences on increasing ligand length. Though bound non lies out-of-line with the trend and has a higher entropy than oct, the line is linear within error. It is probable that the calculated value is an overestimate and correlated motions might lower the magnitude of this data point (**Fig.3.14**). If so, the favourable 1st order entropy-entropy compensation effect would be reduced and the value for $T\Delta S_{Conf}$ can be extrapolated to approximate -2.5 to -3.0 kJ/mol. Note that the entropy values of other ligands in the series would also undergo adjustments in accordance to their lengths.

When positioning a restriction at the terminus of a linear alcohol, bound ligand dynamics affect TS_{Conf} in a non-additive, non-linear manner that cannot unequivocally be reduced to a group effect. It was postulated that a double bond at this position would be energetically equivalent to removal of an n-alkanol rotor. If $T\Delta S_{Conf}$ values for n-alkanols are offset against terminal olefins (**Fig.3.20),** the values are close enough (assuming the upper margin of error) to partially agree with this hypothesis. However, the torsional entropy component does not yield the expected 4.5 kJ/mol gain. The (limited) entropic benefits are due to a combination of factors dependent on ligand length. This is illustrated by the observation that 7C and 8C ligands from both restricted panels have almost identical $T\Delta S_{Conf}$ values.

Modifying saturated ligands with a central restriction results in free TS_{Conf} values being non-additive on addition of a rotor, although slopes largely maintain linearity. There is little difference between bound and free 3Z-olefins, and $T\Delta S_{Conf}$ values for ligands within this panel are the same within error. This indicates that introduction of a central double bond alleviates torsional entropy losses on binding to a similar extent, irrespective of ligand length. As previously discussed, this is probably due to pre-organisation of the ligand to the shape of the binding cavity and the effective shortening of the compound into two shorter segments. It is likely that the latter could be adapted for other protein-ligand binding interactions to mitigate torsional entropy losses in alkane chains, but the former is dependent on matching geometries of the binding partners and is therefore system specific. However, favourable entropic gains do not approach $T\Delta\Delta S_i$ values of 13.6 kJ/mol obtained from the decomposition of ITC values (§3.1.0).

Common conceptual models involving protein-ligand binding such as the induced fit and the lock and key hypotheses inculcate the notion that favourable interactions invariably involve a measure of immobilisation. However, MUP is a promiscuous protein whose ability to bind multiple binding partners is due to a lack of specificity. Discrimination is typically bestowed by a tight, well-defined binding pocket that complements the shape of the ligand and multiple protein-ligand polar bonds that act to restrain its position. As these characteristics are not possessed by MUP, ligands do not undergo extensive conformational restriction as expected and large gains in torsional entropy are not realised upon pre-organisation. The results in this chapter only account for torsional entropy differences and do not factor other contributions from T&R, solvent or protein. The total quantification of entropic effects is difficult to assess with rational design. Even simple modifications result in marked differences in dynamics and have the potential to perturb the underlying network of interactions constituting the system. Therefore, the system must be considered holistically before any final conclusions regarding purported benefits are drawn. The global entropic signature is composed of many contributions that have the potential to accumulate or cancel, and the magnitude of these disparate forces are likely to dwarf those derived from the ligand alone.

3.3.8. Convergence of total TS_{Conf} and $T\Delta S_{Conf}$ values

To assess convergence, plots were created that visualise changes in conformational entropy estimates as the number of data points are increased (**Fig.3.21**). Calculation of the entropy is notoriously difficult as phase space has to be thoroughly sampled to obtain an accurate estimate and true convergence would require infinitely long simulations [248,249]. However, we are not concerned with pinpoint accuracy and are content if clear (1st order) trends can be established. The entropy depends on the density of the number of states possessed by a given observable such as a dihedral angle. Without prior knowledge of the correct entropy value, it is impossible to know whether convergence has been attained. As simulation length is increased and/or additional independent repeats are added to the calculation, newly accessed areas of phase space are factored into the calculation and the entropy estimate increases. e.g. In **Fig.3.21,** smaller datasets (e.g green dashed lines) generally have lower entropy estimates than larger datasets (black and blue dashed lines). Given a finite dataset, calculated values oscillate as the population balance of existing states shift. However, the magnitude of fluctuations diminishes about a stable average as data size is increased and further entropy increases only occur upon the system accruing significant populations of underrepresented states. It is expected that convergence would be much more difficult for protein dihedral angles than those in the ligand. Adequate sampling for the former is complicated by conformational restrictions imposed by other residues within the polypeptide chain (§3.2.5).

Even with a small number of data points, conformational entropies in the free state show marginal differences compared to entropies calculated with larger datasets. Additionally, gauche(-) and gauche(+) states within free dihedrals between adjacent methylenes possess

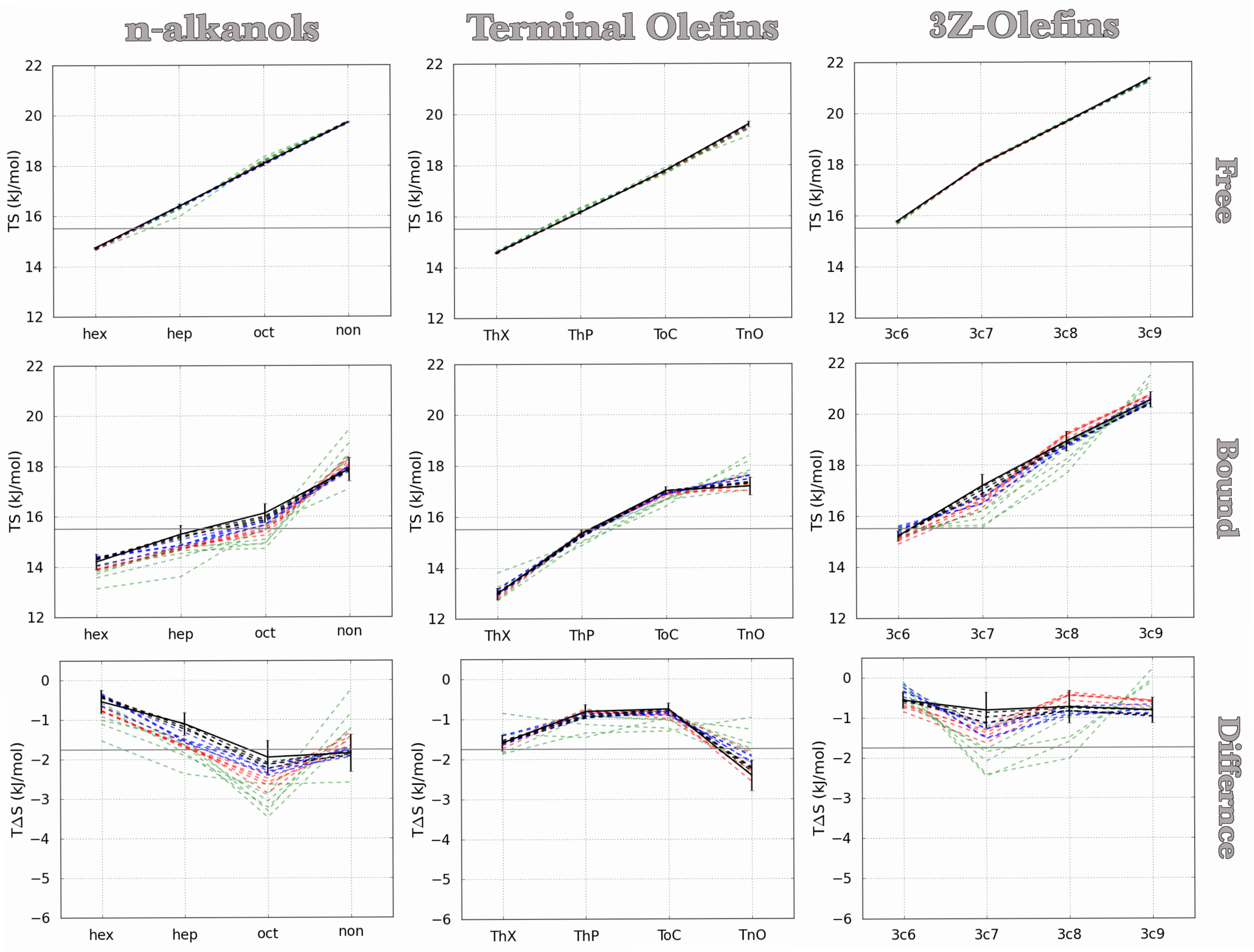


Fig.3.21. Graphs show the convergence of the total TS_{Conf} and $T\Delta S_{Conf}$ values over 600 ns of aggregate simulation time for all 3 panels of ligands. Solid black lines show the value calculated for the full 600 ns. Coloured dashed lines indicate how the entropy total for a given ligand changes with the number of points included in the calculation. The number of points available was split into 150 ns categories, which were further subdivided into 30 ns segments. Lines coloured green contain less than 150ns worth of data points; red 300 ns; blue 450 ns; and black 600 ns. The scale of the difference plot is adjusted to allow optimal visualisation of fluctuation in the data.

almost identical populations (**Fig.A1.1-3**). As the energetic favourability of these two states are indistinguishable, any imbalance would be indicative of imperfect sampling [219]. So, it is unlikely that further sampling would significantly affect the entropy calculation for the free species. As expected, the principal source of variability occurs for bound entropy estimates and this is propagated through to the relative entropy differences. Apparent fluctuations in the latter are amplified because the differences between bound and free states are so small. Therefore, assessment will focus on the bound species. The balance between bound trans, gauche(-) and gauche(+) conformations is primarily affected by the ligands interactions with the architecture of the binding cavity, and it is only possible to assess convergence within the context of the available dataset. Terminal olefins seem to be well estimated by the method, whilst n-alkanols and 3Z-olefins undergo comparatively greater fluctuations. It is possible that further increases in sampling could unlock additional states that affect the entropy calculation. However, **Fig.3.21** indicates that the magnitude of fluctuations decreases with greater numbers of data points. This suggests that corrections would not be considerable, but it would only take around 1 kJ/mol to significantly affect the shape of the trend line as the differences between ligands constituting a panel are so small. This is particularly evident for 8C and 9C ligands which have a greater number of flexible subunits and would thus be harder to converge. Longer ligands are also more likely to be affected by correlated motions in second or higher order calculations.

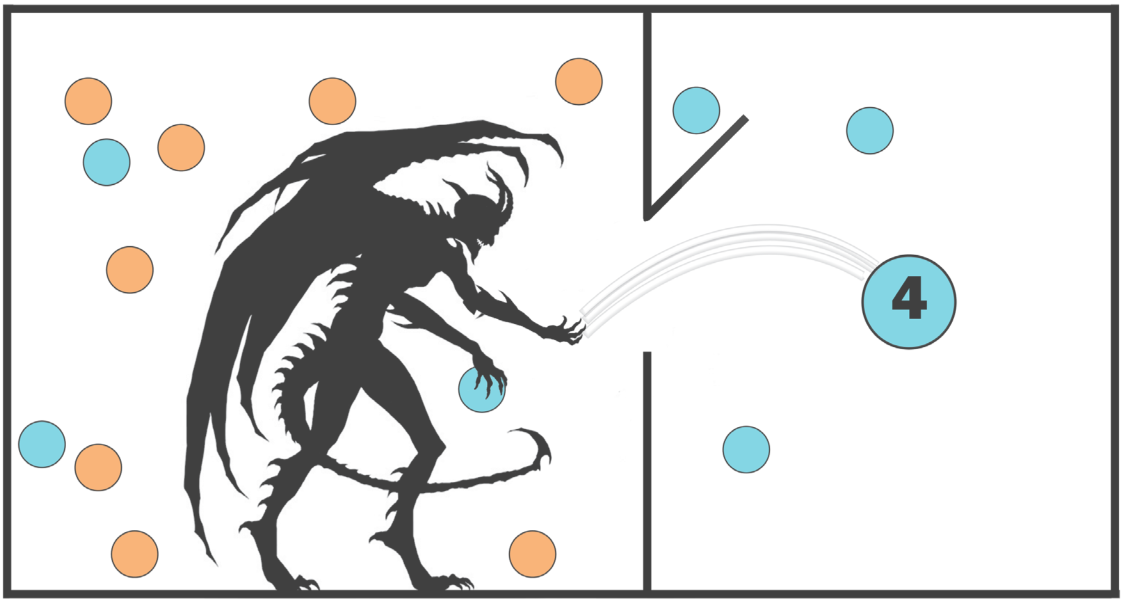
3.5.0. Conclusion

The binding of alcohols to MUP is interesting as the linearity of the global thermodynamic values suggest that this is an uncomplicated system whose binding signature is governed by simple differences. As the only perturbation to the composition of the system is a discrete, incremental difference in ligand structure, the magnitude and additive nature of the globally measured values were thought to arise from the ligand alone.

This chapter concentrated on assessing the internal conformational entropy of three ligand panels as this was considered the most likely source of loss in ligand DOF subsequent to binding. The experimental design of the panels systematically tested the loss of a rotatable bond via addition/removal and through the engineering of a double bond restriction. Both types of modification resulted in $T\Delta S_{Conf}$ values very much smaller than the global values obtained from ITC. This was primarily because the ligands possess significant residual motion whilst bound and the range of internal torsions are very similar to that of the free species. If methods were used to obtain second or higher order entropy estimates, differences would probably be even smaller. Therefore, dihedral $T\Delta S_{Tor}$ values do not approach the magnitude observed for tight binding complexes and thus the conformational entropy is not wholly responsible for the global entropic signature. Fortuitously, ligands within the 3Z-olefin panel were structurally pre-organised to the shape of the binding cavity. This resulted in attenuated $T\Delta S_{Conf}$ losses. Additional entropic benefits were accrued by the ligand being split into two shorter sub-segments by the central restriction. While this did not translate into any large conformational entropy gains within MUP, the modification might be beneficial if implemented in other systems. However, this would have to be assessed by thorough benchmarking.

Various authors have suggested rotor restriction as a method to improve binding affinities. However, the results indicate that this is not a rule of thumb that can be applied without detailed reference to structural features pertaining to both the protein and ligand. Insertion of a double bond cannot be neatly reduced to a group effect, as this ostensibly simple unit modification results in complex dynamic behaviour that is dependent on its position, the length of the ligand and the architecture of the binding pocket. These factors have the propensity to affect measured thermodynamic values in an unpredictable manner. This highlights the power of MD techniques to visualise and ascribe rationale to the effects of structural modifications whose purported benefits have been primarily adduced through thermodynamic accounting. With regards to drug design, the granularity offered by MD is fine enough to offer a level of detail that better informs the choice of structural modifications and their consequences. Global values and partial decompositions often obfuscate the subtlety of ligand dynamics and their relationship within the coupled network of polymer interactions that constitute protein-ligand binding.

As the contribution of ligand internal DOF to the global entropic signature was minimal, chapter 4.0 extends the analysis by considering the contributions from ligand translational and rotational entropy.



Maxwell's Daemon

Chapter 4.0: Ligand Translational & Rotational Entropy

4.1.0. A process of elimination

As changes in the conformational entropy were insignificant on binding simple alcohols to MUP, this chapter continues to investigate the underlying causes for the decrease in global entropy across ligand panels (**Fig.3.1**). Causal arguments linking the increase of ligand length to a decrease in the global entropy of binding are still cogent due to the linear change observed on perturbing the system (§1.3.3). Thus, the most likely causative agents for the observed trends are likely to originate from translational and rotational (T&R) contributions to the binding entropy. Translational penalties are thought to scale weakly with ligand size and the contribution is expected to be very similar for compounds within the tested panels [172]. However, disparities between the principal rotations of the different ligands are much more likely to offer the correct linear gradient and requisite magnitude of entropic losses. This view is reinforced by the fact that differences in conformational entropy could be attributed to differential ligand dynamics (Chapter 3.0). If summed T&R contributions do not equal the global entropy signature, protein and solvent contributions must be considered in an iterative process of elimination.

The majority of methods to calculate T&R entropies make the assumption that bound ligands occupy a single bound minimum within the binding site. Hence, some form of harmonic function is generally used to approximate the bound distribution [164,190,250–252]. This is likely to lead to complications in the case of MUP as the protein is a promiscuous binder and the multiplicity of available partners indicate that ligands are likely to adopt several bound minima. This is supported by the fact that the electron density of 3Z-olefin ligands bound to MUP is much more diffuse than their saturated analogues and bound compounds were modelled with two different poses [180]. Other examples of complexed crystal structures that contain ligands with alternate poses can be found in the literature e.g. T4 lysozyme, HIV PR, T4 lysozyme, neuraminidase, thymidylate synthase and cytochrome P450cam [253,253–256]. This phenomenon is germane to rational ligand design as the simplistic assumption of a single bound minimum introduces errors into entropy estimates and reduces the ability of computational methods to predict the suitability of drug-like compounds. Moreover, assessing inhibitor dynamics and the consequent binding site interactions could potentially provide insights into tackling drug resistance associated with protein mutation [61].

4.1.1. Promiscuous proteins & ligands: Shake, rattle & roll

There is considerable evidence that protein-ligand interactions cannot be simplistically reduced to highly specific one-to-one monogamous relationships, and proteins and their ligands frequently interact with a variety of partners. This is known as promiscuity [161,162,257–262]. There are multiple classifications of promiscuity, but lack of protein substrate specificity is likely to have the have the greatest impact on small molecule inhibitor design.

A survey by Gao et al. (2013) analysed all ligand bound protein structures in the RCSB PDB and created a non-redundant dataset containing 20,414 members and 9,485 unique ligands [161]. The protein binding pockets and their interactions with bound ligands were then analysed using pocket similarity (PS) scores which were generated by structurally aligning backbone geometries and assessing the chemical composition of binding site residues. These measurements coupled with a global structural similarity metric indicated that highly similar protein cavities bind the same cognate ligands in spite of low overall structural homology. e.g. Adenosine diphosphate (ADP) binds to ADP pockets of disparate protein receptors. Furthermore, structurally unrelated proteins have the ability to bind similar ligands, and 52 - 70% of protein complexes in the dataset could be matched to another pocket containing a similarly bound cognate ligand [161].

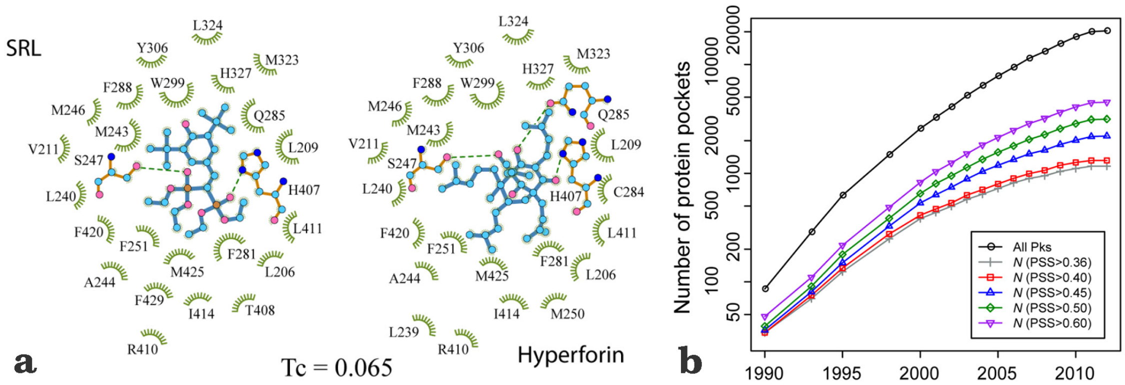


Fig.4.1. (a) - Binding of two structurally dissimilar ligands (Hyperforin and SRL) to the hydrophobic binding pocket of the promiscuous pregnane X captor nuclear receptor protein. A Tanimoto coefficient (Tc) of 1.0 indicates identical compounds. **(b)** - Logarithmic scale showing number of representative pockets found by year. A pocket similarity (PS) score of 1.0 denotes identical pocket structures, whereas a score of 0.0 is indicative of total dissimilarity. Significant similarity is observed for pockets with a PS-score greater than 0.36. Images taken from reference [161].

Clustering allowed the authors to conclude that a subset of only 1,315 pockets could be used to encapsulate the degenerate structural space of all protein binding pockets in the dataset. From the 1990s to the year 2000, the number of new representative pockets submitted to the PDB showed an exponential increase. However, after this period, the rate began to decline and will most probably plateau in the coming years (**Fig.4.1.b**) [161].

Additional analyses indicated that proteins tend to interact with multiple ligand types, often with disparate chemical structures. At heart, this is the definition of a promiscuous protein and these entities were found to compose > ~33% of the dataset. Smaller ligands can access alternative locations within larger pockets and create favourable interactions with other amino acid residues that happen to possess complementary physio-chemical characteristics. This adaptability is further aided by conformational changes associated with ligand and protein flexibility, whilst directional polar bonds that afford specificity made up only ~28% of all interactions. Thus, ligands scaffolds have significant leeway to adapt and conserve favourable protein-ligand interactions within the context of a number of structurally diverse binding pockets (**Fig.4.1.a**) [161]. e.g. Protein kinase A (PKA) uses a single hydrophobic interface to bind multiple A-kinase anchoring proteins (AKAPs) and this enables it to translocate to different compartments within the cell. Binding is entropically driven as each AKAP can adopt a variety of binding poses because the set of hydrophobic contacts it possesses are able to match a variety of alternate contact points on PKA [160].

Promiscuity has serious implications for the rational design of small molecule inhibitors that are able to retain their function because proteins targets often develop mutations that confer resistance, and unintended physiological side effects are likely to be due to off-target interactions. The propensity for both protein and ligands to be promiscuous, complicates the engineering of high affinity complexes possessing clinically acceptable levels of specificity [161,257,260,261,263]. Natural compounds tend to have greater structural complexity (chiral centres and multiple ring systems) than synthetically designed inhibitors, and this is likely to result in highly specific interactions [264,265]. Most importantly, binding promiscuity highlights the necessity to take into account the possibility that the protein can adopt different conformations and could well bind ligands in multiple locations. Thus, the number of variables and their interdependence on each other render this a multi-dimensional problem and this demonstrates that real advances in drug design can only be made by defenestrating static perceptions of binding.

4.1.2. The practicalities of calculating T&R entropies

4.1.2.1. Why is it important to calculate contributions from external DOF?

Endpoint free energy methods such as MM(P/G)BSA and linear interaction energy (LIE) seek to minimise the computational cost associated with pathway methods such as FEP and TI by only considering the difference between free and bound states. Theoretically, this avoids the expensive simulation of unphysical intermediate states along the pathway that ensure adequate exploration of phase space [266–271] (§2.12-4). Whilst pathway methods yield a direct evaluation of the free energy, endpoint methods rely on this value being "built up" from component enthalpies and entropies. An advantage of this approach (with reference to MM(P/G)BSA) is that, facilities such as per-residue free energy decomposition affords greater insights into relationships between structure and function [272–280]. However, the larger scientific community generally regards accurate quantification of the entropy as being too technical, difficult and/or time consuming. Hence, this portion of the calculation is sometimes neglected entirely, or simple models such as the QHA or normal mode analysis are used to recover the entropy associated with internal DOF [281–284]. Indeed, the MM(P/G)BSA tutorial on the amber website states that:

"The entropy contribution can be found by performing normal mode analysis on the three species but in practice entropy contributions can be neglected if only a comparison of states of similar entropy is desired such as two ligands binding to the same protein." [285]

This approach presents a couple of serious issues that limit accurate evaluation of the free energy, and consequently impedes the discovery of efficacious drug-like compounds. Firstly, in the case of ligands that possess a single bound minimum, the entropy contributions from external DOF (i.e. T&R) has the potential to dwarf the conformational component which only accounts for internal DOF. Secondly, the true entropic contribution is very difficult to predict in the case of ligands that adopt multiple bound minima. As demonstrated in chapter 3.0, significant differential dynamics can arise from small disparities in structure. Hence, if the ligand retains substantial residual motion, it is also likely that this difference is not captured by the conformational entropy, but is instead reflected within external DOF. Furthermore, such a paradigm would mean that it would be impossible to ascertain whether the entropic component could be safely neglected with regard to homologous ligands.

The most compelling reason for the evaluation of external entropies is the ability to probe the relationship between structure and dynamics. Approaches such as quantitative structure-activity relationships (QSAR) and others have factored in structural complementarity [286,287]. Some steps have been taken to redress this static notion of binding by accounting for alternate molecular conformations e.g. conformation-activity relationships (CAR) [288–290]. However, given the prevalence of promiscuous binders, it is clear that only accounting for internal DOF is not enough. As demonstrated in this chapter, the quantification of external entropic components provides greater insights into the relationship between structure and dynamics.

4.1.2.2. The Sackur-Tetrode equation

The simplest method of assessing translational entropy losses in bimolecular binding is via the use of the Sackur-Tetrode (ST) equation (**eqn.4.1**). Independently derived by Otto Sackur and Hugo Tetrode, it allows the calculation of the entropy of a monatomic ideal gas; a species that possesses only three translational DOF. The equation can be written as the following, where U is equal to the internal energy, m the mass, k_B the Boltzmann constant and *h* Planck's constant.

$$S(E,N,V) = Nk_B \log\left(\frac{V}{N}\left[\frac{4\pi mU}{3Nh^2}\right]^{\frac{3}{2}} + \frac{5}{2}\right) \qquad \textbf{(eqn.4.1)}$$

The two main components to the equation are related to the potential and kinetic energy. The volume relates to the spatial distribution of particle positions (r), whilst the inner bracketed term associated with U encompasses the distribution of particle conjugate momenta (p). The corrective 5/2 term prevents the overcounting of states in the case of non-distinguishable particles and is necessary to avoid the Gibbs paradox. The quality of the ST equation is such that, the theoretical values it provides for ideal gasses are currently considered superior to those obtained by experiment [291–297]. An expanded discussion regarding the derivation of this important equation can be found in the appendix (§A2.1.1).

4.1.2.3. Rigid and flexible single molecule partition functions

As discussed in §2.1.3, there are a number of ensembles that consist of collections of microstates whose collective properties describe the equilibrium state of the system. By appropriately defining the partition function (Z) of a system it is possible to relate the microscopic fluctuations of particles to various macroscopic properties via statistical mechanics. The probability of a microstate Ω_i with energy E_i being accepted in the ensemble is proportional to the appropriate Boltzmann factor ($e^{-\beta E_i}$), where -β equals $1/k_BT$. Defining a molecular partition function suitable for *in silico* simulations necessarily involves building up an entire complex entity from more primitive elements. In the case of a single molecule, these elements would correspond to the available DOF: translational, rotational and vibrational. To simplify what might otherwise be an intractable calculation, the assumption of weak coupling is commonly used i.e. a small amount of energy transfer is allowed between DOF, but correlations are considered weak enough, that these elementary components are considered independent from one another. Quantifying the amount of correlation and assessing its impact on the final calculation is the topic of many other bodies of work (§3.1.2). However, this underlying assumption allows contributions from the main DOF to be calculated separately and the total entropy is yielded when they are summed (**eqn.4.2**) [291,293,298].

$$S_{Tot} = S_{Tr} + S_{Rot} + S_{Vib} \qquad (\textbf{eqn.4.2})$$

This gives MD the power to decompose dynamics into tractable parts and analysis of these components offer a more granular view that can better guide rational drug design [293]. Despite the success of the ST equation at predicting the entropies of ideal gasses, it does not translate well to data generated by molecular dynamics or Monte Carlo simulations, which are at heart, single molecule experiments. In some studies, the equation has been applied to biological problems along with its companion ideal-gas method to calculate the rotational entropy [171,252,299–302]. However, criticism has been levelled at this approach due to its inconsistent application and because it did not adequately account for the reaction taking place in solution e.g. A key question is whether the results are dependent on atomic mass or not. Additionally, the ST equation gives the entropy for N indistinguishable molecules, rather than being based on the single molecule partition function which derives the binding free energy from standard chemical potentials [164,303,304]. The most lucid exegesis to date is offered by Zhou and Gilson (2009) and the reader is referred there for further information on the associated issues [303]. The remainder of this subsection and the next summarises some of the key points from that work with reference to other sources when appropriate.

With respect to single molecule partition functions, it is useful to be aware of two statistical mechanical models. These are known as the rigid rotor harmonic oscillator approximation (RRHOA) and the flexible molecule (FM) approach. Each model contains discreet expressions for translational, rotational and vibrational contributions and these have to be applied to receptor, ligand and the bound complex to obtain the binding entropy. The RRHOA assumes that the molecule is essentially rigid and that internal motions are modelled as small vibrations via the use of a harmonic oscillator. In terms of this partition function formulation, freezing larger internal vibrations of a molecule and assuming inflexibility allows *the kinetic and potential energy contributions to the partition function to be factorised and evaluated separately*. Thus the potential energy is assessed by the effective volume in coordinate space and this is multiplied by the contribution from the momenta which includes a mass dependent term (**eqn.4.3**). Note the similarities to the ST equation (**eqn.4.1**).

$$Z_{Tr} = V\left(\frac{2\pi m}{\beta h^2}\right)^{\frac{3}{2}} \qquad \textbf{(eqn.4.3)}$$

The entropy arising from the principal rotations of the rigid molecule can now be calculated using three principal moments of inertia (I_1 I_2 I_3) about the COM (**eqn.4.4**).

$$Z_{Rot} = 8\pi^2\left(\frac{2\pi}{\beta h^2}\right)^{\frac{3}{2}}\left(I_1 I_2 I_3\right)^{\frac{1}{2}} \qquad \textbf{(eqn.4.4)}$$

Additionally, the harmonic approximation allows each internal vibration to be quantum mechanically modelled via the angular frequency (ω_i). ω_i =(k/m)$^{1/2}$, where k is the oscillator spring constant (**eqn.4.5**).

$$Z_{vib} = e^{-\beta E_0}\prod_i \frac{e^{\frac{-\beta h\omega_i}{4\pi}}}{1-e^{\frac{-\beta h\omega_i}{2\pi}}} \qquad \textbf{(eqn.4.5)}$$

In contrast to the flexible molecule formulation, all three components of the RRHO partition function are mass dependent and this dependence cancels out when they are used to calculate the total entropy. An incorrect estimate of the total entropy will be obtained if the RRHO equations are combined with alternative expressions that do not result in mass cancellation [303–306].

On the other-hand, the FM partition function eliminates the need to calculate the contribution from momenta at the outset and is therefore mass independent. Most MD simulations are run at room temperature and the mass of the system remains constant. Moreover, due to the equipartition theorem, each DOF makes a fixed contribution to the kinetic energy and this is unaffected by entities moving from the unbound to the bound state. Therefore, a classical statistical thermodynamic treatment is more appropriate than a quantum one. This means that the only thing left to evaluate is the potential energy which is evaluated by considering the configurational integral over spatial coordinates i.e. the *effective* volume. The assumption of rigidity in the RRHOA can be discarded as there is no longer any unwanted correlation between the potential and kinetic energy and the requirement to calculate fixed moments of inertia evaporates. The translational component contributes a factor of V to the total partition function and the rotational contribution, 8π^2. The full range of internal vibrations can be considered using **eqn.4.6** in which the Jacobian (J) is a corrective factor utilised when transforming from Cartesian to internal coordinates.

$$Z_{\text{int}} = \int dr\, J(\mathrm{r})e^{-\beta E(\mathrm{r})} \qquad \textbf{(eqn.4.6)}$$

The full FM partition function can now be written as **eqn.4.8** [196,303].

$$Z_{\text{int}} = V\, 8\pi^2 \int dr\, J(\mathbf{r}) e^{-\beta E_{(\mathbf{r})}} \qquad \textbf{(eqn.4.8)}$$

A further distinction between rigid and flexible formulations is the way DOF before and after binding is treated. Rigid models often assume that both ligand and receptor have 3 translational and 3 rotational DOF when free in solution. Subsequent to binding, the complex as a whole retains 6 external DOF, whilst the ligand's external DOF can be treated as 6 new vibrational DOF that are incorporated into the complex. In contrast, the FM approach allows the ligand to retain its external DOF and this can be more easily calculated from the ligand's motion relative to the protein [164,190,250,271,303,304]. Additionally, while the total entropy returned by both approaches should be similar, values obtained from vibrational, T&R subcomponents will differ due to the different partition schemes utilised [303].

4.1.2.4. Translational entropy is dependent on the standard concentration

In the case of protein-ligand binding under the NPT ensemble, the thermodynamic potential that describes the macrostate of the system is the Gibbs free energy. The standard free energy of binding (ΔG°_b) is determined experimentally under standard state conditions when the three species are at equilibrium and ΔG°_b is zero (**eqn.4.9**) i.e. a temperature of 298 K, 0.1 MPa pressure and 1 M standard concentration (C°). The choice of C° is arbitrary, but a value of 1 M is most commonly used for convenience. The reciprocal of this value (V/1 M) is the standard volume (V°) and this equates to 1,660 Å^3 in the case of a single molecule [164,190,250,291,303].

$$\Delta G^\circ b = \text{-kBT} \ln(C^\circ\ Ka) \qquad \textbf{(eqn.4.9)}$$

The standard enthalpy of binding is *independent* of C° and is given by **eqn.4.10**.

$$\Delta H_b = -k_B\, \mathrm{T}^2 \left(\frac{\delta \ln K_a}{\delta T} \right)_P \qquad \textbf{(eqn.4.10)}$$

Per Contra, the standard entropy of binding (ΔS°_b) is also *dependent* on the choice of C° and therefore V° (**eqn.4.11**).

$$\Delta S_b^\circ = k_B \ln(\mathrm{C}^\circ \mathrm{K}_a) + \mathrm{k}_B\, T \left(\frac{\delta \ln K_a}{\delta T} \right)_P \qquad \textbf{(eqn.4.11)}$$

In the case of a simple monatomic ligand binding to a receptor (at constant temperature), the translational entropy can be calculated as the ratio of the standard volume (V°) that the ligand has access to when free, and its bound volume (V_b) (**eqn.4.12**).

$$\Delta S_b^\circ = k_B \ln(C^\circ V_b) = k_B \ln\left(\frac{V_b}{V^\circ} \right) \qquad \textbf{(eqn.4.12)}$$

Hence, the translational entropy is the only entropic subcomponent that possesses a dependence on the standard concentration [303,304]. Other contributions such as those arising from solvent, rotational and vibrational DOF can be calculated separately and added to obtain a 1st order entropy estimate.

4.1.2.5. A brief overview of methods used in the literature

Various approaches have been used to calculate translational and rotational entropies and all of them cannot be described due to space limitations. Irudayam et al. (2009) [164] have conveniently categorised some of the available literature in terms of experimental methods [29,163,170,171,301,307,308,308–316], approximate estimations [251,299,302,309,317–321] and computational approaches [250,271,322,323]. The last category is further subdivided into approaches that rely on numerical integration [324,325], normal-mode analysis [326], docked ensembles [327] and distribution functions [240,322,328–331].

Many of the experimental approaches are cleverly designed and prompt deeper discussions regarding the concepts underlying T&R entropy losses. However, such techniques are inevitably hamstrung in terms of broad applicability, due to the inherent assumptions accompanying the decomposition of the global entropy into external and internal contributions [308,332,333]. Whilst, gas phase entropies are somewhat easier to characterise, the application of similar techniques to solvated systems present considerable difficulties due to questions regarding the extent to which solvent hinders gross motions. Various T&R entropy losses for molecules in solution have been proposed, ranging from 9 to 63 kJ/mol. Lundquist and Toone (2002) state that, the commonly cited figure of ~44 kJ/mol provided by Jencks is a rough rule of thumb that is likely to be an overestimate; the true value being unlikely to exceed ~25 kJ/mol [172,334]. Thus, it should be realised at the outset, that there is a lack of experimental data on decomposed T&R contributions that offer unequivocal conclusions with respect to complex macromolecular systems.

4.1.2.6. Early experimental & *in silico* entropy decompositions

Finkelstein and Janin (1989) proposed a novel method to calculate translational and rotational entropies from the RMS fluctuations derived from crystallographic B factors [251]. Their formulae assumed that the relative motions of two molecules within a complex are similar to that observed in crystal structures and the translational and rotational entropy of a bound molecule could be obtained from RMS displacements and amplitudes along each principal axes. Thus, $S_{Tr} = R \ln(\delta_x \delta_y \delta_z/1660 Å^3)$ and $S_{Ro} = R \ln(\delta_\theta \delta_\psi \delta_\varphi/8\pi^2)$, where 1,660 $Å^3$ and $8\pi^2$ correspond to free translational and rotational configurational volumes respectively [250,251]. A translational entropy estimate of ~28 kJ/mol was obtained for molecules possessing bound volumes of 8.0 x 10^{-3} to 1.6 x 10^{-2} $Å^{-3}$ in lysozyme crystals [251]. Whilst the formulae and concepts regarding the relative motions of the bound ligand to protein are solidly grounded, it is questionable whether small RMS fluctuations from cryocooled crystals provide an accurate measure of the bound volume and this assumption has been queried [250,271,323,335]. In the same paper, F&J compared solid naphthalene to that of a very tightly bound complex; the gas to an ideal solution and sublimation to complex disassociation. The comparison of protein complexes to crystals was extended by Searle and Williams who produced a particularly interesting manuscript that decomposed the entropy of fusion from weakly ordered hydrocarbon crystals [170]. They estimated that 9 to 15 kJ/mol came from T&R, and 1.6 to 3.4 kJ/mol from internal DOF (§3.3.4). In order to match the F&J result, they concluded that molecules in the crystal had to either retain large amplitude motions or that the melted product was more ordered than expected. However, it is now known that the insights that (cryocooled) crystallography generates on dynamics is limited. The static view it provides, imparts the impression that proteins and their bound ligands are far more ordered than the actuality (§A2.1.2).

An alternative approach to the treatment of bimolecular binding emerged from a series of investigations into the dimerisation of insulin. The T&R contributions to binding were decomposed using the ST equation and a primitive rigid body model (that used featureless cylinders) were used to calculate the disassociation of the insulin dimer into monomers. The entropic cost for restricting *external* DOF was estimated at ~153 kJ/mol and was considered far too high, because the global entropy of binding equated to only ~12 kJ/mol and thus, the authors suggested that solvation must play a significant part [299]. However, a rough estimate based on the burial of SASA indicated that global entropy losses were only partially ameliorated [321]. Several authors had suggested that relative residual motions and vibrations in the complex would further reduce the large negative penalty associated with T&R [37,171,317]. For example, Erickson (1989) estimated a combined estimate of ~46 kJ/mol by assuming that bonded subunits could undergo relative displacements of ~2 Å in each direction (using **eqn.4.12** for translation). He further suggested that co-cooperativity would further decrease this value to ~29 kJ/mol [317]. The alternative approach implemented by Tidor and Karplus (1994) modelled residual motion as 6 new vibrational modes via normal mode analysis [336]. A total entropic loss of -84 kJ/mol was computed; the bulk of which came from a large unfavourable T&R contribution (-114 kJ/mol). This was partially compensated by a favourable vibrational component (30 kJ/mol) which arose from changes to the density of states and consideration of new vibrations within the complex. This experiment highlighted the necessity to account for all DOF when calculating the total entropy. A strength of the approach over that proposed by F&J is the use of quantum mechanical equations which would capture changes associated with narrow energy wells, whilst a disadvantage is the use of the harmonic approximation [240]. There is also the question as to whether this model might be more appropriate to covalent binding, which would see a far greater reduction in residual motion than non-covalent binding [164,240,250,319].

4.1.2.7. The utility & pitfalls of approximations

Whatever the approach, the primary problem associated with the calculation of T&R contributions to TΔS° is accurate estimation of the *effective* free and bound configurational volumes. Access to long timescales was hampered in early MD simulations due to computational costs, and thus, recovery of the ergodic distributions associated with protein DOF was difficult, if not impossible. As discussed in §3.1.2, approximations such as the ad hoc method proposed by Schlitter [202] and the QHA [200,337] circumvented the sampling issue by assuming that a functional form was sufficient to capture the underlying distribution. Uncertainty inherent in such approximations means that these methods only provide an estimate of the *maximum* entropy. The QHA is believed to provide a tighter upper bound as it is supposed to better account for anharmonicities. However, the utility of the maximum entropy estimate as a proxy for the true entropy has been questioned [203,338] and testing has demonstrated that these approximations can significantly overestimate contributions arising from principal rotations and torsional DOF [203,250,339]. Nevertheless, both these methods can be extended to separate translation from rotational and vibrational components via the use of translational and rotational RMS fits. However, it is difficult to deconvolute rotational and vibrational contributions as they are intrinsically linked and the fitting procedure can introduce significant errors into the calculation [203,250,339]. The problem is exacerbated in the case of flexible ligands and most examples in the literature focus on relatively rigid compounds [164,190,250,271,319,325,326,340]. In principle, histogramming can capture multi-modal distributions with greater accuracy as there is no assumption of any functional form, but this approach has historically been limited by sampling and difficulties in accessing minima separated by unfavourable energy barriers [190,250,339]. Some authors have estimated the ligand's bound translational volume (for use in **eqn.4.12**) by integrating under histograms generated from COM motions along each principal axes, or in the case of residual rotational motion, Euler angles [240,328,339]. This approach assumes that motions along the three orthogonal axes are independent from one another and care must be taken to account for correlations between them.

There are other techniques that utilise **eqn.4.12** such as FEP, potentials of mean force, etc. However, free volume methods form a distinct category to what has been described thus far, and the salient points of this approach will be discussed below.

4.1.2.8. Free volume methods

There are currently two opposing schools of thought regarding the correct method to calculate T&R entropy contributions and a greater proportion of the literature is devoted to what Irudayam and Henchman (2009) [164] term the molecule frame (MF) approach [186,240,250,251,271,299,302,304,318,321–323,326,327,329–331,336,339,341,342] versus system frame (SF) techniques [317,319,325,343]. The Fundamental difference between the two is that MF theories treat the ligand as a "particle in a box", wherein the free ligand is able to explore the entire system volume i.e. the standard volume of 1,660 $Å^3$. However, SF proponents argue that the "free volume" accessible to the ligand is much smaller due to the effect of solvent-caging and thus calculated T&R entropies should be much smaller than that calculated by MF approaches. Gilson (2009) makes the important conceptual point that free solute molecules are confined in solvent cages for very short periods of time and they will explore the entire system volume during the course of an experiment. The complexed entity can also access the full system volume along with its internalised ligand. Moreover, a bound ligand's translational motions are confined within the smaller volume of the cavity and are highly correlated with the protein. Thus, this correlation could significantly lower the calculated MF entropies which typically do not account for this higher order correction [303].

SF methods can involve controversial concepts such as the cratic and communal entropies. In order to assist the decomposition of the entropy of binding into terms associated with solvation, conformational DOF, etc it was proposed that the translational entropy could be estimated via the use of the cratic correction [311,344–346]. The model was applied to the problem of say, an insulin dimer fixed in space separating into two well separated monomers that are also fixed in space [347]. In a *hypothetically dilute* solution, the number of molecules is increased from 1 to 2 and the translatory contribution can be separated out by use of the equation $S_{Cratic} = -R \ln x$, wherein x is equivalent to the mole fraction of the solute in water. As water possesses a concentration of 55 M in the 1 M standard state, x = 1/55. Thus the cratic correction equates to ~10 kJ/mol at 300 K [347]. This approach attracted severe criticism from several authors, and Holtzer's (1995) theoretical analysis thoroughly addressed the associated issues and concluded that the theory had no foundations in thermodynamics [186,303,304,347]. As some of the very early estimates of the translational entropy (e.g. insulin) were very large, the smaller value returned by the cratic correction (~10 kJ/mol) was used to argue that the entropies obtained from the ST equation were grossly overestimated and thus not applicable to studies in solution [311,319,348]. Nevertheless, SF theories are difficult to implement due to the practical limitations of defining the free volume of the ligand within the mean frame of mobile solvent molecules, and there is disagreement as to the correct method [164,319,343]. For instance, an early approach proposed by Amzel (1997), suggested that $T\Delta S°_{Tr}$ could be estimated by summing the cratic and communal entropies and the entropic difference arising from the change in free and bound volumes [319].

The communal entropy arises from cell theories of liquids: an early theoretical framework that has been extended to address SF difficulties in calculating localised configurational volumes. If the total solvated system volume (V) is subdivided into N smaller cells of volume v = V/N, the motion of any particle is captured by a potential function that localises it within a single cell, and this is assumed to be independent of the motions of particles in neighbouring cells. This allows the total partition function of large, complex, inhomogeneous liquids to be calculated from the product of single molecule partition functions obtained from the potential function in each independent cell [319,349]. Particles in the solid state are also not permitted to exchange cells and are distinguished from the liquid state by *greatly reduced* free volumes (v_f). In contrast, particles in a gas readily exchange between cells, and are thus considered to possess extra "communal entropy". However, particles in a solid are distinguishable because the cells can be labelled and this is reflected in the equations for the partition function and entropies (**eqn.4.13-14**). The thermal de Broglie wavelength, Λ = h/(2πmkT) contains the momentum [298,319].

$$Z_{gas} = \frac{V^N}{N!\Lambda^{3N}} \qquad Z_{solid} = \frac{V^N}{N^N\Lambda^{3N}} = \frac{v^N}{\Lambda^{3N}} \qquad \textbf{(eqn.4.13)}$$

$$S_{gas} = Nk_B \ln\frac{v}{\Lambda^3} + \frac{5}{2}Nk_B \qquad S_{solid} = Nk_B \ln\frac{v}{\Lambda^3} + \frac{3}{2}Nk_B \qquad \textbf{(eqn.4.14)}$$

In the case of a gas and a solid with the same number density, the difference between the two entropies (k_B) corresponds to the communal entropy. This is relevant, because S_{solid} is used for the bound ligand which remains localised within its cell. The equation for S_{liquid} is identical to S_{gas}, but the particle's accessible free volume should be smaller than that in the gaseous state because the aqueous medium is much more crowded (**eqn.4.15**).

$$S_{liquid} = Nk_B \ln\frac{v_f}{\Lambda^3} + \frac{5}{2}Nk_B \qquad \textbf{(eqn.4.15)}$$

It is very difficult to evaluate the configurational integral of a particle in the mean field of its neighbours in the condensed phase, but the cell model allows v_f to be easily assessed by integrating over the potential function within a single cell. However, the concept of the communal entropy has been criticised as an artificial construct and in attempting to give it a formal definition, Kirkwood (1950) could not overcome the issues with this term [298,350,351].

Later SF theories proposed by Henchman removed the troublesome communal entropy by only permitting single-cell occupancy. Additionally, a novel method to estimate a particles T&R configurational volumes via a 6D anisotropic harmonic potential was developed [164,349,352,353]. In order to calculate T&R configurational volumes of a molecule in the mean field of its neighbours, averaged pairwise, force and "torque" constants between all the atoms in the cell are calculated and halved to prevent overcounting along the 6 principal axes (x, y, z, θ_x, θ_y, θ_z). Free and bound volumes are then used in conjunction with the cratic entropy to calculate T&R entropies for several protein-ligand binding interactions, including the T4 lysozyme mutant (T4LM) and MUP [164]. Both these systems are addressed in the results section as the former is a good example of a system that binds ligands in a single bound minimum (§4.3.2), whilst the latter is likely to be characterised by multiple binding minima (§4.3.3).

4.1.3. Entropy naming conventions:

There are many labels in the literature that describe the various subcomponents of the total global entropy. Depending on the partitioning method, certain labels have subtle distinctions e.g. conformational versus configurational as defined by Gilson et al (2007) [173,188]. Translational and rotational entropies are calculated by measuring momentum and position distributions, but in a classical treatment the momentum contribution cancels out when the difference between products and reactants is evaluated [303,304]. Hence, Gilson et al. (1997) proposed that a distinction should be made via use of the alternative terms positional and orientational entropy [304]. The scheme used in this chapter is tabulated in **Table.4.1**.

Entropy Component	Entropy Label	Bound	Free	Difference
Global ITC	Global	na	na	$T\Delta S_{Glo}$
Translational	Positional	$TS^{\circ}_{Po\bullet B}$	$TS^{\circ}_{Po\bullet F}$	$T\Delta S^{\circ}_{Po}$
Combined Rotational & Conformational	Orientational	$TS_{Or\bullet B}$	$TS_{Or\bullet F}$	$T\Delta S_{Or}$
• Principal Rotations	Rotational	$TS_{Ro\bullet B}$	$TS_{Ro\bullet F}$	$T\Delta S_{Ro}$
• Conformational	Internal	$TS_{In\bullet B}$	$TS_{In\bullet F}$	$T\Delta S_{In}$

Table.4.1. Entropy partitioning scheme used in this work with associated labels and symbols. Note that the orientational entropy includes both internal and external DOF.

Here, the method used to calculate the orientational binding entropy ($T\Delta S_{Or}$) implicitly includes a contribution from internal DOF ($T\Delta S_{In}$) in addition to that associated with the principal rotations ($T\Delta S_{Ro}$) of the molecule. This cannot be separated in the first instance but if the conformational entropy is obtained in a separate calculation (e.g. chapter 3.0), it should theoretically be possible to recover the entropic term associated with the principal rotations of the molecule via **eqn.4.16**.

$$T\Delta S_{Ro} = T\Delta S_{Or} - T\Delta S_{In} \qquad \textbf{(eqn.4.16)}$$

4.1.4. Objectives & Overview

4.1.4.1. Objectives

This chapter focuses on assessing the external entropies (positional & orientational) of ligands within n-alkanol and 3Z-olefin panels as the results from chapter 3.0 indicated that the structural differences between these compounds would provide the most interesting contrast. There are two key questions underlying this investigation that are intrinsically intertwined with one another:

Q1. Why does a relatively simple perturbation (i.e. introduction of a cis-3-4 double bond into unsaturated linear alcohols) result in improved global entropies of binding (**Fig.3.1)** of 3Z-olefins compared to n-alkanols? As the entropy contribution from internal DOF ($T\Delta S_{In}$) did not account for the large difference between the two panels, the external entropy contribution ($T\Delta S°_{Po}$ and $T\Delta S_{Or}$) will now be quantified. This will allow resolution of the following simple hypotheses:

> ***Hypothesis-1:*** *Extending ligand length by a single methylene group yields an enthalpic gain due to increased protein-ligand van der Waals contacts. However, a compensating entropic penalty entropic penalty of 5.4 kJ/mol is paid due to the addition of a rotor which inevitably becomes restrained on binding* [36]*. Only the entropic contribution is considered in this chapter.*
>
> ***Hypothesis-2:*** *Disabling a rotor via introduction of a double bond avoids the entropic penalty on binding as this debt has been paid during the process of chemical synthesis* [36,178].
>
> ***Hypothesis-3:*** *Pre-organisation of the ligand (i.e. 3Z-olefins) so that the structure complements the shape of the binding site ameliorates entropic penalties by virtue of less strain being imposed on the ligand* [174,177–179].

Q2. Beyond obtaining measurements that quantify the ligand contribution to the global entropy of binding, it is also desirable to characterise the dynamics of bound compounds because this affords predictive capabilities that can guide rational drug design. As discussed, there are many examples of promiscuous proteins (§4.1.1), and techniques that illuminate the relationship between structure and dynamics can offer insights into how inhibitor modification affects specificity and efficacy. Hence, the question: "How and why do structural modifications affect the dynamics of bound ligands within the context of the binding cavity?"

4.1.4.2. Chapter overview

* **Question one** is answered by developing two new methods that quantify $T\Delta S^{\circ}_{Po}$ and $T\Delta S_{Or}$:

1. Calculation of $T\Delta S^{\circ}_{Po}$ is controversial and there is much debate in the literature regarding the correct technique. As all current methods usually involve the (quasi-)harmonic approximation in some manner, the method presented in this work eschews any reliance on a functional form to describe the underlying distribution from which the positional entropy is calculated. As the protein-ligand bound state is often characterised by multi-modal distributions the approach allows more accurate estimates of $T\Delta S^{\circ}_{Po}$.

2. Defining the principal rotations of flexible molecules (such as linear alcohols) is beset with complications because it is impossible to define the principal rotational axes for molecules that drastically change shape with time. The complexity of this problem has resulted in researchers relying on assessing the conformational entropy, whilst neglecting potentially larger contributions from molecular rotations. This issue is particularly pronounced for promiscuous binding interactions, wherein the ligand possesses multiple bound minima. Hence, an approach that deals with these issues was developed and tested. The method obtains entropies for flexible compounds by calculating entropies on a per-bond basis and then summing these values to provide a 1st order estimate for the total $T\Delta S_{Or}$ of binding. As per-bond entropies afford greater insights into the factors governing ligand dynamics, this allows an avenue through which Q2 can be addressed.

* **Question two** addresses the "how and the why" underlying the different global entropic signatures of the two panels tested and is more difficult to explain because the answer is highly dependent on the relationship between the architecture of the binding cavity and the dynamics of the ligand; the latter being affected by its size and structure. In order to elucidate the underlying physical rationale for dynamic differences between bound compounds, the results are split into four main sections that consecutively deal with:

1. **Positional entropy:** Provides data on the extent of ligand residual motion within the cavity and provides evidence of significant differences in positional distributions that are correlated with ligand size and structure. The proposed method is iteratively validated by:

    i. Randomly generated points: to calibrate bin volumes

    ii. Benzene bound to T4LM: to test the method against a ligand that binds within a single minimum. This system also allows comparison with a variety of other methods used in the literature.

    iii. 3Z-olefin and n-alkanol ligands binding to MUP: to test the method with ligands that bind within multiple minima. An additional level of rigour is instituted by checking whether $T\Delta S^{\circ}_{Po}$ trends across panels are captured. This component contributes significantly to the global entropy.

2. **Protein-ligand H-bonding:** In the predominantly apolar binding pocket, the single polar moiety possessed by these linear alcohols plays a crucial role in determining the positions occupied by the molecule and the orientations that its constituent bonds can adopt. Again, clear differences in H-bond patterns are discerned between both panels and a correlation with ligand size is uncovered.

3. **Orientational entropy:** There are systematic differences in the trends exhibited by 3Z-olefins compared to n-alkanols and this component also contributes significantly to the global entropy. After exploring the preceding two points the rationale behind differential patterns in per-bond $T\Delta S_{Or}$ is uncovered. Furthermore, on solving the final piece in the puzzle, the "how and the why" regarding differential positional entropies and H-bonding is retroactively comprehended.

4. **Convergence issues and sampling:** This is addressed at appropriate points throughout the chapter as this is a critical factor associated with the evaluation of any *in silico* calculation. Nonetheless, a more detailed section near the end analyses the implications of convergence in greater detail and this allows a comparison of the calculated total ligand entropy (internal + external contributions) to global ITC values. Finally, predictions regarding the total system decomposition (ligand + protein + solvent contributions) are made.

As the relationships between the factors described in these four sections are *intrinsically intertwined*, it is difficult to explain one without referring to the others. Nevertheless, the order chosen best presents the detailed "how and the why" for the difference in global entropic binding signatures between 3Z-olefin and n-alkanol ligands.

4.2.0. Methods

4.2.1. Simulation setup, sampling times & protocols.

4.2.1.1. ff03 force field

In order to investigate ligand positional and orientational contributions to the global entropic binding signature of MUP, the MD simulations obtained in chapter 3.0 (n-alkanols, 3Z-olefins and apo receptor only) were extended from 0.6 µs/ligand to 1.2 µs/ligand. As flexible ligands possessed significant residual motion when bound, they are likely to occupy multiple bound minima; a factor likely to be crucially dependent on sampling. Five simulations had to be completely rerun because the ligand began to escape from the pocket on extending running time (**Table.4.2**). All new and extended simulations were run using the AMBER 12 GPU code.

Bound Ligand	Repeat
hep	01
hep	05
3c8	05
non	03
non	04

Table.4.2. List of simulations obtained in chapter 3.0 that were discarded and rerun from the beginning.

All data was obtained using the protocol and parameters detailed in §3.2.3. Simulations were rechecked for stability via RMSD and the monitoring of other variables such as energy, temperature, density, etc.

4.2.1.2. ff99Sb-nmr force field

1. MUP: In order to check results obtained using the ff03 force field (§4.3.3), six 100 ns simulations were set up using exactly the same settings and protocols used for the ff03 based simulations as detailed in §3.2.3. The only difference being the use of the ff99SB-nmr force field. This was applied to the MUP-hep complex and apo receptor only.

2. T4 lysozyme mutant: The binding of benzene to this protein was assessed via six 200 ns simulations of both the complex and apo receptor (§4.3.3). The structure used for apo and holo simulations was obtained from the PDB (181L) [354]. TIP3P was used for the solvent model and crystallographic waters were retained. Benzene was parameterised using the Gaussian 09 and the R.E.D III suite of tools as described in §2.2.2 and §3.2.1. Apart from this, identical settings to the ff03 simulations were used.

4.2.1.3. Positional entropy: Free simulations

In order to obtain the translational entropy difference between bound and free states, a new set of simulations were run in which the ligand sampled a free volume corresponding to the 1 M standard state, 1,660 Å^3. A dummy atom incapable of reacting with its surroundings was situated at the centre of the simulation box and restrained in place with a strong harmonic potential (100 kcal/mol). The ligand was positioned alongside this atom and solvated with TIP3P waters using a box size of 12.0 Å. The ligand was constrained within the required volume by restraining all heavy atoms to remain within 7.345 Å of the dummy atom via a flat-welled parabolic restraint (20 kcal) with steep sides. Production runs were carried out using NPT conditions. Force field and simulation lengths used for various systems are detailed in §4.2.1.1-2.

4.2.1.4. Orientational entropy: Free simulations

To avoid biasing bond rotations and conformations of the free ligand, the protocol used for unrestrained simulations in §3.2.3 was utilised. Force field and simulation lengths used for various systems are detailed in §4.2.1.1-2.

4.2.2. RMS fitting of complexed simulation trajectories

The protein binding pocket is used as the frame of reference against which relative translational and rotational movements of bound ligands are quantified. The overall translations and rotations of the protein in solution are removed by RMS fitting backbone atoms (C, C_α, O and N) to a reference structure using ptraj. The reference structure was created by averaging all frames from every simulation for that system. e.g. in the case of bound MUP, protein backbone atoms were averaged over all frames obtained from every n-alkanol and 3Z-olefin simulation. (1,200,000 frames per complex x 8 complexes x 6 repeats = 57,600,000 total frames). The averaged reference structure was then manually reorientated using UCSF Chimera so that its final position matched that depicted in the front view (**Fig.4.2**):

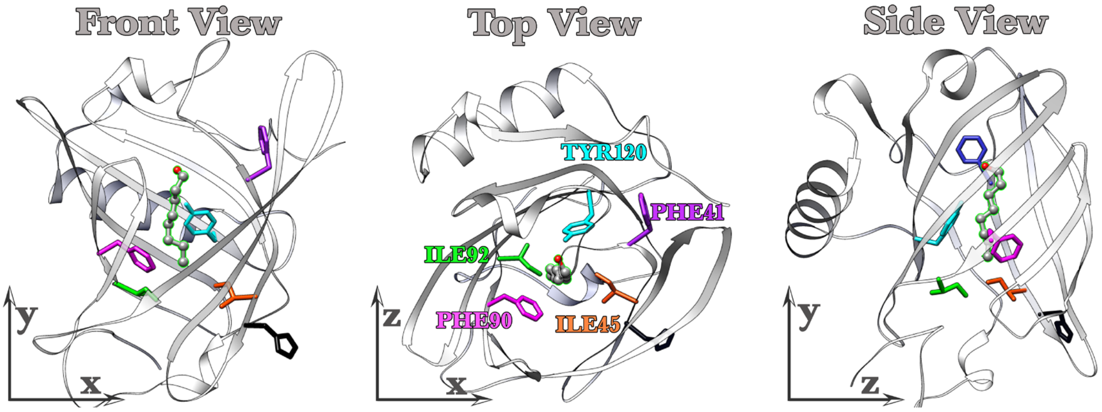


Fig.4.2. The protein was RMS fitted to an averaged structure as described in the methods so that the ligand's COM distributions depicted in **Fig.4.11-14** could be visualised. As the averaged structure has a known position and orientation, it can be rotated so that the 2D projections through the three principal planes display COM densities in relation to the architectural features of the calyx. Residues used for orientation are colour coded as per the labels in the centre panel.

This procedure affords two advantages. Firstly, a better quality RMS fit is obtained when fitting the protein to the averaged reference structure compared to fitting to the first frame (**Fig.4.3**). Secondly, as all simulations have been fitted to the same reference structure, COM density plots for different bound ligands can be easily compared and contrasted.

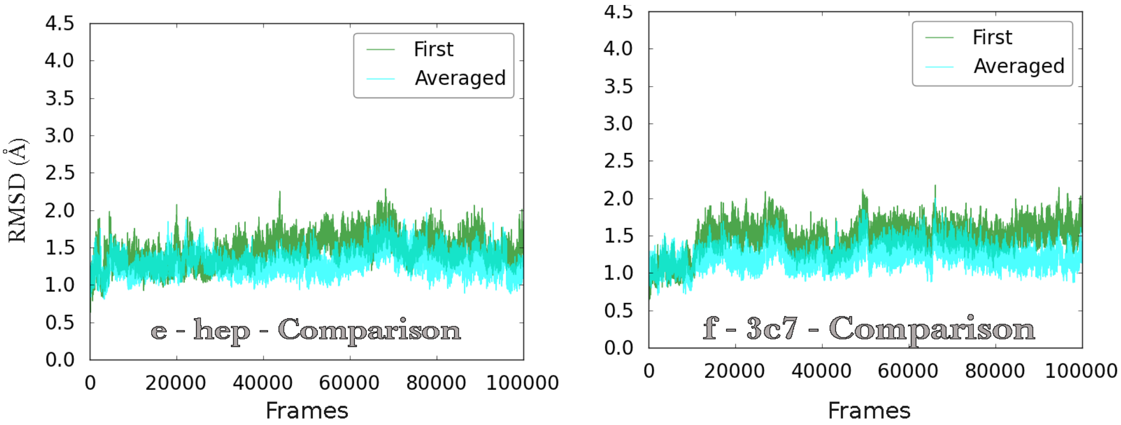


Fig.4.3. The graphs compares the RMSD obtained by fitting a 100 ns simulation to the averaged structure to that obtained via fitting to the first structure in the trajectory. Fitting to the averaged structure (cyan) provides a better alignment throughout the whole trajectory as opposed to fitting to just the first frame (green).

Reference frame "jitter" assessment: Subsequent to the RMS fit, the protein is localised in terms of both position and orientation in all trajectory frames. However, the bound ligand retains full translational and orientational motion because its coordinates have been shifted relative to the protein backbone atoms. As the motions of the bound ligand are defined in relation to RMS fitted backbone of the protein, the accuracy of every COM position is dependent on the frame of reference remaining static from snapshot to snapshot. The COM displacements of the protein backbone (N, C_α, C and O) were assessed with a view to calculating the entropic cost to the quality of the fit over the full 1.2 μs concatenated trajectories. However, the fit in the case of both T4LM and MUP systems was so good that the COM distribution was too small to measure using the positional entropy method proposed in §4.2.4.2. Hence, a rough estimate of the volume used by the protein's COM distribution was generated by multiplying distances obtained from the minimum and maximum of the distribution across each orthogonal axis.

4.2.3. Mean shift clustering

The clustering of trajectories was accomplished with python scripts that utilised the mean shift algorithm implemented in the scikit-learn machine learning library [355]. Mean shift clustering is non-parametric method that identifies densely populated regions of an n-dimensional probability distribution function by performing a gradient assent until convergence is reached. A bandwidth parameter defines the radius of a kernel whose initial position is randomly assigned. The COM is calculated from all the data points that fall within this kernel and the direction of the mean shift vector is ascertained by calculating the density gradient. The kernel is then moved along the vector towards the volume of greatest density and the process is repeated until convergence is attained. The minimum number of points considered for inclusion in the kernel also influences cluster identification. In order to obtain optimally positioned clusters, iterative testing demonstrated that 3,500 points in conjunction with a bandwidth of 0.5 were the best values for the key parameters with respect to the size of the dataset.

Averaged structures representative of ligand and key amino acid residue conformations were obtained by isolating the frames making up each cluster using python scripts and ptraj.

4.2.4. The positional entropy

4.2.4.1. Depiction of the bound ligand's COM motion

In order to get visual representations of the translational displacements of the bound ligand in relation to key orientating residues in the binding pocket, Cartesian coordinates of the groups of interest were stripped from the concatenated, RMS fitted trajectory using ptraj and python scripts. The reference residues extracted were PHE41, HIS46, ILE45, PHE90, ILE92 and TYR120. A further python script calculated the centre of mass of all extracted groups. The function hexbin from the python matplotlib library was utilised to generate three 2D COM density plots, so that the 3D distributions are in effect projected onto three mutually perpendicular planes. The three COM projections (**Fig.4.11-14)** correspond to the views obtained from the RMS fitted protein (**Fig.4.2**).

4.2.4.2. Positional entropy calculation

As detailed in §4.1.2.4, the positional entropy is calculated using **eqn.4.12** and measures the change from the free ligand's effective volume (1,660 Å^3) at the 1 M standard state to the smaller effective volume the ligand can access after binding. The primary problem associated with the calculation of the positional entropy is accurate quantification of the *effective* volume. It would not be accurate to treat a 3D cloud of points generated from COM displacements as a hard sphere because points near the edges are expected to have lower density than those at the centre of a given minimum. Traditionally, some form of (quasi-)harmonic approximation has been used to model this. However, the probabilistic approach proposed in this chapter to assess the density of states offers several advantages when coupled with adequate sampling. The method is known as 3D histogramming (3Dh) and uses a 3D grid to discretise 3D space into smaller cubes (**Fig.4.4**). The positional entropy can then be calculated via the Shannon entropy equation. The unitless result is converted into the thermodynamic entropy via the introduction of the temperature and the gas constant into the equation (**eqn.4.17**). P$_i$ is calculated by dividing the number of points in each smaller cube volume by the total number of points.

$$TS^{\circ}_{Po} = -RT\sum_{i=1}^{bins} p_i \ln p_i \qquad \textbf{(eqn.4.17)}$$

TΔS°$_{Po}$ is then calculated by taking the difference between endpoint states (**eqn.4.18**).

$$T\Delta S^{\circ}_{Po} = TS_{Po\bullet B} - TS_{Po\bullet F} \qquad \textbf{(eqn.4.18)}$$

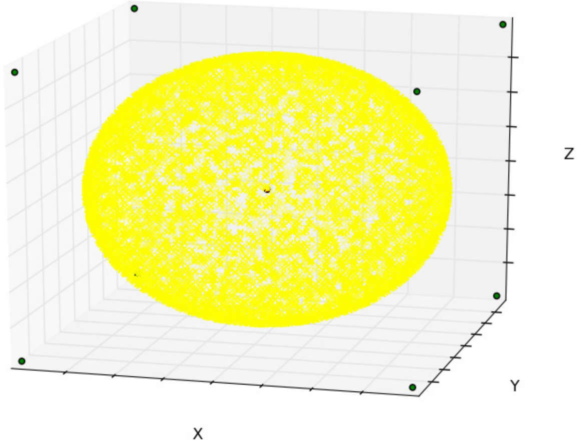


Fig.4.4. Example of 3D binning with eight bins along each orthogonal axis. As an illustrative example, a randomly generated cloud of points (yellow) was created to fill a volume of 1,660 Å^3. §4.3.1 contains details regarding parametrising the optimal number of bins. The grid discretising the 3D volume can be drawn from the frame edges.

As detailed in Chapter 3.0, entropy calculations are dependent on the choice of bin size or in this case, cube volume. In order to establish the correct cube volume a simple toy problem is analysed in the first results section (§4.3.1). This is followed by assessing the method with T4LM, which chiefly binds benzene within a single minimum (§4.3.2), and finally MUP; a promiscuous protein that is expected to allow ligands to bind within multiple minima (§4.3.3).

4.2.5 The orientational entropy

4.2.5.1. Bond vector isolation and setup

In the case of the bound simulations, ligand coordinates were stripped from RMS fitted trajectories and a python script was used to isolate the coordinates of individual bond vectors defined by pairs of heavy atoms. The Cartesian coordinates of these isolated bonds were written into separate files after removing translational motions by moving each vector to the origin. This procedure preserves the orientation of the bond vectors which are clustered around the origin. The density of states explored by each bond can be mapped onto the surface of a sphere. This can be visualised using the technique described in §4.2.5.2 and the per-bond orientational entropies can be calculated using the method detailed in §4.2.5.3. As the bound trajectories have already been fitted to a common frame of reference in the first instance, direct visual comparisons can be made between the densities of orientational states of analogous bond vectors obtained from the different simulations. Note that orientational motions are less susceptible to shifts in the fitting procedure compared to positional displacements.

In the case of the free ligand, the ptraj command autoimage was used to remove translational motion from the unfitted trajectory. Then the same procedures used for the bound ligand were applied to obtain visualisations and entropies.

4.2.5.2. Depiction of bond vector motion

Dr G. Thompson kindly provided a python script that iteratively subdivides the twenty faces of an icosahedron into triangular bins of equal area. Data points obtained from the dynamic movement of each vector are clustered within the triangular bins formed by this process using a hierarchical ray casting method (**Fig.4.6**). Initially, points are coarsely apportioned between a few large bins. Subsequently, each bin is subdivided into four smaller bins and the points are more finely reapportioned into the new bins. This procedure is iteratively repeated until the desired bin size is obtained and this considerably saves on calculation time. In big O notation, the algorithmic efficiency of a simple linear search would be $O(4^n)$, whilst the hierarchical search is 4,000 times faster as it is $O(4n)$ [356].

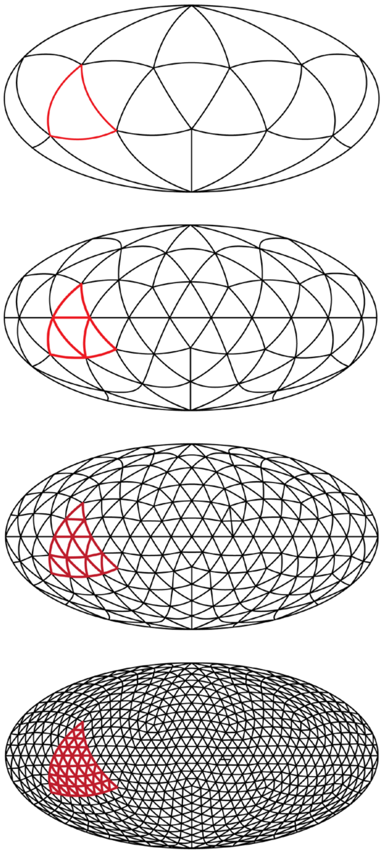


Fig.4.5. Hammer projections of the iterative subdivision of an icosahedron that allows visualisation of orientational densities of ligand bond vectors. Triangles marked in red illustrate the process used to hierarchically cluster the data points so as to increase the efficiency of the process. Panels (top to bottom) contain 20, 20 x 4, 20 x 4^2 and 20 x 4^3 bins respectively.

Differences in orientational density of bond vectors can then be visualised in 2D using the equiareal Hammer projection from the matplotlib Basemap toolkit. The problem of correctly preserving the shape and area of the Earth has resulted in a variety of map projections. Early projections were limited by the perspective of the viewer and could only depict a single hemisphere at a time. More advanced techniques such as the Hammer projection double longitudinal values so that the entire globe can be viewed from a single perspective. The resulting 2:1 ratio is equiareal and minimally compromises distance accuracy. Overall shape distortion is moderate, whilst the centre is free from artefacts [357–359]. Note that a distinct procedure is used to calculate the per-bond orientational entropy §4.2.5.3.

4.2.5.3. Orientational entropy calculation

Per-bond entropies: The orientational entropies of all bond vectors were calculated using a modified version of the program “order” by Dr C MacRaild and Dr G Thompson [360]. The program assesses the density of points created by the terminus of each bond vector as it moves across the surface of a sphere by binning the data via the method discussed below. Note that orientational entropies calculated from the bond vectors isolated using the procedure described in §4.2.5.1, contain inseparable contributions from both principal rotations and internal DOF. However, the pure rotational component can be isolated via **eqn.4.16**.

If a sphere is divided by two parallel planes separated by distance (h), the portion between them is known as the zone or frustum which has an area of 2πrh. If subdivided into many such horizontal zones, frustums near the equator would have a larger surface area than those at the poles (**Fig.4.6**). In spherical coordinates, a fixed variable known as the θ_{step} specifies the number of bins along an arc following the zy-plane. Thus, in order to ensure the accuracy of the entropy calculation, all bins must be equiareal and the number of bins within each frustum is allowed to vary along the xy-plane. This results in a different value for δφ for each frustum and the appropriate number of φ_{steps} is calculated using **eqn.19-22**. In practice the value for θ_{step} will be very small - typically 1° - 2°.

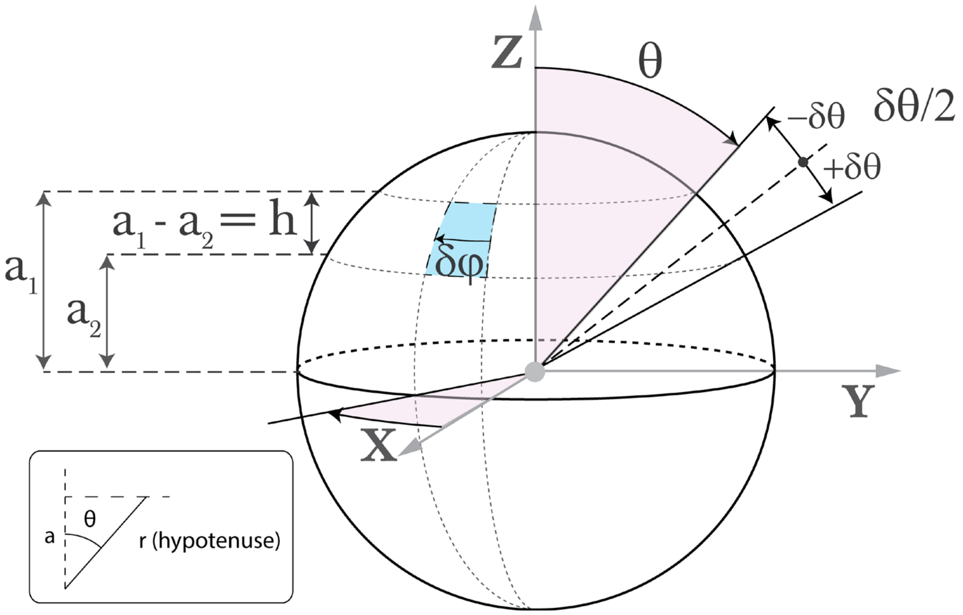


Fig.4.6. Depiction of spherical coordinates using physics based notation. See text for details.

$$\frac{\cos_{\pm\delta\theta/2}}{2} = \frac{a_{\pm\delta\theta/2}}{r}$$
$$a_{\pm\delta\theta/2} = r\cos\theta_{\pm\delta\theta/2} \qquad \textbf{(eqn.4.19)}$$

Therefore,

$$h = a_{+\delta\theta/2} - a_{-\delta\theta/2}$$
$$h = r\cos(\theta + \tfrac{\delta\theta}{2}) - r\cos(\theta - \tfrac{\delta\theta}{2}) \qquad \textbf{(eqn.4.20)}$$
$$h = r[\cos(\theta + \tfrac{\delta\theta}{2}) - \cos(\theta - \tfrac{\delta\theta}{2})]$$

Using the trigonometric identity in **eqn 4.20**, h can be calculated using spherical coordinates.

$$\sin a \sin b = \frac{\cos(a-b) - \cos(a+b)}{2}$$
$$\frac{2\sin\theta\sin\varphi}{2} = \cos(\theta - \tfrac{\delta\theta}{2}) - \cos(\theta + \tfrac{\delta\theta}{2}) \qquad \textbf{(eqn.4.21)}$$
$$\therefore$$
$$h = 2r\sin\theta\sin\tfrac{\delta\theta}{2}$$

As the area of the frustum is $4\pi r^2 \sin\theta \sin \delta\theta/2$, the number of φ_{steps} for each frustum can be calculated using eqn4.22.

$$\varphi_{steps} = \frac{\sin\theta}{\sin^{\delta\theta}\!/_{2}} \qquad \textbf{(eqn.4.22)}$$

The optimal bin area was assessed by running calculations with different increments for the θ_{step}. Orientational entropy differences were invariant after a hundred θ_{steps}, so a θ_{step} corresponding to 180 (i.e. 2°) was selected (**Fig.4.7**).

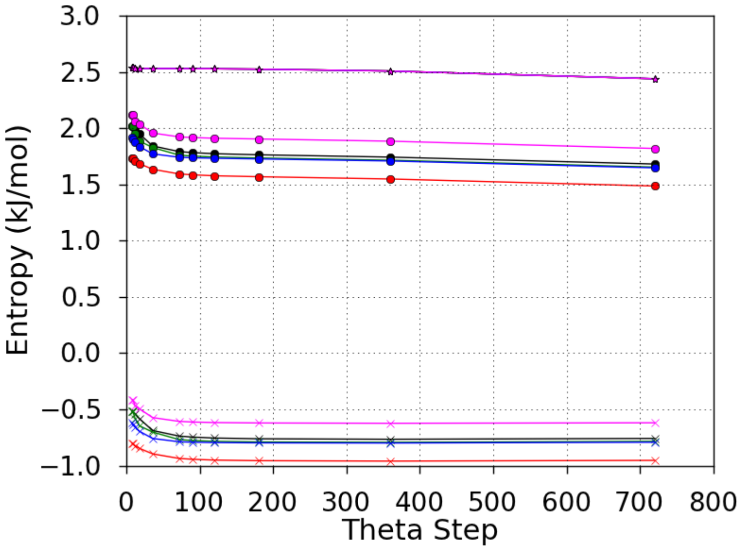


Fig.4.7. Per-bond orientational entropies plotted versus θ_{step} for all bond vectors isolated from bound nonan-1-ol. The calculated entropy differences are invariant when using θ_{step} > ~100 steps.

Calculating total ligand $T\Delta S_{\text{Or}}$ contributions: Subsequent to binning the data, order parameters were calculated on a per-bond basis for free and bound states using the method proposed by Best and Vendrusculuo (2004) [361]. The orientational entropy was calculated using **eqn.4.23** as described by Yang and Kay (1996) where q represents the molecular coordinates and v the area of a given bin [96]. Entropies were calculated using units of K_B.

$$S = -\int_{v} P(\mathrm{q})\ln\{\mathrm{p(q)}\}\mathrm{dv} \qquad \text{(eqn.4.23)}$$

Per-bond orientational entropies can then be summed to get a 1st order estimate and the total value obtained for the free state can be subtracted from the bound state to generate the total ligand $T\Delta S_{\text{Or}}$ in a similar way to the positional entropy (**eqn.4.18**). All per-bond orientational entropy values are given in units of k_B, whilst summed entropy totals are reported in units of kJ/mol at 300 K.

4.2.5. Hydrogen bond analysis

Hydrogen bond (H-bond) analyses on the MD generated ensemble of structures were accomplished using the ptraj module from AMBER, and custom python scripts. Analysis was performed on all 1.2 x 10^6 frames for all ligands from both panels.

4.3.0. Results & Discussion

4.3.1. Positional entropy: a toy problem to calibrate bin volumes

As described in §4.2.4.2, 3D histogramming (3Dh) creates a cubic frame centred on the COM distribution, so that space can be discretised in 3-dimensions by gridding each orthogonal axis to create smaller cubes. Technically, this outer frame can possess any volume that is greater than the standard volume (1,660 Å^3) as cubes without data points are excluded from the calculation due to the logarithmic relationship in **eqn.4.17**. In practice, it was found that the frame edge lengths could equal the diameter (14.49 Å) of standard volume which is represented by a spherical distribution in the free state. However, this can create issues when evaluating the bound state as COM distributions can be non-spherical. Whilst the effective volume occupied by bound ligands in the cavity is smaller than V°, the range of points along some axes can exceed the diameter of the sphere describing the V°. To allow for this, an additional 8 Å were added to each frame edge to yield a total length of 22.49 Å. As equivolume cubes are created by equally subdividing the three principal frame edges into 1D bins, the optimal number of bins is dependent on frame edge length and this would have to be empirically calibrated.

In order to facilitate this, two datasets were created:

1. **Analytical:** Theoretical entropy values were generated analytically using **eqn.4.12**. The free volume corresponded to 1,660 Å^3 and a range of bound volumes (0.5 to 1650 Å^3) were input into the equation to systematically represent various degrees of positional restriction smaller than V°.

2. **Synthetic:** "Synthetic" datasets were then numerically generated for the same range of volumes using 1.2x 10^6 data points scattered in a 3D Gaussian distribution. $T\Delta S^{\circ}_{Po}$ values were then calculated via 3Dh (**eqn.4.17-18**).

The reasoning behind the use of analytical and synthetic datasets is that the former easily provides a range of idealised reference $T\Delta S^{\circ}_{Po}$ values, whilst the latter generates 3D COM distributions that possess the practical limitations associated with calculating the *effective* volume (§4.2.4.2). Thus, the entropy calculation can be iteratively run on the synthetic datasets whilst varying the number of bins along each axis. As synthetic datasets utilise a Gaussian distribution, they have the potential to approximate the idealised volumes used in the analytical calculation. So, on approaching the optimal bin number, $T\Delta S^{\circ}_{Po}$ values generated by the synthetic trials should converge to reference $T\Delta S^{\circ}_{Po}$ values yielded by the analytical calculation.

Sphere Volume 1660 Å^3 (1.2x10^6 points)						
Bin no	**Cube Vol (Å^3)**	**Avg points/Cell**		**Bin no**	**Cube Vol (Å^3)**	**Avg points/Cell**
5	62.504	9600.00		50	0.063	9.60
10	7.813	1200.00		55	0.047	7.21
15	2.315	355.56		60	0.036	5.56
20	0.977	150.00		65	0.028	4.37
25	0.500	76.80		70	0.023	3.50
30	0.289	44.44		75	0.019	2.84
35	0.182	27.99		80	0.015	2.34
40	0.122	18.75		85	0.013	1.95
45	0.086	13.17				

Table.4.3. For the example volume of 1,660 Å^3, changing the number of bins along each principal frame edge alters the cell volume and consequently, the average number of data points that fall into it.

The frame edges were subdivided into bins in iterative synthetic trials that varied the number of bins through a range of 5 to 85. The upper and lower limits corresponded to discretised cube volumes of 62.5 Å^3 to 0.013 Å^3 respectively (**Table.4.3**). The results are plotted in **Fig.4.8.** It is difficult to assess the optimal number of bins by merely examining $TS^{\circ}_{Po{\bullet}F}$ or $TS^{\circ}_{Po{\bullet}B}$ as the values never converge (**Fig.4.8.a**). However, on considering relative entropy differences ($T\Delta S^{\circ}_{Po}$), convergence can be easily gauged (**Fig.4.8.b**).

The larger the volume under consideration, the more rapid the entropy convergence and thus, fewer bins are required to accurately quantify the entropy of a large spread of points (**Fig.4.8.b**). For smaller *effective* volumes (< 10.0 Å^3), convergence only occurs when the number of bins is > ~45 to 50. This is because the cube volume is inversely proportional to the number of bins, and a small number of bins results in large cube volumes that coarsely represent the underlying distribution. As *smaller volumes tend to have highly localised, dense distributions*, the data points tend to occupy one bin rather than being distributed across several. This accounts for the jagged fluctuations seen at small volumes (< ~10 Å^3) when bins are < ~45 to 50. Conversely, the representation becomes oversharpened above the optimal number of bins. The number of bins could be continually increased with a concomitant change in the entropy. However, this change becomes very small after the optimal point and over binning can result in artefacts being introduced into the data representation. Despite the large size of the synthetic dataset (1.2 x 10^6 points), local variations in the structure of the distribution result in synthetic

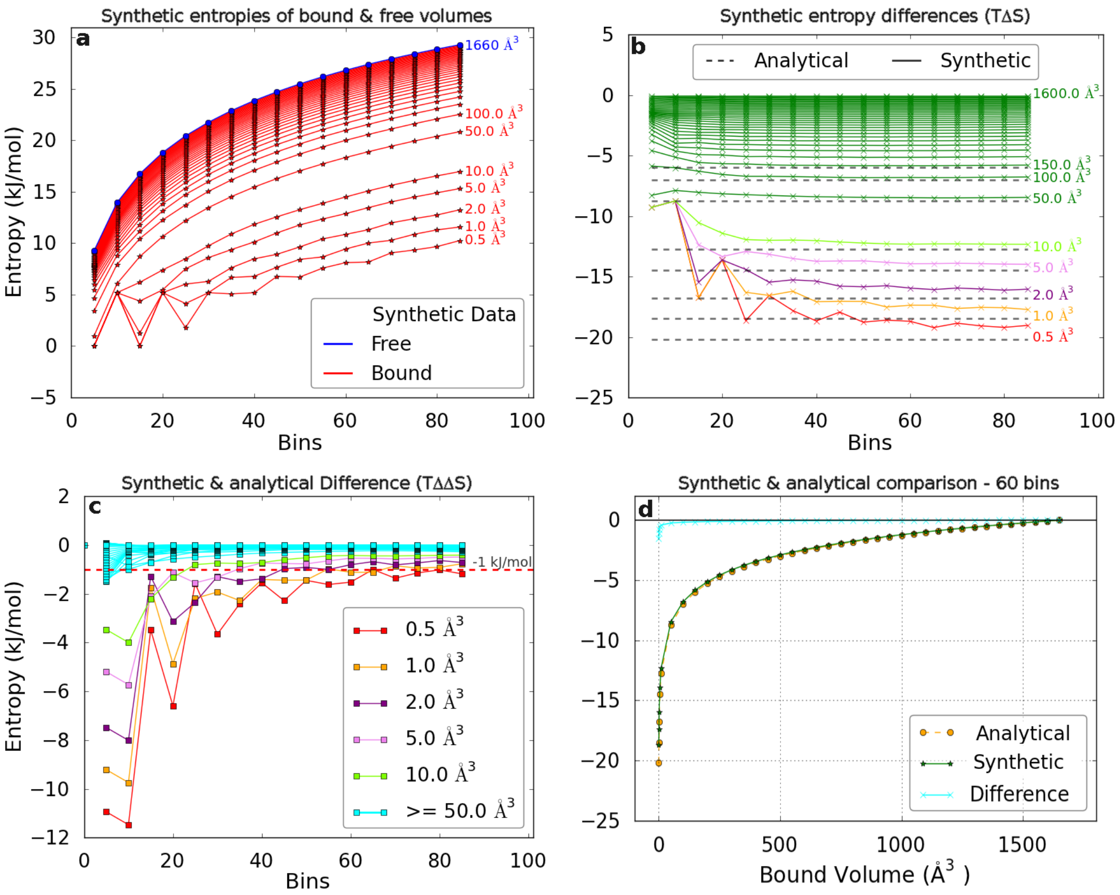


Fig.4.8. (a) TS°_{Po} values generated from synthetic datasets via 3Dh on varying the number of bins. Smaller bound volumes are labelled in red, whilst the largest free volume (V°) is coloured blue. **(b)** $T\Delta S^{\circ}_{Po}$ values obtained for the range of synthetic bound volumes. Grey dashed lines mark reference $T\Delta S^{\circ}_{Po}$ values obtained analytically for bound volumes less than 150 Å^3. **(c)** The difference between $T\Delta S^{\circ}_{Po}$ values obtained from synthetic and the analytical calculations i.e. $T\Delta\Delta S^{\circ}_{Po}$. The red dashed line marks a 1 kJ/mol limit to indicate the negligible amount of error .**(d)** Synthetic $T\Delta S^{\circ}_{Po}$ values obtained for all bound volumes (with 60 bins) compared to the analytical $T\Delta S^{\circ}_{Po}$ values. The greatest difference between the two ($T\Delta\Delta S^{\circ}_{Po}$) chiefly occurs within the smallest bound volumes.

values never exactly matching those of the analytical values.

The difference between entropy values calculated from the synthetic data to that obtained analytically ($T\Delta\Delta S^{\circ}_{Po}$) is plotted in **Fig.4.8.c** for all bound volumes tested. The resolution necessary to differentiate entropies obtained for bound volumes < 2.0 Å^3 is impaired. However, all the other bound volumes possess $T\Delta\Delta S^{\circ}_{Po}$ negligible differences of less than 1.0 kJ/mol when the number of bins is > 55 to 65. As it is unlikely that most non-covalently bound compounds would be localised to extremely small volumes < 2.0 Å^3, this should not present an issue.

On considering the slope of the lines for the smallest reliable volumes (10, 5 and 2 Å), the optimal number of bins for the toy problem was deemed to be 60 along each frame edge (**Fig.4.8.b**). Finally, synthetically calculated $T\Delta S^{\circ}_{Po}$ values are compared against analytically generated TΔS°_{Po} values for the selected number of bins (**Fig.4.8.d**). The optimal bin number will have to be recalibrated for different systems as changes in the shape of the underlying distribution may affect this variable; an eventuality that 3Dh is particularly adept at handling.

The chief benefits of 3Dh are threefold:

1. 3Dh is quick and simple to execute in contrast to most other methods and the associated CPU and memory requirements are very small, even when evaluating massive datasets. Furthermore, it is flexible as the frame size can be altered to suit systems that possess a variety of shapes

2. Methods that generate external entropies by summing (or taking the product) of 3 entropy calculations [164,339] for each principal axis generate larger values. This is because they do not implicitly account for correlations between DOF which are not independent. 3Dh avoids this issue by utilising a single entropy calculation that simultaneously accounts for all 3 DOF.

3. Methods based on the harmonic approximation typically assume a single bound minimum and excessively smooth the multimodal distribution commonly observed in the bound state. Calculating the effective volume via histogram assessment of the density of states is likely to capture the complicated topography created by multiple minima (§4.3.3), or indeed the subtle variations associated with a single minimum (§4.3.2). This is because 3Dh is unhampered by any assumptions regarding the underlying functional form of the distribution, and is consequently unaffected by anharmonicities typically encountered in bound COM distributions [203,324].

The next two sections test 3Dh on T4LM and MUP; macromolecular systems that bind ligands within single and multiple minima respectively.

4.3.2. T4LM positional entropy: a single bound minimum

The T4 lysozyme mutant (T4LM) is a particularly interesting system in which the results obtained from various methods can be compared and contrasted. The mutant (C54T, C97A, L99A) was created to demonstrate that proteins could be engineered to bind specific ligands, and any structural changes accompanying the modification could be stabilised by the protein's intrinsic flexibility [362]. The buried apolar cavity measures 150 Å^3 and its ability to bind small hydrophobic ligands (such as substituted benzene analogues) have been extensively studied by ITC and *in silico* investigations [252,253,340,354,363]. In this case, it is fortunate that a number of estimates have been made regarding T&R

entropy losses on binding benzene. ITC measurements for this binding interaction provide global values of: ΔG° = -21.8 kJ/mol; ΔH° = -26.4 kJ/mol and TΔS° = -4.6 kJ/mol [252]. The study analysed 16 compounds and suggested that the tight fit offered by the cavity was likely to result in a highly unfavourable $T\Delta S°_{Po+Or}$ that was not compensated by a favourable desolvation contribution. Thus, most studies to date have assumed benzene possesses a single, highly localised bound minimum. As protein reorganisation was also observed in various complexed crystal structures, there is a possibility that protein flexibility might contribute favourably to the overall global entropy of binding and ameliorate unfavourable $T\Delta S°_{Po+Or}$ [252,354].

Note that author abbreviations used for methods described in various papers are listed in **Table 4.4**.

4.3.2.1 Flexibility & timescales

In order to demonstrate the link between timescales and *in silico* entropy estimates, only the translational component is considered in the remainder of this section. The T4LM system was extensively benchmarked by Carlsson and Aqvist (2005) using Schlitter's harmonic approximation, the QHA, and two simpler functional forms that assume the effective bound volume can be described by a Gaussian or Uniform distribution [250]. The F&J approach (§4.1.2.6) was also tested, but $T\Delta S°_{Tr}$ was found to exceed the entropy calculated for the free state, and was thus deemed unphysical. An interesting aspect of the analysis performed on T4LM is the impact of protein flexibility on the entropy calculation. Hermans and Wang (1997) had systematically tested a method known as the restrain release (RR) approach [340]. Benzene was restrained in the binding site and slow-growth thermodynamic integration was used to calculate the free energy of binding as positional restraints were gradually removed. This allowed an estimate of $T\Delta S°_{Po+Or}$ to be made, but complexities in the method (associated with alchemical transmutation) meant that a relatively static protein structure had to be utilised, and to ensure convergence, benzene was actively discouraged from exploring alternative binding modes. $T\Delta S°_{Po+Or}$ estimates were obtained for several simulations in which the protein was restrained to different extents and $T\Delta S°_{Po}$ values are tabulated in **Table.4.4**. Despite high protein-ligand shape complementarity, the different restraint schemes indicated that the binding site retained significant flexibility. Hence, C&A also tested various methods by comparing the results obtained from unrestrained simulations to ones that imposed a 50 kCal/mol Å^2 restraint on all protein atoms [250,340]. Restraints dramatically reduce protein flexibility and thus limit ligand access to alternative binding modes, and $T\Delta S°_{Po}$ is penalised by ~4.5 kJ/mol compared to the unrestrained simulations (**Table.4.4**). This should be considered unsurprising as crystallographic structures indicated a differential protein dynamic response on binding 9 different ligands, with the largest changes occurring in helix F (**Fig.4.9.g**). Moreover, the propensity for protein reorganisation meant that, some closely related ligand analogues with minimal structural differences possessed significantly different binding modes [354]. Hence, protein dynamics is likely to promote cavity adaptability and the motion of shifting sidechains would result in ligand accessible boundaries being in a state of flux. Thus, residual ligand motion is likely to be promoted to a greater extent than that suggested by the averaged crystal structure pose.

Other simulations on the ultra-tight binding biotin-streptavadin complex also demonstrated that biotin could access multiple binding modes that possessed plausible, alternative H-bond interactions to that observed in the crystal structure. The root cause was also determined to be due to protein flexibility. Furthermore, as solvent friction and caging effects in explicit water simulations resulted in (~ x10 times) slower protein conformational exploration compared to implicit models, it is likely that very short simulations underestimate the effective bound volume [240,364]. Despite highly favourable enthalpic interactions, biotin possessed bound translational amplitudes of 4-6 Å; a value significantly higher than the F&J crystal structure based estimate (0.25 Å) [240].

			TΔS$_{Po}$ (kJ/mol) at 300K			
	Publication	**Abbrv**	**Restr**	**Unrestr**	**Diff**	**Sim Time(ns)**
	Carlsson & Aquist (2005):	C&A				
MF	• Finkelstein & Janin (1989)	F&J	-32.3	-	-	na
	• Hermans & Wang (1997)		-25.5	-21.3	-4.2	~0.2
	• Schlitter		-22.0	-17.2	-4.8	2.0
	• QHA		-22.1	-17.2	-4.9	2.0
	• Uniform		-23.0	-18.5	-4.5	2.0
	• Gaussian		-21.7	-17.2	-4.5	2.0
	Tampi (2015):					
	• 3D Histogram	3Dh	-11.7	-14.7	-3.0	1,200
SF	**Siebert & Amzel (2004)**	S&A	-12.6	-14.7	-2.1	0.8
	Irudayam & Henchman (2009)	I&J	-	-9.4		1.0

Table.4.4. TDS°$_{Po}$ values obtained from the various methods discussed in this section are listed by paper. Note that Tampi (2015) is from this chapter and the F&J values were recalculated by C&A. Entropy values are categorised by method type (SF or MF) and results from simulations that utilise protein restraints are included to demonstrate the effect of protein flexibility. With regards to 3Dh and S&A methods: the restrained and unrestrained columns refer to the amount of restraints used in RMS fitting the protein reference frame in 3Dh. For S&A, it refers to the method used to approximate the ligand configurational volume (see main text).

4.3.2.2. Approximations, MF and SF approaches

In contrast to the MF methods discussed, both the SF methods generate much lower estimates for TΔS°$_{Po}$ and are closer to -10 kJ/mol, inline with the cratic correction

(§4.1.2.8). As the global entropy change on binding benzene to T4LM is -4.6 kJ/mol, this would suggest that MF theories do indeed overestimate losses in $T\Delta S°_{Po}$.

However, there are two additional factors that should be taken into account:

1. There is no unequivocal experimental or computational estimate of the translational component for this particular system. Therefore, it is plausible that the highly unfavourable MF $T\Delta S°_{Po}$ estimates are actually correct in the first place and the many DOF in the protein favourably compensate this unfavourable component.

2. 3Dh is the epitome of a MF method, in which the ligand is treated as a "particle in a box". Despite this heritage, the approach yields $T\Delta S°_{Po}$ estimates of -11.7 and -14.7 kJ/mol. This is a result that SF proponents declare to be impossible due to fundamental MF methodological flaws associated with defining the free volume [164,311,319,325]. What is the rationale for this contradiction?

Point-2 will be assessed, because point-1 is impossible to prove without calculating 2nd to 3rd order entropy contributions for every component (protein + ligand + solvent) and comparing this against the global entropy of binding,.

The simulation timescales explored by all the other methods are extremely short (**Table.4.4**). If the (~600 to 1,200x) additional sampling is considered, it is obvious that simulations run at shorter timescales have the propensity to return severely underestimated bound volumes (**Fig.4.9.a**). If sampling is < 2 ns, some kind of functional form must be used to compensate for gaps in the data. Even at 20 and 100 ns, the COM is poorly represented compared to fuller distributions captured from 600 to 1,200 ns datasets. Note that the shape of the cavity volume is also better matched by the 1.2 μs COM distribution, and the light grey areas corresponding to areas of low occupation indicate that benzene can shift its position to greater extents than previously suggested (**Fig.4.9.f,g**). Additional minima are not fully revealed under ~100 ns and the true extent of COM displacements is not uncovered till ~350 to 400 ns (**Fig.4.9.a**). Thus, it is very likely that a functional form used in conjunction with undersampled data, will result in underestimated distributions without *a priori* knowledge. This would result in $T\Delta S°_{Po}$ being more negative and unfavourable.

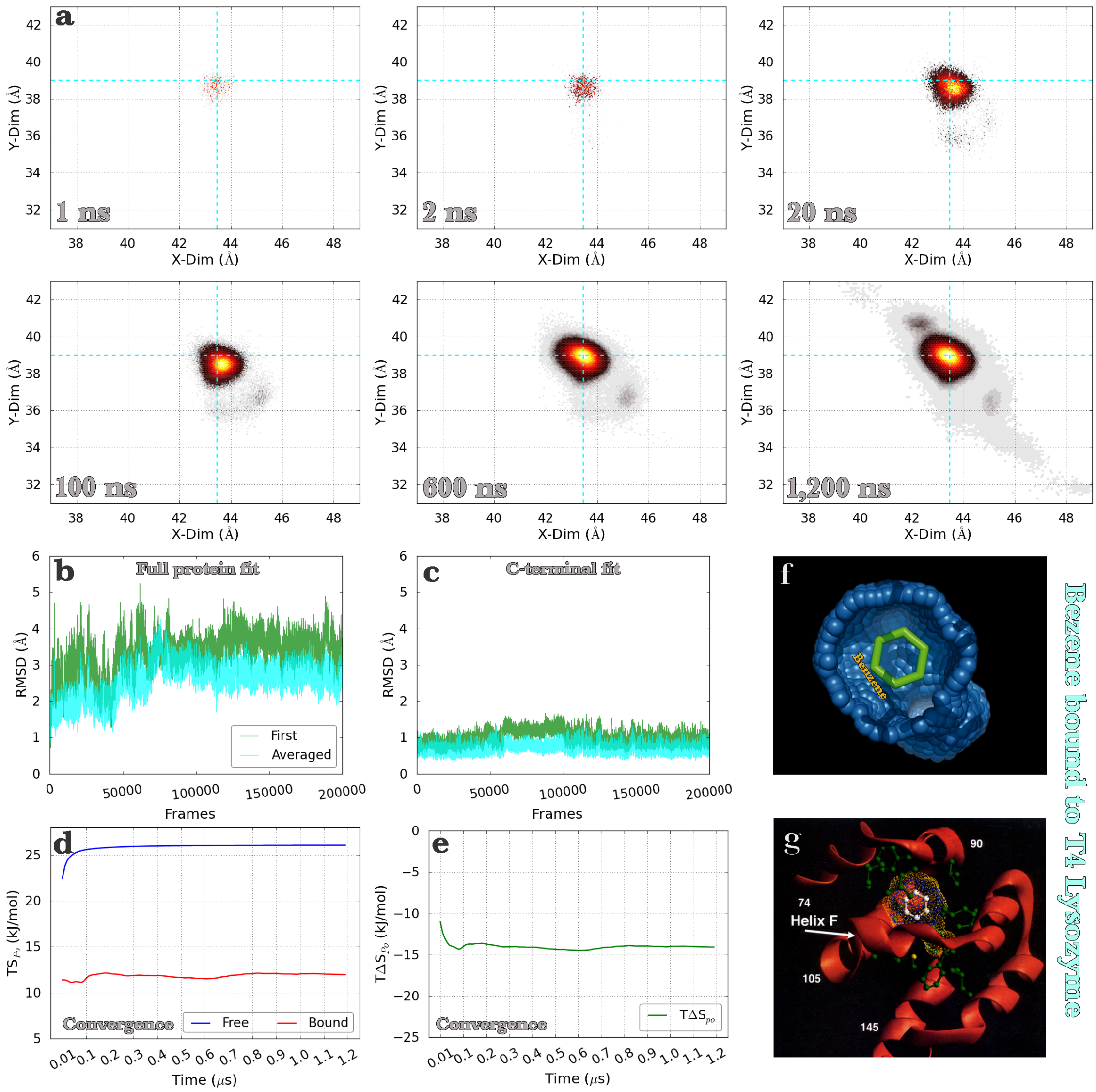


Fig.4.9. (a) The effective 3D bound volume of benzene bound to T4LM is shown as a 2D COM density distribution through the xy projection. The top two rows show the growth of the effective bound volume as the simulation is extended. Note that the centre of the distribution (dashed blue lines at 1.2 µs) shifts because the trajectory is concatenated from 6 independent 200 ns simulations (§4.3.6). **(b-c)** A comparison of RMS fitting trajectory frames to the whole protein versus just the C-terminal ligand binding region. Furthermore, fitting to an averaged reference structure (cyan) generates a much better fit compared to the first frame (green). **(d)** Convergence of positional entropy for bound and free states taken in 10 ns increments over the full 1.2 µs dataset. A bin size of 60 was used to calculate the entropies. **(e)** Convergence of $T\Delta S^{\circ}_{Po}$. **(f)** Cutaway of cavity binding surface (blue) showing bound benzene (green). if present, the lower right protuberance can easily accommodate benzene substituent groups. As demonstrated by panel-a, protein reorganisation allows infrequent entry of the aromatic group into this zone and another at the top left. Image adapted from Basse et al. (2010) [365]. **(g)** Image taken from Morton and Matthews (1995) depicts benzene bound to the carboxy-terminal domain of T4LM. The cavity volume and van der Waals surface of benzene are coloured with yellow and blue dots respectively [354].

3Dh captures $T\Delta S^{\circ}_{Po}$ values by measuring the displacements of the ligand relative to the proteins frame of reference. In order to check the error associated with the fitting procedure, calculations were executed on trajectories that were RMS fitted to the entire protein, versus C-terminal domain residues (84-162) only (**Fig.4.9.b,c**). The former represents more restrictive fitting restraints than the latter and $T\Delta S^{\circ}_{Po}$ values of -11.7 and -14.7 kJ/mol were generated respectively. The reference frame "jitter" accompanying the RMS fit improved a hundredfold (1.7 x 10^{-2} to 8.5 x 10^{-4} $Å^3$) when fitting restraints were relaxed to exclude residues distal from the binding cavity. Thus, the C-terminal fit generated the more accurate value and as 8.5 x 10^{-4} is such a small volume, the error associated with the fit is negligible.

3Dh is less accurate with small datasets (< 100 ns) and the convergence plots indicate that most of the uncertainty arises from the steep change to the free volume estimate (**Fig.4.9.d**). This converges relatively quickly (~210 ns), but estimates of the bound volume change and fluctuate much more gradually as more data points are added to the calculation. These two variables result in $T\Delta S^{\circ}_{Po}$ losses being underestimated (~ -11 kJ/mol) with poor sampling under ~10 ns (**Fig.4.9.e**). Benzene-T4LM is particularly suited for methods based upon the harmonic approximation as it possesses a single minimum, but 3Dh generates reduced $T\Delta S^{\circ}_{Po}$ penalties (> -17 kJ/mol) compared to other MF methods because it better captures the detailed shape and density of the distribution via a probabilistic approach. However, with small datasets, $T\Delta S^{\circ}_{Po}$ is underestimated because missing data cannot be extrapolated via a functional form. Whilst approximations provide this facility, the results returned are accompanied by estimation errors when the true size of the distribution is unknown. In this particular case, $T\Delta S^{\circ}_{Po}$ values converge quickly due to the single bound minimum. Other MF methods based on approximations are thus, also likely to yield similar entropy estimates to 3Dh when 100 to 200 ns simulations are considered. So, why do SF methods [164,325] produce comparatively lower $T\Delta S^{\circ}_{Po}$ estimates than these MF approaches, despite possessing the worst sampling <= 1.0 ns?

In the case of S&A, complexed simulations were run in the gas phase using 10.0 kJ/mol $Å^2$ restraints on all C_{α} atoms. Using cell theory they calculated the loss of translational DOF by summing the entropic contribution arising from ligand differences in bound and free configurational volumes; the cratic correction, and the communal entropy (§4.1.2.8). It is noteworthy that they obtained bound ligand configurational volumes (0.095 $Å^3$) that were significantly smaller than that in the free state (0.245 $Å^3$). They also make the point that the cratic correction does not account for inequalities between endpoint configurational volumes and if this was not calculated, $T\Delta S^{\circ}_{Po}$ would have been more favourable by 2.4 kJ/mol (**Table.4.4**). Thus, the correction was deemed a simplistic measure that failed to account for the differences in the intramolecular interactions the ligand makes with solvent compared to the protein [325]. Why do S&A obtain an identical value to 3Dh? On one hand, they restrain the protein, but as simulations are performed in the gas phase, sampling of protein internal DOF could potentially be improved. Furthermore, they found that fitting the Boltzmann factor of the ligand over a grid gave better results than a quadratic function (unrestrained and restrained values in **Table.4.4** respectively). This would avoid errors associated with the harmonic approximation. Hence, these multiple confounding factors make it difficult to pinpoint the precise reason for the similarity.

The remainder of this section focuses on the more recent formulation proposed by I&J which utilises a harmonic approximation [164]. The subscript "Tr" will be used because the partition scheme includes the momentum (§4.1.3).

I&H assert that use of force and "torque" constants yield lower effective bound volumes compared to ligand displacements that are defined relative to the reference frame of the protein. They also suggest that bigger losses in $T\Delta S^{\circ}_{Tr}$ generated by MF methods are due to large, protein reference frame shifts that artificially inflate ligand residual motion. However, the argument is based on proposals articulated by Gilson with reference to separating internal and external DOF via the use of the BAT coordinate system (§3.1.2). This was shown to improve estimates provided by the QHA compared to the use of Cartesian coordinates which tended to distort *molecular conformations and rotations* [203,304]. Additionally, Gilson's comments chiefly refer to the 3 atom "anchored Cartesian" coordinate system used in his body of work [182,203,204,304,324,366,367]. These criticisms are not applicable to the case of 3Dh, because portions of the protein backbone were RMS fitted to an averaged reference structure and the amount of reference frame "jitter" was negligible (**Fig.4.9.b,c**).

The SF approach proposed by I&G assumes a single bound minimum based on well-defined ligand poses observed in crystal structures and this leads to a fundamental difference in partitioning [164]. In the case of protein-ligand binding, the free ligand is defined as possessing hindered translational (cratic + "vibrational") and rotational (orientational + librational) motion within its solvent cage. Librations are defined as hindered rotations and are analogous to the vibrational component. After binding, the ligand is assumed to remain within a single bound minimum and this is used to justify zero cratic and orientational contributions i.e. only high frequency vibrations and librations remain. The partitioning scheme used is somewhat involved, and a simplified breakdown is presented below. As all the ligands tested were relatively rigid, internal contributions were ignored.

$$S_F = S_{Tr}\text{ (vibrational + cratic correction)} + S_{Ro}\text{ (librational + orientational)} + S_{int}$$

$$S_B = S_{Tr}\text{ (vibrational + }\textbf{no}\text{ cratic correction)} + S_{Ro}\text{ (librational)} + S_{int}$$

The "vibrational" contribution to translation is computed from the differences in bound and free configurational volumes. These calculated volumes are inversely proportional to the averaged force constants "felt" by the ligand in the mean field of its neighbours (§4.1.2.8). It is purported that a free ligand such as benzamidine (with two polar groups) will form stronger interactions with its neighbours, and thus possesses a small configurational volume of 0.053 Å^3. Conversely, benzene has a larger volume of 0.403 Å^3 because apolar interactions contribute less to the force constant. Thus, the "vibrational" entropy of the former (18.0 kJ/mol) is smaller than the latter (21.5 kJ/mol). The situation is similar for the bound state and the near equivalence of endpoint configurational volumes yield minimal differences (< ~1.5 kJ/mol). Hence, the *difference* in "vibrational" contribution to translation is dwarfed by the cratic correction (~10.0 kJ/mol), and is largely invariant across several protein-ligand binding systems (**Table.4.5**) [164].

Irudayam & Henchman (2009)			**Translational entropy (kJ/mol) at 300K**					
			Free	**Bound**	**Diff**	**Free**	**Bound**	**Total**
	Protein	**Ligand**	**Vib**	**Vib**	**Vib**	**Cratic**	**Cratic**	**$T\Delta S_{Tr}$**
	T4LM	Benzene	21.5	21.7	0.2	10.0	0.0	-9.8
	FK506	BUT	18.5	18.4	-0.1	10.0	0.0	-10.1
	Trypsin	Benzamadine	18.0	16.7	-1.3	10.0	0.0	-11.3
	MUP	IBMP	21.7	22.4	0.7	10.0	0.0	-9.4

Table.4.5. Decomposition of translational entropy values generated by I&J's SF method for several protein-ligand binding interactions [164]. Values were calculated from Tables.3-4 in that work. Abbreviation BUT stands for 4-hydroxy-2-butanone.

Interestingly, I&J compute the contribution from each vibrational DOF separately and thus, do not take the correlations between them into account. As correlations reduce the magnitude of the calculated entropy, the already negligible vibrational contribution should become even smaller and irrelevant. By this argument, it might seem that there is little need in rigorously calculating the "vibrational" component via an intricate array of equations as the entire translatory contribution is "captured by the cratic correction". However, can this fundamental entropic contribution be reduced to a group effect that fits any protein-ligand binding interaction? Intuitively, the disparate systems tested should yield a range of $T\Delta S^{\circ}_{Tr}$ values > 1.9 kJ/mol due to disparities in the way bound ligands interact with different binding cavity architectures.

The near equivalence of endpoint configurational volumes suggests that:

1. On taking the "ligand's point of view", there is equivalence between solvent-ligand interactions that exist prior to binding, and protein-ligand interactions after binding. This is because the averaged force constant is influenced by both electrostatic and dispersive interactions. However, this is at odds with the phenomenon of proteins that bind with an enthalpic signature; a phenomenon that contradicts the "classical" hydrophobic effect (§1.3.2) [93]. In the case of the apolar association of ligands (such as IPMP) with MUP, the enthalpic binding signature arises from the inequality of dispersive interactions between endpoint states because the cavity is suboptimally hydrated [36,89]. If the interaction is analysed from the protein "point of view", it can be seen that the barrel-like structure of the β-clam fold and lack of bound waters result in the calyx possessing unsatisfied van der Waals interactions. Thus, on binding, the presence of the ligand creates favourable protein-ligand interfacial interactions that overshadow lost solvent-ligand contributions. This inequality results in the favourable enthalpic binding signature [34,36,93]. It just so happens, T4LM also possesses a suboptimally hydrated cavity to which hydrophobic ligands bind with an predominantly enthalpic signature... [252,354,365]

2. There is equivalence between the endpoint "vibrational" entropies (~21.8 kJ/mol) of benzene-T4LM and IBMP-MUP (**Table.4.5**). In the case of the bound ligand, this suggests that IBMP "feels" almost identical averaged forces to benzene despite being bound to a structurally distinct cavity. Whilst T4LM offers a relatively tight fitting hydrophobic binding site [354], MUP's cavity is much more spacious as evidenced by its ability to promiscuously bind different ligands of various sizes [36,89]. Indeed, microsecond length MD simulations demonstrated that IBMP, a larger congeneric ligand retained significant residual motion in the pocket [91]. Furthermore, it is unclear why IPMP-MUP does not have a larger force constant compared to benzene-T4LM, because IBMP contains 3 electronegative atoms which have the potential to make H-bonds. Conversely, benzene would principally interact with its environment via van der Waals interactions. So if the assumption of a single bound minimum was discarded and IPMP-MUP simulations were extended, enhanced sampling would result in more H-bonded poses and these interactions would increase the averaged force constant. The bound configurational volume and thus the associated entropy would then be decreased. IPMP would then paradoxically possess bound "vibrational" entropy that was less favourable than benzene, even though the latter is clearly ensconced in a much smaller pocket. Regardless, this component is much too meagre to significantly affect $T\Delta S^{\circ}_{Tr}$. More importantly, if the method was expanded to include multiple minima, how would that affect the cratic correction? As it could no longer be set to zero in the bound state, $T\Delta S^{\circ}_{Tr}$ would be less than ~10 kJ/mol; significantly less, in the case of MUP due to the reasons discussed above. Furthermore, bound orientational entropies would also be > 0. The difference in the orientational entropy is of a similar magnitude to the cratic correction and when all these ligand-centric contributions are accounted for, $T\Delta S°_{Tr}$ is likely to fall significantly short of the experimental techniques the method was validated against [164]. As the global entropy is only -4.6 kJ/mol, it could be argued that this would make the approach more accurate. However, the many DOF available to protein are likely to generate contributions that dominate that from the ligand. There should therefore be more than a mere tremor in the force constant, and the proposed equality between endpoint configurational volumes should be questioned. Moreover, in the case of systems characterised by multiple bound minima, this SF approach is likely to require detailed analyses to establish the validity of the cratic correction.

4.3.2.3. Summary

In the case of benzene-T4LM, long 1.2 μs simulations demonstrate that benzene retains considerable ability to shift its position due to protein flexibility (**Fig.4.9.a**). This results in an effective bound volume much greater than the SF estimate of 0.450 Å^3. Most importantly, despite identifying a larger bound volume, 3Dh generates $T\Delta S°_{Po}$ values that are more favourable than other MF methods. As SF proponents have argued that this was not possible due to fundamental flaws associated with MF approach, this result validates the conceptual approach of MF theories. The reason why MF entropies were more unfavourable than expected, is very likely to be due to errors associated with using (quasi)-harmonic approximations which have been demonstrated to be unreliable when dealing with multimodal distributions [203,339]. Additionally, such approximations are unlikely to compensate for undersampled trajectories when the correct sizes of the Boltzmann weighted ratio of stable states is unknown (§4.3.8.1).

If the analogy that one human year is equivalent to seven dog years is extended to computer life spans; mechanical/electronic failures, Moore's law and planned obsolescence result in an average workstation lifetime of ~7 years. Thus, assuming an average human life expectancy of ~67 years, every human year equates to ~10 computer years. So, computers built in 2005 and 2010 would currently be a hundred, and fifty human years respectively. On considering higher operational temperatures and rapid development cycles, GPUs should age at an even faster rate. Given the exponential acceleration of computational capabilities (§1.4.3), there is little need for the use of harmonic approximations or other simplistic functional forms to calculate the entropy of moderately sized systems in the year 2015. These were ingenious, respected methods that were relevant at a not too distant time in scientific history. But in the information age, shouldn't such approximations coupled with sub-microsecond simulations be regarded as an anachronism?

4.3.3. MUP positional entropy: multiple bound minima

The binding of ligands to MUP is likely to be characterised by multiple bound minima, and 3Dh's ability to generate reliable $T\Delta S^{\circ}_{Po}$ values is further tested in the remainder of this section.

4.3.3.1. Analysis of n-alkanol & 3Z-olefin COM distributions

If the full COM distribution (grey to orange/yellow) is considered, it is apparent that ligands bound to MUP retain considerable freedom of movement. **Fig.4.11-12** shows the COM translational displacements of n-alkanols and 3Z-olefins in relation to the COM of key amino acids which are colour coordinated according to the key provided in the bottom row. A zoomed in representation of the figures is provided in **Fig.4.13-14** without amino acid residues. Ligand COM density plots in these two figures have been normalised with respect to each other to allow cross-comparison. The following trends are observed:

1. Wide spread (light grey) areas of low population surrounding higher density minima suggest that the ligand can sample multiple positions and orientations within the pocket. Visual inspection indicates that these are dissimilar to the bound crystallographic ligand poses published by Malham et al. (2005) [36].

2. The accessible volume represented by the COM in the 2D plots is larger for n-alkanol 6C and 7C ligands as their smaller size allows them to explore a greater proportion of the cavity. It is more difficult to discern the distinctions in the overall spread of distributions possessed by the 3Z-olefin ligands.

3. Whilst smaller ligands can easily occupy spacious regions near the W loop in their entirety (§3.3.6), longer ligands from both panels tend to increase occupation of high density minima further within the depths of the calyx.

4. The distributions depicted in **Fig.4.13-14** demonstrate that 3Z-olefins have a broader, more evenly distributed spread of COM minima than their n-alkanol counterparts which tend to possess highly occupied minima. This indicates that 3Z-olefins have the ability to rapidly shift their COM, and the resulting distributions in the zy plane form a rough "kidney shape" corresponding to the rise and fall of the ligand's COM within the cavity. If the bottom left quadrant (demarcated by dotted cyan lines) is examined, it can be ascertained that 3Z-olefin ligands maintain a significant cluster of points in this zone that is "off-centre" from the y-axis. On the other hand, n-alkanols dwell in more localised minima that become reduced in number as the length of the ligand is increased. With the exception of hep, no saturated ligand maintains a stable cluster in the bottom left quadrant.

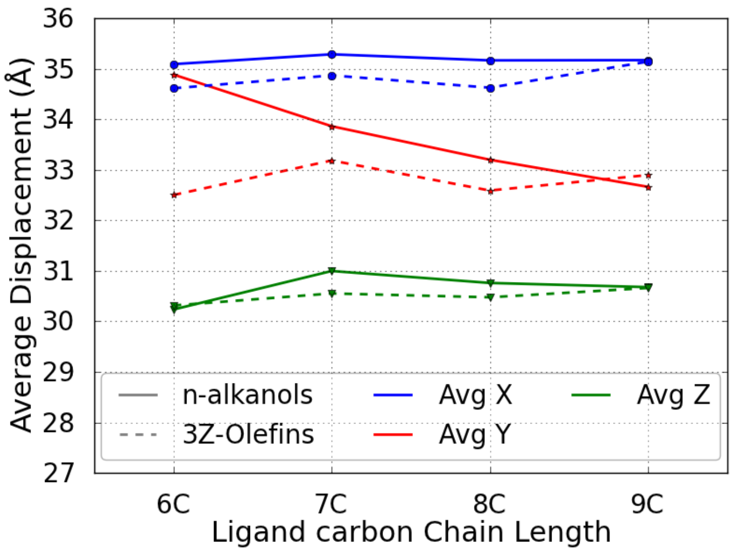


Fig.4.10. Average ligand COM displacements. Whilst all 3Z-olefins possess roughly the same average displacement value along each axis, n-alkanols progressively become more localised within the calyx (y-axis) as carbon chain length increases. Note that zig-zag trends are likely to be the result of odd/even alternation of ligand length.

The average displacements along the three principal axes for ligands in both panels are illustrated in **Fig.4.10.** While there are small inter-panel differences along the x and z axes, the most obvious disparity occurs along the y axis. Whilst, all 3Z-olefins maintain relatively constant positions along the principal axes, n-alkanols increase occupation of positions deeper within the cavity as ligand size increases. When this observation is coupled with the concomitant reduction in stable minima, there is a strong indication that the size and structure of n-alkanols predispose them to become "trapped" by binding site residues.

Malham (2012) compared crystals of MUP complexes and found that the electron density of bound 3Z-olefin ligands was less defined than n-alkanols. This fact supports the observation that unsaturated ligands are able to access a greater number of bound poses than their saturated analogues [180].

5. The range of COM motion along the three axes can be ordered as y > z > x for all the ligands (**Table.4.6**). A large contribution to the range along the z axis stems from displacements towards the cavity entrance which is situated in the cleft formed by the *d* and *e* strands. This location was proposed by Timm et al. (2001) as being associated with ligand ingress and egress [72]. Further examination of the COM distributions along the zy plane indicates that some of the bound ligands (hex, hep, and non) visit areas close to the proposed

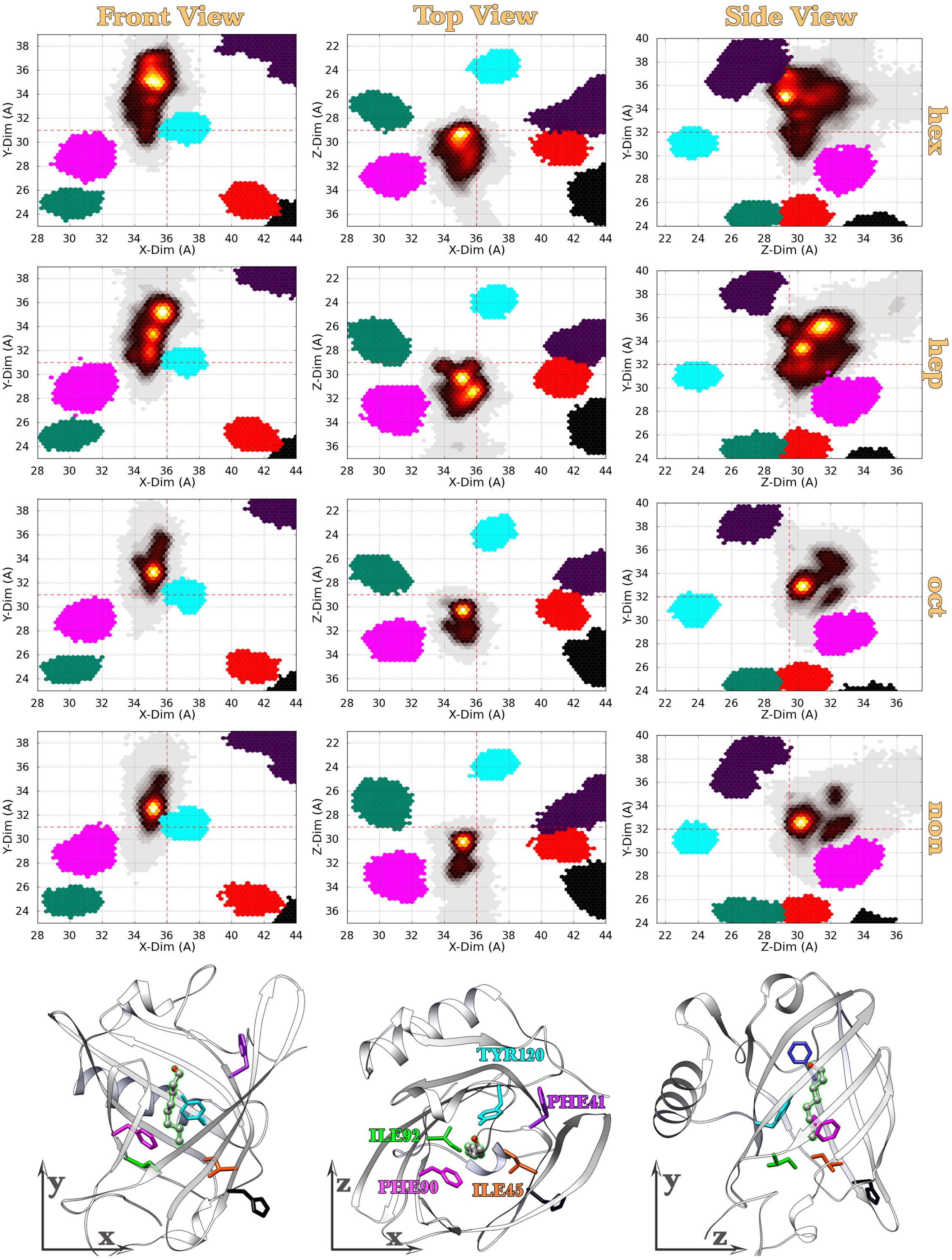


Fig.4.11. Two dimensional density plots representing the COM motions of n-alkanol ligands bound to MUP. Middle panel in bottom row contains key for residue coding. Each graph characterises the motion over a concatenated 1.2 μs trajectory. Density maps are not normalised.

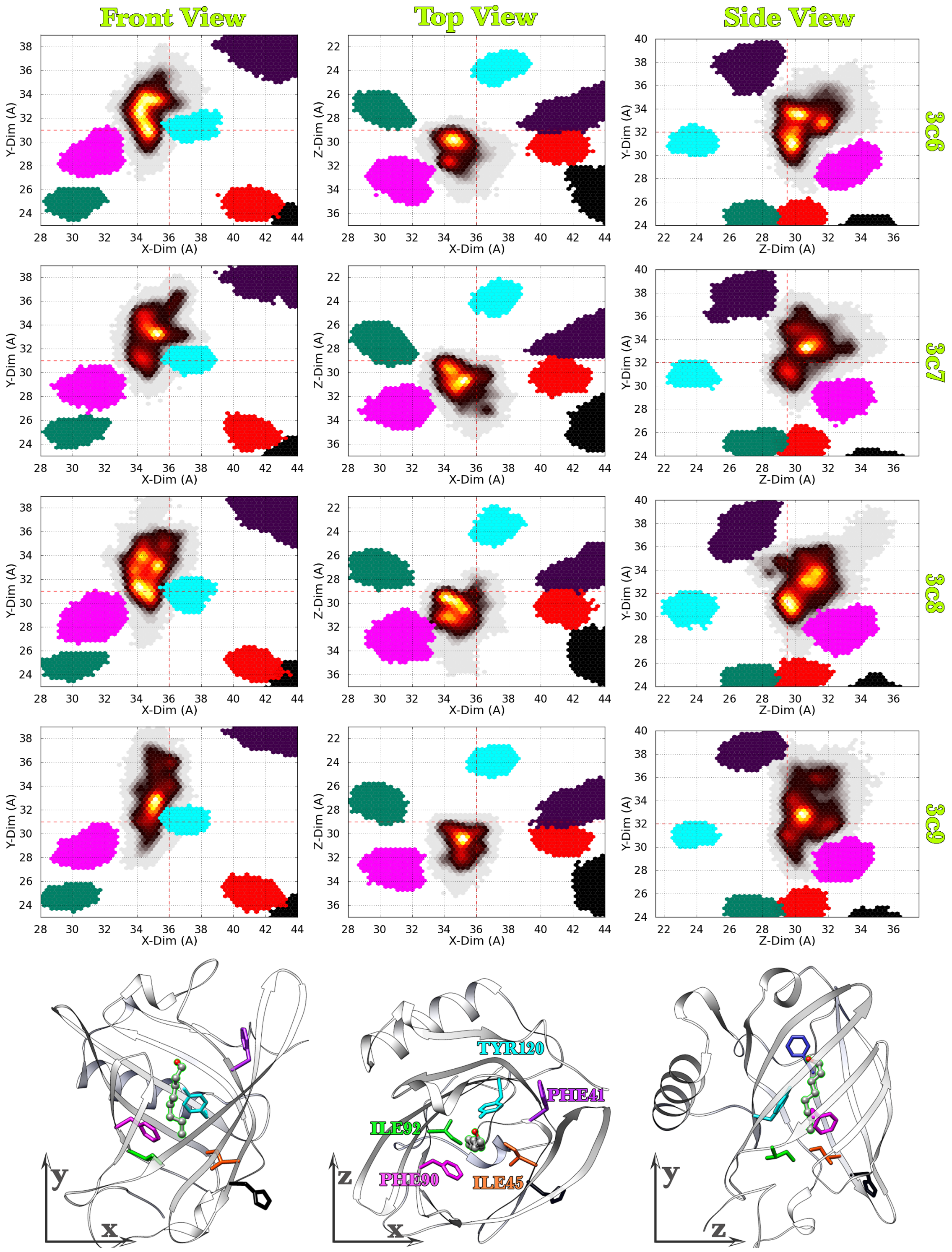


Fig.4.12. Two dimensional density plots representing the COM motions of 3Z-olefin ligands bound to MUP. Middle panel in bottom row contains key for residue coding. Each graph characterises the motion over a concatenated 1.2 μs trajectory. Density maps are not normalised.

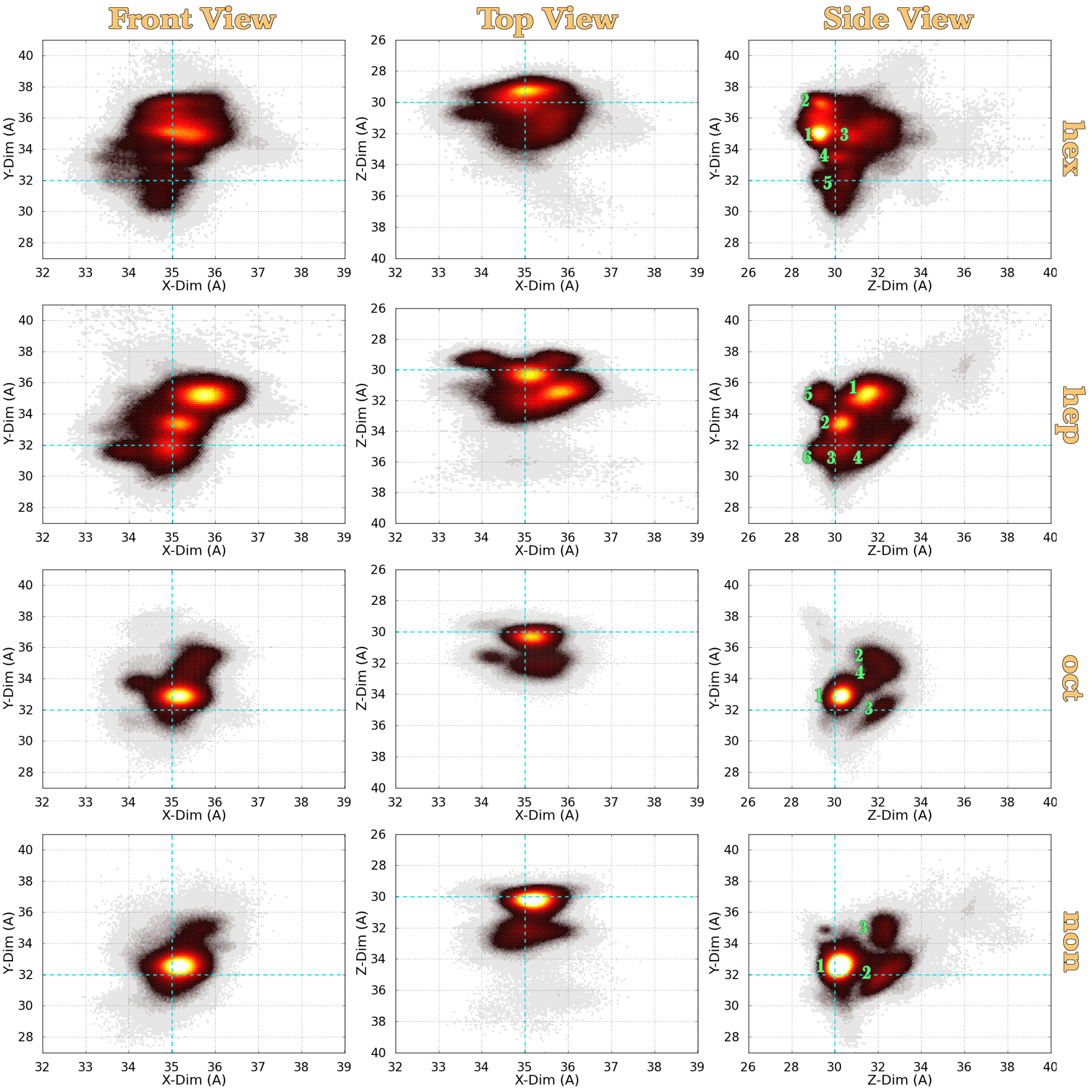


Fig.4.13. Magnification of two dimensional density plots depicting the COM motions of n-alkanol ligands bound to MUP. Each graph characterises the motion described over a concatenated 1.2 μs trajectory. Densities are normalised so that the maximum count per hexagonal bin is capped at a maximum of 5,000, to allow cross-comparison between ligands in both panels. Clusters isolated via mean shift clustering are marked with green numbers in the zy projection (§4.3.3).

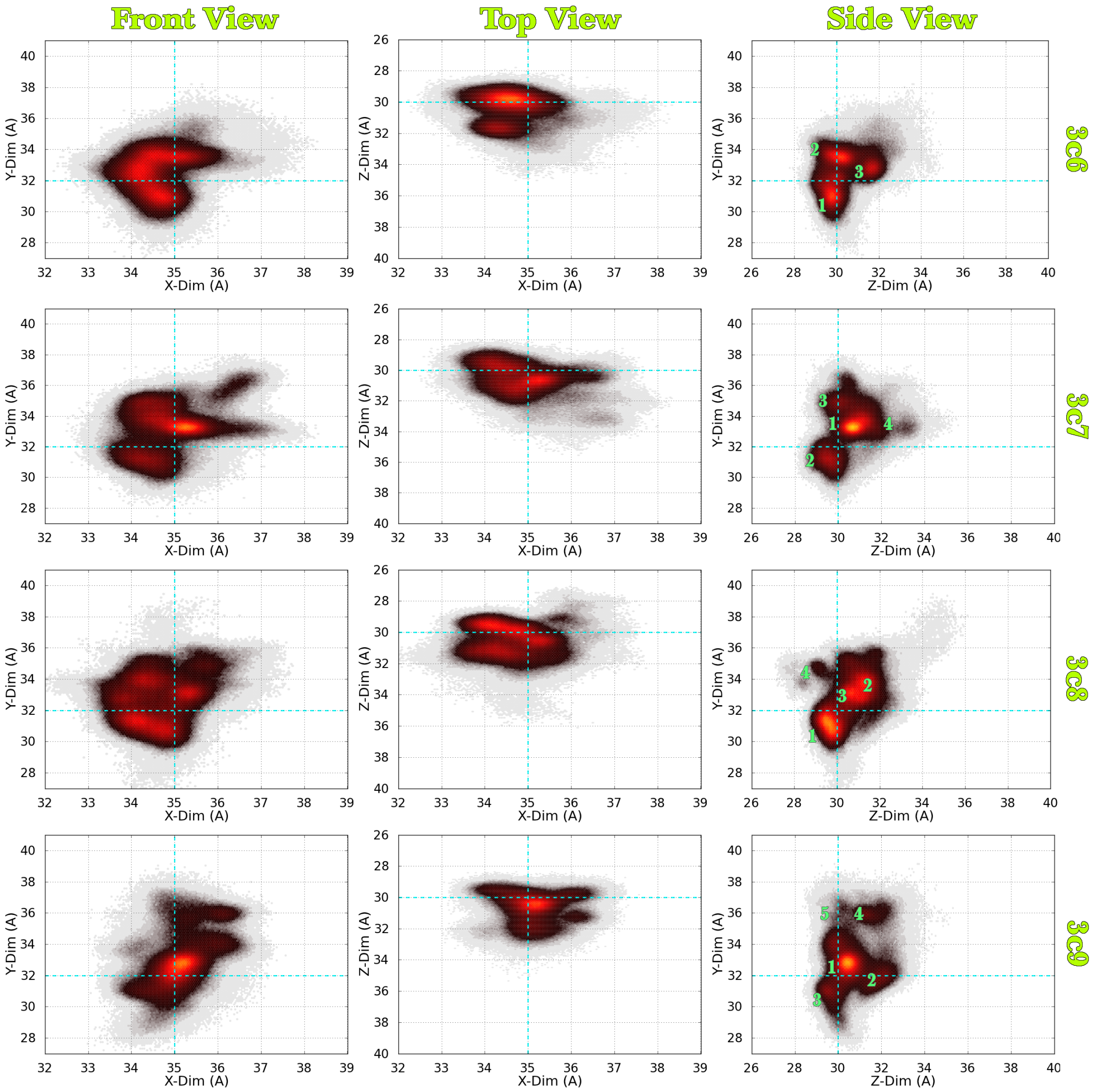


Fig.4.14. Magnification of two dimensional density plots depicting the COM motions of 3Z-olefin ligands bound to MUP. Each graph characterises the motion described over a concatenated 1.2 μs trajectory. Densities are normalised so that the maximum count per hexagonal bin is capped at a maximum of 5,000, to allow cross-comparison between ligands in both panels. Clusters isolated via mean shift clustering are marked with green numbers in the zy projection (§4.3.3).

gateway to greater extents than other ligands. For example, in the case of hep the zy coordinates are 36,37 (**Fig.4.11-12**). Attempted egress is a stochastic phenomenon as there is a lack of multiple restraining polar interactions within the calyx. Nevertheless, these excursions constitute a minority of snapshots (< 1%) of the 1.2 μs concatenated trajectories and tests demonstrated that the positional entropy calculation was negligibly perturbed.

	Angstroms (Å)			
	X Range	**Y Range**	**Z Range**	**Clusters**
hex	6.1	14.3	12.5	5
hep	8.0	14.6	11.9	6
oct	4.4	11.5	6.5	4
non	4.8	13.1	11.9	3
	X Range	**Y Range**	**Z Range**	**Clusters**
3c6	6.1	11.4	7.2	3
3c7	5.8	10.5	7.6	4
3c8	5.8	13.1	9.0	4
3c9	4.6	11.9	7.0	5

Table.4.6. The range of ligand COM displacements (Å) along each axes. The range is distinct from the average values depicted in **Fig.4.10** as it reports on the spread of values along each axis, as opposed to the mean of the distribution.The number of high density states identified by mean shift clustering is tabulated in the final column.

4.3.3.2. Positional entropy of n-alkanol & 3Z-olefin ligand panels

The positional entropy was calculated for free ($TS^{\circ}_{Po\bullet F}$) and bound states ($TS^{\circ}_{Po\ Po\bullet B}$) using **eqn.4.17**. Data points from concatenated 1.2 μs trajectories were used to generate the more accurate value for each ligand by maximising sampling. As demonstrated in **Fig.4.15.a-c**, values of TS°_{Po} obtained for free and bound states are not meaningful in themselves. Only relative differences ($T\Delta S^{\circ}_{Po}$) between the two possess physical significance (**eqn.4.18**). It is difficult to visually identify the optimal number of bins by examining the change in TS°_{Po} on increasing the number of bins along each axis (**Fig.4.15.a,c**). However, the data obtained for $T\Delta S^{\circ}_{Po}$ flattens after a certain point and the correct answer is more apparent. Thus, 60 bins were selected on the basis that changes in $T\Delta S^{\circ}_{Po}$ (i.e. $T\Delta\Delta S^{\circ}_{Po}$) on increasing bin size were minimal at this point (**Fig.4.15.b,d**).

The precision of the method was measured by performing the entropy calculation on the 6x 200 ns repeats constituting the joined trajectory and generating the standard error. The margin of error is greatest (<= 0.70 kJ/mol) in the case of bound ligands because protein conformational changes drive shifts in the architecture of the binding site, and this limits or abets ligand translocation within the confines of the cavity. This results in relatively greater variability between individual repeats than that observed for the free ligand because movement between all localities is not equally favourable. Conversely, all regions within the free simulations are equally favourable. The extremely small (< 0.15 kJ/mol) standard error observed for free ligands indicates that the free volume is sampled homogenously and variations between individual repeats

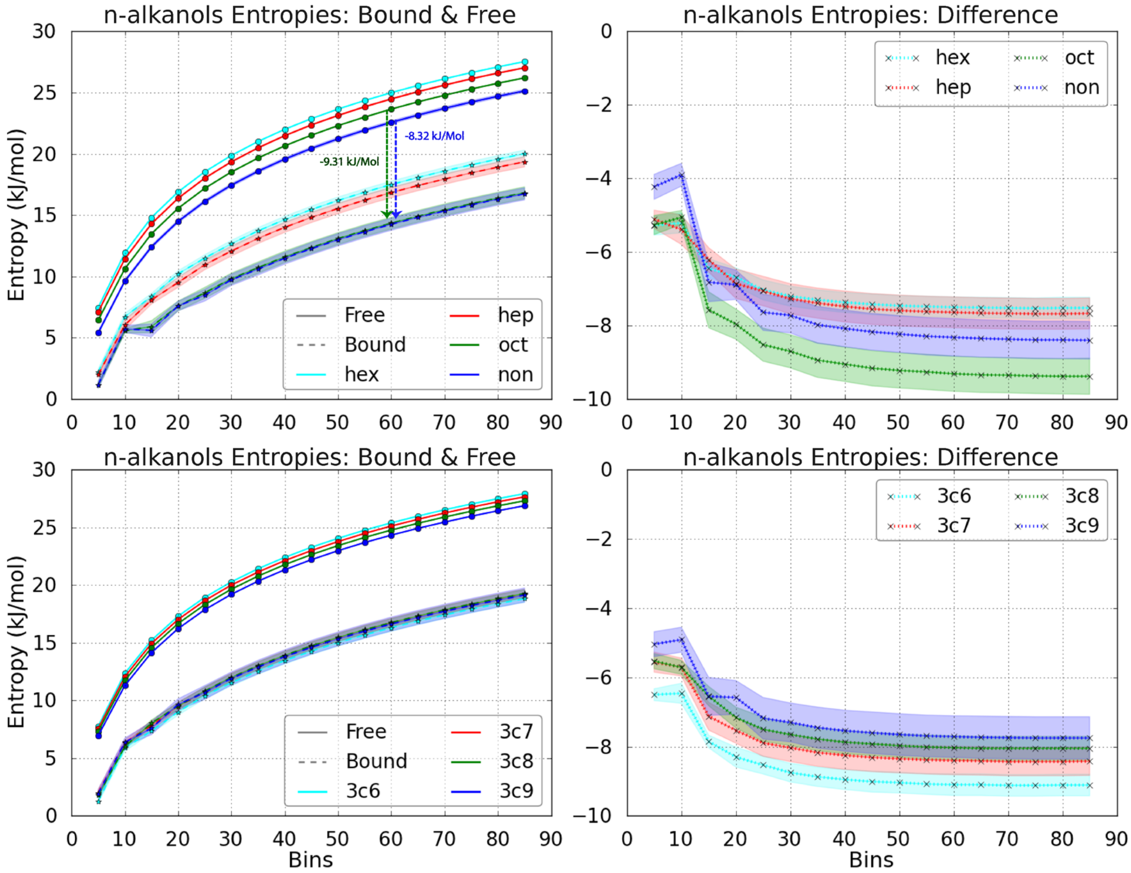


Fig.4.15. (a) Endpoint entropies for n-alkanols n-alkanol bound & free states. Dashed green and blue lines demonstrate how the inter-panel free entropy differential dominates the entropy change on binding. Identical bound entropies result in oct counterintuitively losing more entropy on binding than non. **(b)** $T\Delta S^{\circ}_{Po}$ values obtained for n-alkanols. Standard error is plotted as background shaded area. The error for free state is so small, it can barely be seen. **(c)** Endpoint entropies for 3Z-olefins bound & free states. Note that all 3Z-olefins have near-identical bound values. **(d)** $T\Delta S^{\circ}_{Po}$ values obtained for 3Z-olefins. As bound entropy values are so similar, the free entropy once again dominates the change in binding entropy.

are minimal (**Table.4.7**). In both cases, the method employed obtains consistent results within 200 ns sampling time. The COM deviations of the protein backbone (N, C_{α}, C, O) were also assessed with a view to calculating the entropic cost to the quality of the fit. However, the distribution of points was too small to reliably measure because, the largest amount of frame "jitter" occupied a negligible volume of ~6.1 x 10^{-4} Å^3.

Of particular interest is the rank order of the different ligands in terms of $T\Delta S^{\circ}_{Po}$. Intuitively, it would be expected that by virtue of size, shorter ligands would suffer less restriction on binding, than longer ligands. The COM density distributions for n-alkanols and 3Z-olefins support this idea, though $T\Delta S^{\circ}_{Po}$ disparities between ligands within the latter panel are smaller than those in the former (**Fig.4.16**). When free, the smallest ligands do have greater amounts of translational motion than larger ligands as expected. However, values for $T\Delta S^{\circ}_{Po}$ indicate that some of the shorter ligands suffer *greater* positional restriction than their longer counterparts (**Fig.4.15.b,d**). The rank order of all ligands within the 3Z-olefin panel is the precise opposite of what was anticipated, and the position of oct and non are inverted with respect to size based expectations (**Fig.4.15.b,d**). This counter-intuitive result has physical significance and the examples can be explained by the fact that when ligands in the bound state have *near identical values* for $\mathrm{TS}^{\circ}_{\mathrm{Po\bullet B}}$, the entropy differential *between the same ligands in the free state* ($\mathrm{TS}^{\circ}_{\mathrm{Po\bullet F}}$) *dominates* the value obtained for $\mathrm{T\Delta S}^{\circ}_{\mathrm{Po}}$. This is illustrated for the case of oct and non in **Fig.4.15.a**.

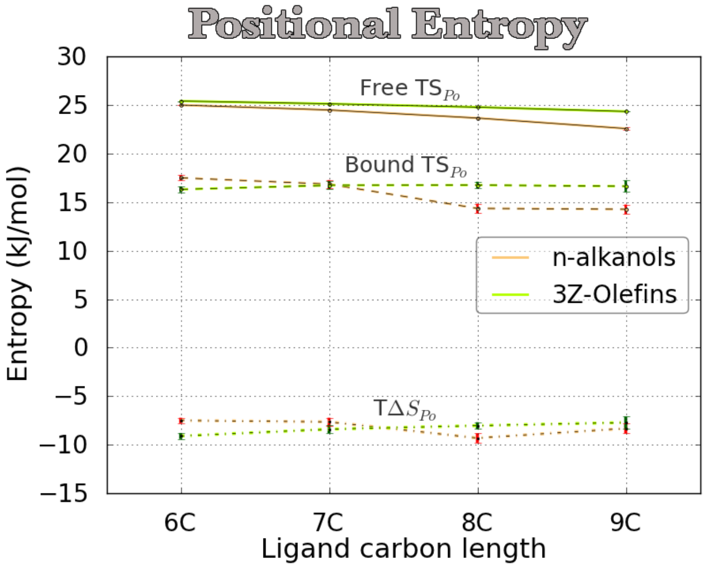


Fig.4.16. Positional entropies ($\mathrm{TS}^{\circ}_{\mathrm{Po\bullet F}}$,$\mathrm{TS}^{\circ}_{\mathrm{Po\bullet B}}$ & $\mathrm{T\Delta S}^{\circ}_{\mathrm{Po}}$) obtained for ligands in n-alkanol and 3Z-olefin panels obtained using 60 bins along each axes.

	Positional Entropy (kJ/mol)			
n-alkanols	**Bins**	**$\mathrm{TS}_{\mathrm{Po\bullet B}}$**	**$\mathrm{TS}_{\mathrm{Po\bullet F}}$**	**$\mathrm{T\Delta S}_{\mathrm{Po}}$**
hex	60	17.50 ± 0.28	25.00 ± 0.02	-7.50 ± 0.28
hep	60	16.85 ± 0.42	24.49 ± 0.05	-7.64 ± 0.42
oct	60	14.35 ± 0.48	23.66 ± 0.03	-9.31 ± 0.48
non	60	14.26 ± 0.49	22.58 ± 0.13	-8.32 ± 0.51
3Z-Olefins		**$\mathrm{TS}_{\mathrm{Po\bullet B}}$**	**$\mathrm{TS}_{\mathrm{Po\bullet F}}$**	**$\mathrm{T\Delta S}_{\mathrm{Po}}$**
3c6	60	16.31 ± 0.31	25.41 ± 0.00	-9.10 ± 0.31
3c7	60	16.73 ± 0.41	25.13 ± 0.00	-8.40 ± 0.41
3c8	60	16.75 ± 0.35	24.78 ± 0.01	-8.00 ± 0.35
3c9	60	16.63 ± 0.60	24.34 ± 0.01	-7.71 ± 0.60

Table.4.7. Positional entropies ($\mathrm{TS}^{\circ}_{\mathrm{Po\bullet F}}$,$\mathrm{TS}^{\circ}_{\mathrm{Po\bullet B}}$ & $\mathrm{T\Delta S}^{\circ}_{\mathrm{Po}}$) with standard error, obtained for ligands in n-alkanol and 3Z-olefin panels obtained using 60 bins along each axes.

Calculated $\mathrm{T\Delta S}^{\circ}_{\mathrm{Po}}$ values indicate that whilst this component contributes a significant amount of energy (-7.50 to -9.31 kJ/mol) to the global entropic signature, differences in-between ligands are not substantial enough to account for the linear trend observed in the ITC data (**Fig.4.16**). Although $\mathrm{TS}^{\circ}_{\mathrm{Po\bullet B}}$ for all 3Z-olefin ligands are equivalent within error, the two larger ligands (oct and non) within the n-alkanol panel are restricted to a greater extent than the shorter members. The resulting disparity between the two groups is ~1.8 kJ/mol at its largest. As the difference is so small, further changes to this trend with increased amounts of sampling cannot be ruled out. However, the magnitude of inter and intra panel $T\Delta S°_{Po}$ differences are minor and the correlation between ligand size and the amount of restriction is weaker than expected because compounds retain considerable translational freedom when bound. For the same reason, the overall magnitude of $T\Delta S°_{Po}$ is less than that obtained in other studies (§4.3.3.3) [91,164].

4.3.3.3. Assessing the accuracy of the positional entropy calculation in MUP

As discussed in §4.1.2.5 there are few (if any), experimental decompositions that yield unequivocal T&R entropy estimates in the literature. Thus, the accuracy of the calculation is difficult to determine. In the case of MUP, the focus has centred on the binding of IPMP and IBMP and these interactions have been assigned a combined T&R penalty of -25.8 kJ/mol [34,89,163]. The value is obtained from the calculation of intrinsic free energies of the cell surface glycolipid ganglioside (GM1) fragments to the B-subunit of the cholera toxin (CTB) [313]. The quantity was later treated as a system-independent contribution that could be applied to various other binding interactions such as the binding of IPMP and IBMP to MUP [34,89,163], and histamine to recombinant histamine-binding protein (rRaHBP2) [368]. The validity of broadly applying this contribution in a system-independent manner should be questioned due to differences in ligand size, structural features, and protein binding site architecture. As GM1 possesses more hydrophilic groups than IPMP and IBMP, *it can, and does* make more H-bonding interactions to CTB (PDB ID: 3CHB) [369,370] compared to hydrophobic ligands bound to MUP. Hence, these multiple, polar directional restraints are likely to be associated with greater T&R entropy penalties on binding, and in the case of MUP, the estimate of -25.8 kJ/mol could be an overestimate.

Irudayam & Henchman calculated the combined loss of external DOF on binding IPMP to MUP to be -22.2 kJ/mol at 298 K, of which -9.3 kJ/mol is derived from the translational component (tables 3 to 7) [164]. However, in the discussion (p5882, section 6) they confusingly state that the value is -25.0 kJ/mol and hence, a good approximation to the (-25.8 kJ/mol) contribution originally proposed by Turnbull et al. (2004) [313]. As the tabulated data reports a value of -22.2 kJ/mol, this is likely to be their actual result. Hence, the claim that their SF method generates a near-identical value to the "experimental estimate" (obtained from a completely different system) is vitiated by a (not insignificant) 3.6 kJ/mol differential [164]. Furthermore, the concerns raised in §4.3.2.2 regarding the validity of this particular SF approach are still pertinent.

Roy & Laughton (2010) performed long MD simulations (1.2 μs x 3 replicates) of IBMP bound to MUP; the larger analogue of IPMP [91]. They observed significant residual motion of IBMP within the calyx and noted that the ligand could take up a larger array of orientations than originally anticipated and that some of these poses were characterised by a lack of H-bonding to TYR120. This directly challenges the notion of a single bound minimum and a rough estimate of the ligand entropy was generated via the method proposed by Schlitter. Ligand translation and rotational tumbling were removed via a series of RMS fits. They concluded that that the total entropy loss on binding was ~ -22 kJ/mol. -13 kJ/mol could be attributed to combined internal and rotational contributions, to which internal DOF only contributed -1 kJ/mol. Using the ideal gas approximation, a rough estimate of -10 kJ/mol was obtained for the loss of translational DOF. However, the magnitude of the entropic penalty could potentially be overestimated for the following reasons:

1. The Schlitter method is based upon the harmonic approximation which is known to account poorly for anharmonicities and multimodal distributions. The binding of IBMP is likely to be characterised by both these factors.

2. The calculation uses 1 microsecond of data with a frame spacing of 1 ns to create a 1,000 frame dataset. Hence, this is equivalent to 1 ns of sampling and this would affect evaluation of the entropy which is critically dependent on correct evaluation of the density of states (§4.3.6). Furthermore, note that the use of such a large frame spacing has been shown to lead to convergence issues when using Cartesian coordinates in conjunction with a covariance matrix (**Fig.2 in** p341 of Baron, 2012) [189]. On viewing figure S12 in the supplementary material of that publication, it is clear that despite significant similarities, IBMP is able to explore different regions of phase space in different 1.2 μs repeats [91]. Obviously, this would affect the distribution and consequently, the entropy result. If more frames (> 1.2 x 10^6) had been utilised, a more ergodic distribution would have been obtained, and the greater number of accessible states is likely to have reduced the entropic penalty.

An examination of the range of values (-7.5 to -9.31 kJ/mol) generated by 3Dh indicates that loss of translational DOF cannot be reduced to a system-independent contribution as there are small differences in the binding of different ligands to the same protein. However, at this stage, an unequivocal conclusion is not yielded by inter or intra-methodological comparisons. Hence, the only remaining solution is to calculate the entropic penalties associated with the remaining ligand DOF and compare this to the global ITC results. As this work assesses the binding of 8 different ligands, any errors in method should be highlighted by discrepancies in entropic trend lines. After calculating the total orientational entropy in §4.3.8, a total system decomposition is presented in §4.4.1-2.

4.3.3.4. Positional entropy convergence and force field dependency

Convergence of TS°$_{Po}$ was assessed to account for the variability of the calculated entropy with increased sampling (**Fig.4.17**). The rapid convergence of TS°$_{Po•F}$ compared to TS°$_{Po•B}$ is apparent and the lack of change after ~200 ns in the former indicates that errors in TΔS°$_{Po}$ would primarily arise from energy barriers preventing bound ligands fully sampling the accessible cavity volume within the simulation. The largest changes in TS°$_{Po•B}$ occur within 300 ns sampling time and the degree of variability decreases on supplying more data points to the calculation. Bound oct takes the longest amount of time to converge and this could be ascribed to higher energy barriers preventing its translocation between the different regions of the cavity compared to shorter ligands. The longest ligand non, is less likely to move between different regions than oct because of its size and thus, the distribution describing its COM displacements has a greater likelihood of being accurately captured. On the other hand, oct has the *potential* to move between different bound minima to a greater degree than non, but it tends to dwell in a given location for greater amount of simulated time. The minima that make up the relevant COM distributions of the other compounds within the n-alkanol panel are built up relatively evenly on extending simulation time. A fact that is supported by the lack of large fluctuations in TS°$_{Po•B}$ upon supplying more data points to the calculation i.e. there are minimal changes in entropy between the datasets > 300 ns (**Fig.4.17**). Thus, these ligands have the facility to transition between available bound minima with higher frequencies than oct. On the other hand, oct tends to remain within stable minima for longer periods of time and larger variations in the T*S*°$_{Po•B}$ convergence

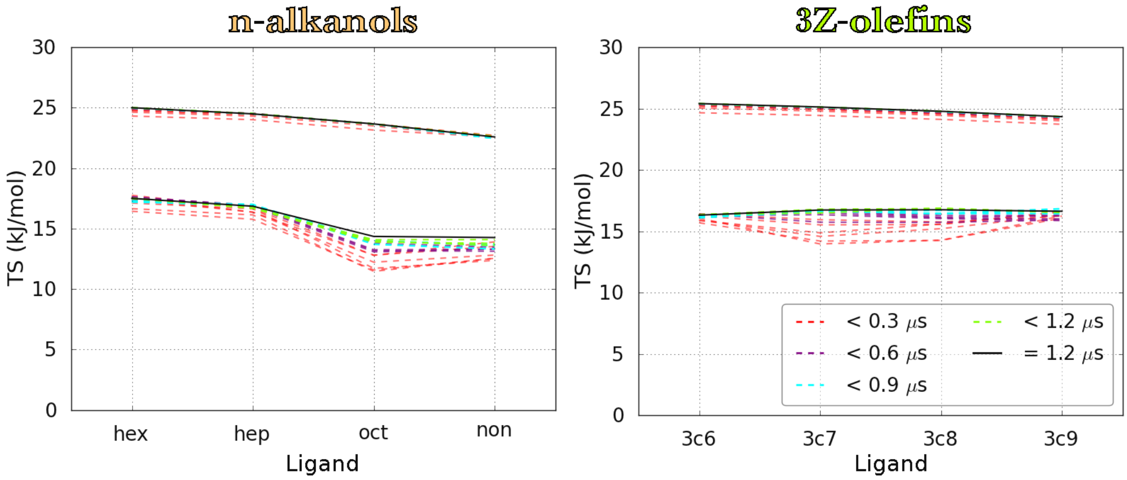


Fig.4.17. Positional entropy (TS°$_{Po•F}$ & TS°$_{Po•B}$) convergence was assessed over 1.2 μs of aggregate simulation time for both ligand panels. The number of points made available to the entropy calculation were iteratively increased by 6 x 10^4 (60 ns) each round. Solid black lines show the value calculated for the full 1.2 μs, whilst coloured dashed lines indicate how the entropy total for a given ligand changes with better sampling. Some of these lines cannot be discerned as they are overlaid and points that possess small separations in-between different dashed coloured lines indicate good convergence. Greater visibility of cyan and green dashed coloured lines is discernable near points that have convergence difficulties e.g. oct.

plot occur because the occupancies of newly unlocked minima increase in an uneven fashion.

The other principal source of error that could well dwarf those already discussed, are inaccuracies in the force field used for the simulations. New force fields are constantly being developed and more recently, the main adjustments have been to the parameters governing important torsion angles rather than fundamental shifts in methodology. It is often not clear which force field offers the greatest level of accuracy until sufficient research is published in the literature and benchmarks performed to validate modifications with experimental data. This can limit early adoption of new developments as researchers would rather use tried and tested parameters. Early ff94 and ff99 force fields overestimated the stability of helical regions due to sub-optimal parameterisation of glycine and alanine residues. Due to computational cost, the torsional parameters of the protein's φ and ψ angles were quantum mechanically fitted to a limited number of low energy conformations of these amino acids and a systematic cumulative error resulted in secondary structure elements such as right handed alpha helices being overly stabilised. The ff99SB force field improved the existing ff99 force field via extensive refitting of these torsional parameters to multiple glycine and alanine tetrapeptide conformations. This resulted in greater rigidity of the backbone and a concomitant reduction in loop dynamics [121,371–375]. The ff03 force field used in this study can be considered a distinct force field as the charge derivation procedure used is fundamentally different [151,371]. Over time there have been various improvements to the ff99SB force field such as the ff99SB-nmr, ff99SB-ildn and the ff99SB-ildn-phi. A recent benchmark rates the ff03 and ff99SB variants as having good agreement with NMR-derived observables. Of particular note were the ff99SB-ildn-phi and ff99SB-ildn-nmr force fields which possessed calculation errors comparable to experimental uncertainty when used in conjunction with the TIP4P-EW water model [375].

Given the discussion above, 6x 100 ns equilibrium simulations of hep bound to MUP were run to verify whether the magnitude of translational motions were diminished when using the ff99SB-nmr force field. The results show that the ligand still retained extensive residual motion within the binding site and the difference in the calculated $TS°_{Po•B}$ value (16.15 ± 0.5 kJ/mol) to that obtained from the ff03 force field is only -0.70 kJ/mol. While there are minor differences in the distributions, the key features are retained (**Fig.4.18**). As the ff99SB-nmr simulations have half the number of data points, it is expected that similarities in the distribution should increase with extended simulation time.

The agreement between the two very different force fields indicates that the radical difference in the method of force field parameterisation does not greatly perturb the

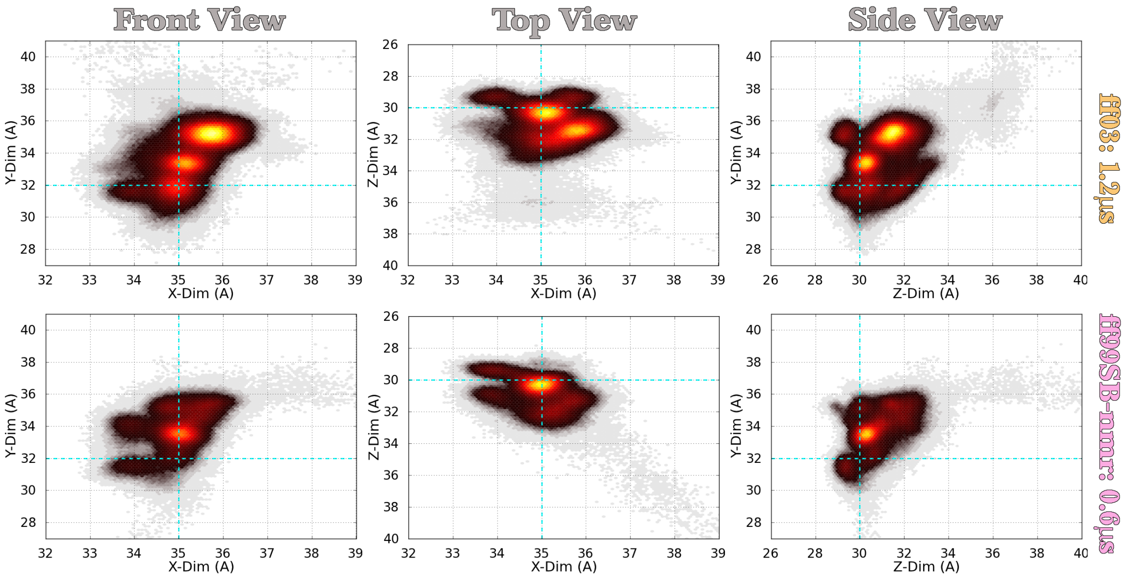


Fig.4.18. Comparison of the COM distributions of bound hep obtained with the ff03 (Top row) compared to the ffs99SB-nmr force field (Bottom row) sampled to 0.6 μs and 1.2 μs respectively. The ffs99SB-nmr simulation is normalised to half (5000 count) that of ff03 which has twice the number of data points. The comparison demonstrates the impact of sampling, and demonstrates the ligand retains considerable residual motion when bound despite differences in force field parameterisation.

range of residual ligand motion observed in MUP. Moreover, it is worth remarking that there is always a new force field being developed, but as the chief modifications are incremental improvements to torsional terms, it is extremely unlikely that the number of bound minima would suddenly be reduced to a single stable minimum. Considering MUP is a promiscuous binder, it is likely that its facility to accommodate a wide range of different shaped ligands would be accompanied by a lack of specificity. The predominantly apolar cavity has a dearth of strategically placed H-bond acceptor/donors that act to synchronously constrain the translational motions of the flexible, hydrophobic ligands tested in this study (§4.3.4).

4.3.4. The architecture of the binding pocket (clustering I)

Before considering the orientational entropy, it would be useful to describe the architecture of the binding pocket and the nature of ligand hydrogen bonding so that the structural factors that affect ligand dynamics are made clear (§4.34-5). An appreciation of these underlying factors and their relationship to dynamics provides a framework through which, the rationale for differences in 3Z-olefin and n-alkanol ligand entropy ($T\Delta S°_{Po}$ and $T\Delta S_{Or}$) can be better understood.

Highly populated bound minima in both panels demonstrate that the ligand's COM is offset at a variable distance from TYR120. The underlying 3D probability distribution was analysed via mean shift clustering and high density clusters were isolated (**Table.4.6**). The locations of identified clusters are numerically marked on the zy projection in **Fig.4.13-14.** Each cluster contained ligand binding poses localised to discrete minima within the binding site and visual analysis revealed that the COM displacements are amplified because the ligand is capable of adopting a variety of compact and extended conformations due to the flexibility of its hydrocarbon tail. Averaged PDB files were generated from every cluster and all ligand poses obtained in this manner were overlaid on a per-panel basis to give a conservative indication of the average positions of minima within the context of the binding pocket (**Fig.4.19**). Every pose thus represents the averaged ligand coordinates obtained from the ensemble of different ligand conformations, positions and orientations isolated from each high density cluster. Though internal flexibility results in localised fluctuations of ligand orientation and conformation, large changes to the average position of the molecule within each extracted cluster is minor. Thus, the image shows idealised poses that have the greatest probability of occurrence in a larger continuum of possibilities and does not show poses around the mean or those at the extrema of the total distribution. The overlay gives a very conservative indication of the range of bound positions to supplement the abstraction provided by the COM distributions (**Fig.4.11-14**).

For the purpose of discussion, it is convenient to subdivide the calyx into three smaller chambers: *cal1*, *cal2 & cal3* (**Fig.4.19**). The largest volume is possessed by *cal1* and this is linked to the proposed point of ligand entry/exit to the occluded binding site and the relationship between the two is further discussed in chapter 5.0. *Cal2* is bounded by TYR120, PHE56 and ALA103 and is smaller in volume than *cal1* due to the intrusion of these residues. The highly hydrophobic environment offered by *cal3* lies within the area situated near the base of the calyx and access to this area is partially occluded by the bulk of PHE90. However, in response to the presence of bound ligand this residue is capable of undergoing a conformational change whereby it swings open to press against the upper part of TYR80 so as to admit portions of the ligand into *cal3*. Chemical shift perturbations (CSP) obtained from NMR spectroscopy indicate that this residue shows one of the greatest ^{1}H ^{15}N changes on ligand binding (Data not shown). The conformational change was also noted in long MD simulations of IBMP bound to MUP by Roy et al. (2010) [91]. The isobutyl tail of IBMP possesses greater bulk than the linear alcohols examined in this work and it is possible that this structural disparity prevents IBMP from translocating to the very bottom of the calyx to a similar extent. In the case of n-alkanols, the staggered nature of the subcavities predisposes saturated ligands to adopt "curved" poses. In contrast, 3Z-olefins are pre-organised to the shape of the cavity and this is likely to facilitate a greater amount of positional displacement (**Fig.3.15-16** and **Fig.4.19**).

As detailed in chapter 3.0, the ligand's internal DOF are only marginally restricted when bound and it maintains the capability of adopting a variety of compact to

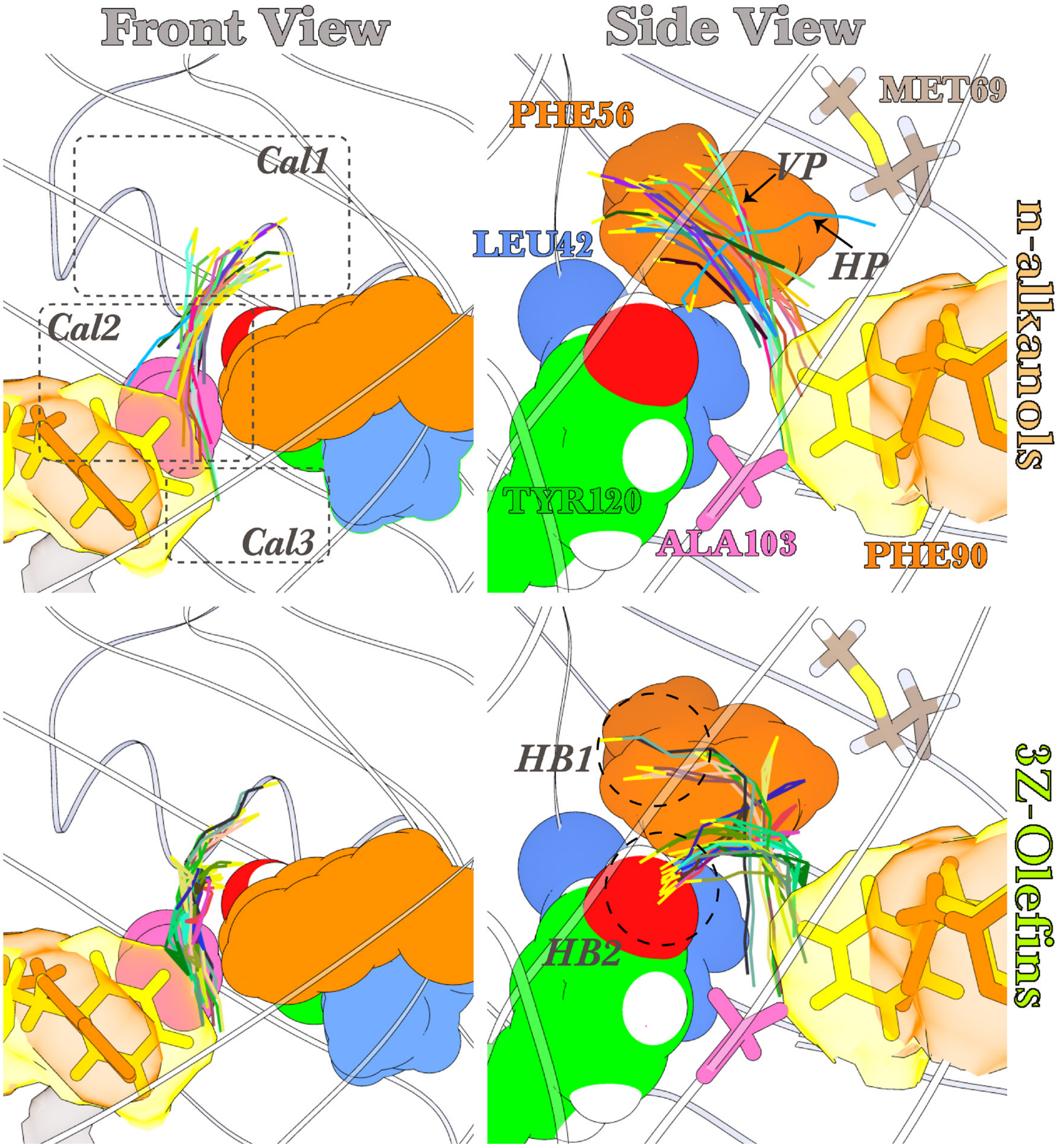


Fig.4.19. The range of ligand translational displacement within the architecture of the binding pocket is depicted by an overlay of 28 averaged ligand binding poses obtained from every ligand COM clusters identified by mean shift clustering. All frames in an identified cluster were averaged to yield averaged ligand coordinates and this gives a conservative indication of positional variation. Each ligand carbon chain is coloured differently, whilst oxygen atoms are coloured yellow. Subcavities are demarcated and labelled as *cal1, cal2 & cal3* in the top left. Examples of vertical and horizontal pose are marked as VP and HP respectively (top right). Longer n-alkanols are distinguished by preferential adoption of vertical poses, whilst 3Z-olefins fluctuate between the two. There are also two principal H-bond localities associated with the poses (HB1 & HB2). n-alkanols predominantly bind to the HB1 zone, whilst 3Z-olefins preferentially H-bond to HB2. The closed and open conformations of PHE90 are overlaid and coloured yellow and orange respectively.

extended conformations. However, visual examination of the simulations and the averaged cluster structures indicate that the ligand adopts two characteristic poses that can range from being roughly perpendicular to the y axis or parallel to it. Examples of both can be observed for oct bound to MUP and are labelled as a vertical pose (VP) and horizontal pose (HP) respectively (**Fig.4.19**). HPs tend to occupy *cal1* and upper portions of the *cal2* subcavities because the volume available in these zones is large. Ligands of all lengths are capable of HPs, but shorter ligands have a greater propensity to adopt this whole-molecule orientation as their smaller size results in a reduced chance of getting trapped by residues near the middle of the calyx. This type of pose has the potential to be associated with ligand entry/exit if the ligand is not sufficiently restrained by H-bond interactions near TYR120 (chapter 5.0). VPs generally result in the ligand being trapped by residues in *cal2 & cal3* whilst the hydroxyl group is usually held by H-bonds. Note that 3Z-olefins can jump between these two poses more readily than n-alkanols for reasons that will be discussed in §4.3.7.

The densities shown in **Fig.4.11-14** chart the displacement of the ligand's COM as it varies its position within the pocket whilst the hydrophobic chain transitions between a farrago of compact, intermediate and extended conformations. This is unsurprising as the difference between dihedral angle distributions for free and bound states were minimal (**Fig.A1.1-3)**. As the hydroxyl head is (usually) restrained by H-bonds, high density minima should correspond to positions in the pocket close to the location of these polar interactions. Interestingly, the cluster analysis indicates that polar hydroxyl group of the linear alcohols can broadly be categorised as occupying two areas. The first (HB1) positions the alcohol moiety close to TYR120 and ALA103, whilst the second (HB2) is located higher along the y axis than the former. This can most clearly be observed in the side view of clustered 3Z-olefin structures (**Fig.4.19**). In the predominantly apolar cavity, the identity of potential protein hydrogen partners is important as this interaction plays a vital role in localising the ligand within the cavity. Hence, this line of enquiry will be investigated in detail within the next section.

4.3.5. Hydrogen bond interactions within the calyx

It is well known that H-bond strength is distance and angle dependent (§1.2.2) and that these interactions play a structural role that usually contributes favourably to the binding affinity. A detailed analysis of H-bond interactions beyond examining patterns in crystallographic structures has not been published to date and prior research has emphasised the role of TYR120 as the principal interacting partner [36,90,136,180]. Most of the binding poses indicate that the ligand's alcohol moiety is close enough to directly, or indirectly be involved in H-bond interactions with TYR120. However, localisation to other areas is also observed and the COM of the ligand can occupy regions that are near the apex of *cal1* or even within the hydrophobic reaches of *cal3*. This is somewhat surprising and the conclusion that must be drawn is that at certain junctures the ligand is H-bonded to other residues or not involved in any such polar interactions at all. IBMP was also reported to display significant residual motion within the cavity and occupied poses that disallowed H-bonding with TYR120 [91,180]. Whilst the linear alcohols have the potential to move around all localities within the pocket, the highest populated COM minima lie within *cal1* and *cal2* subcavities as these areas presumably provide more H-bonding opportunities than *cal3*.

Published crystal structures of n-alkanols bound to MUP indicate that that these linear ligands can be involved in a direct H-bond interaction to TYR120 or are indirectly linked via bridging water molecules [36,90,136,180]. Additional H-bond interactions are also observed with various amino acid backbone atoms such as PHE38 and LEU40. Analysis of the simulation data indicates that the intermolecular H-bond network is more subtle than that implied by the static picture provided by crystals as the dynamic nature of bound compounds within the calyx results in the breaking and reforming of a greater number of these polar interactions than originally anticipated. All frames (1.2x 10^6) from each concatenated trajectory were analysed for the fraction of time H-bonds occurred during the simulation. This allowed separation of H-bonds made directly to the ligand and those facilitated via the presence of bridging water molecules. Direct interactions are discussed below and water mediated H-bonds are discussed in the next chapter (§5.3.4.4). *Note that the percentage occupancies only reflect the probability of that bond existing within the simulation and do not indicate whether the bond exists synchronously with bonds made with other potential H-bond partners.* Tables 4.8-9 demonstrate that n-alkanols predominantly bind to backbone atoms of residues (LEU40 and PHE38) situated within the large Ω loop; a structural feature that functions to form a lid over the open maw of the calyx (**Fig.1.20**). These interactions can readily be observed in crystal structures of MUP (e.g. 1ZND, 1ZNE, 1ZNH & 1ZNK), but it is worth noting that the single structure typically produced by x-ray crystallography is an averaged conformation that is affected by a multitude of other factors and does not provide a comprehensive picture regarding the H-bond network which is very likely to morph in sympathy with protein and ligand dynamics (§A2.1.2). TYR120 is also identified as an H-bonding partner, but the fraction of its occupancy throughout the time course of the simulation is significantly less than that observed for the residues located within the Ω loop. In contrast, 3Z-olefins principally bind to TYR120 and the backbone carbonyl atom of ALA103, though there is a much reduced probability of forming bonds to the Ω loop in the case of 3c7, 3c8 and 3c9. Furthermore, due to significant ligand residual motion ligands from both panels are capable of directly forming a multitude of other transient interactions to alternative residues, albeit with occupancies < 1% (not tabulated). The protein atoms involved in these interactions are principally located within the backbone of the β-clam fold and are rarely unmasked by the bulk of obstructing sidechains as the protein undergoes conformational change. When this is juxtaposed with the forces that drive the ligand to adjust its position and orientation within the shifting architecture of the calyx, the favourable confluence of angle and distance required to form a successful

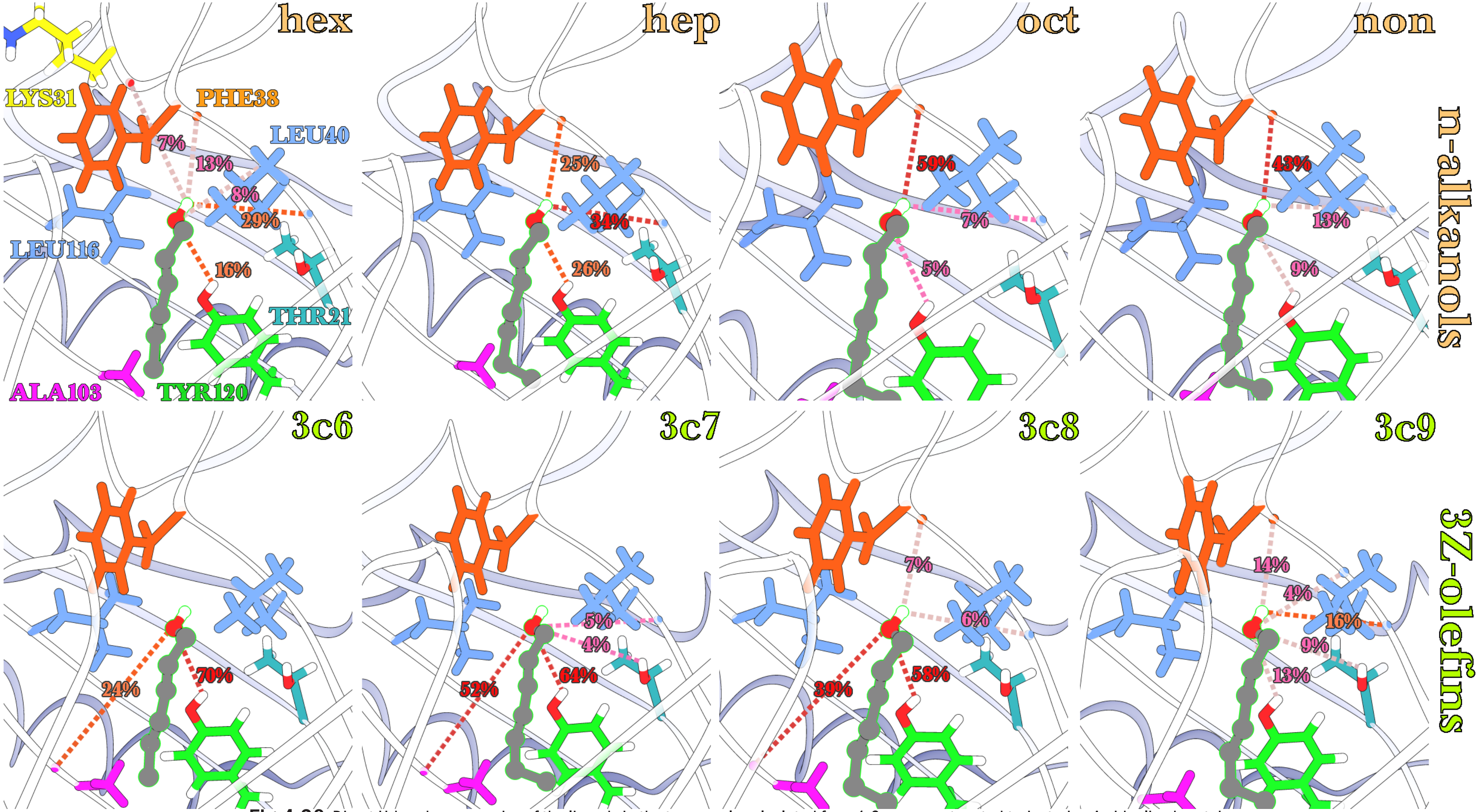


Fig.4.20. Direct H-bond occupancies of the ligands in the two panels calculated from 1.2 μs concatenated trajectories. An idealised protein conformation and ligand pose is depicted to facilitate ease of comparison. Figure is labelled with H-bond percentage occupancies.

H-bond interaction with these partners is severely limited. More importantly, there is a correlation between ligand length and direct H-bond occupancies (**Fig.4.20**).

4.3.5.1. n-alkanol direct H-bonds

All ligands H-bond to TYR120 to some extent, but there are a greater number of interactions to other residues within the Ω loop and some of these have higher probabilities of being involved in H-bonds with the ligand than the intuitively obvious candidate. As the ligand increases in size, the residue with the largest H-bond occupancy shifts from LEU40 which lies lower on the loop, towards PHE38, which is higher and more centrally positioned. The simple rationale underlying this trend can be explained by the following observations (**Fig.4.20**): The smallest ligand hex, spends the majority of simulated time in *cal1* by virtue of its small size and thus makes the greatest variety of interactions with the Ω loop. Consequently, the dynamics of the loop result in this ligand being displaced to the greatest extent compared to the other ligands. Thus, it possesses the largest magnitude of $TS^{\circ}_{Po\bullet B}$ because the shorter hydrophobic tail is much less likely to be hindered by residues situated in *cal2 & cal3* (**Table.4.8**). Increasing the length of the ligand by a single methylene group results in more protein-ligand interactions which act to transfix the hep to a greater extent and counteract the residual translational motion stimulated by Ω loop motions. Therefore, hep possesses roughly equal probabilities of forming H-bonds with residues within the Ω loop and lies near the midpoint of the described trend. *Per contra*, the two longer ligands (oct and non) are disposed to adopting fewer, better defined translational minima because they are more likely to be displaced into *cal2 & cal3* and getting trapped by residues within that region (**Fig.4.20**). Their larger size also allows them to make increased polar interactions to residues near the apex of the calyx. The interactions with the Ω loop are especially interesting because these residues are also associated with ligand internalisation and externalisation (Chapter 5.0).

4.3.5.2. 3Z-olefin direct H-bonds

The cis-3-4 restriction introduces a curvature into these unsaturated ligands and this feature allows them to adopt positions lower down along the y axis so that H-bonds are principally formed to TYR120 and ALA103 (**Fig.4.19** and **Fig.4.10**). Visual examination of the simulations indicates that interactions between these residues often occur synchronously due to their close proximity. These twin restraining forces cause the polar hydroxyl head of the ligand to be more firmly localised within the vicinity of TYR120 and thus, there is also a greater probability of interaction with THR21 which lies nearby. As will be discussed in §4.3.7, the principal driver of residual translational motions in these unsaturated ligands is due to rigid body displacements of the hydrophobic tail and the complementary shape of these ligands to the binding cavity. Indeed $TS^{\circ}_{Po\bullet B}$ for all these compounds were near identical (**Table.4.7**). As

	Direct hydrogen bonding: n-alkanols					
				(%)	**(ps)**	**(ps)**
	Acceptor	**Donor-H**	**Donor**	**Occup**	**Avg LT**	**Max LT**
hex	LEU40@O	hex@H7	hex@O1	29.1	7.2	183
	hex@O1	TYR120@HH	TYR120@OH	15.5	2.1	43
	PHE38@O	hex@H7	hex@O1	12.6	5.2	104
	hex@O1	LEU40@H	LEU40@N	7.6	1.6	22
	LYS31@O	hex@H7	hex@O1	7.3	2.8	86
	Unique Partners under 4% occupancy = 40					
hep	LEU40@O	hep@H9	hep@O1	34.6	8.1	207
	hep@O1	TYR120@HH	TYR120@OH	26.6	3.8	50
	PHE38@O	hep@H9	hep@O1	25.1	4.4	103
	Unique Partners under 4% occupancy = 40					
oct	PHE38@O	oct@H9	oct@O1	59.4	5.4	123
	LEU40@O	oct@H9	oct@O1	6.9	5.2	257
	oct@O1	TYR120@HH	TYR120@OH	5.0	3.1	47
	Unique Partners under 4% occupancy = 26					
non	PHE38@O	non@H10	non@O1	43.1	4.9	99
	LEU40@O	non@H10	non@O1	12.8	7.2	151
	non@O1	TYR120@HH	TYR120@OH	8.5	2.6	36
	Unique Partners under 4% occupancy = 49					

Table.4.8: n-alkanol direct H-bond interactions tabulated using AMBER notation. H-bond lifetimes (LT) are measured in picoseconds and the maximum lifetime assesses the longest contiguous period a particular H-bond is present. The average H-bond LT is calculated by averaging all contiguous periods that an H-bond is present in the entire trajectory. See p271 of AMBER 14 manual for further details. For reference, H-bond lifetimes in pure water have been measured to be 1 to 20 ps [376].

the ligand gets larger, the probability of direct H-bonds with residues in the Ω loop increases, albeit with much lower occupancies than that observed for their saturated analogues. The proposed rationale for residual translational motion for members in both panels would seemingly be contradicted by the direct H-bond occupancies of the longest ligand, 3c9. Like n-alkanols, it only binds to TYR120 and Ω loop residues, yet still manages to maintain significant residual motion. However, this can easily be reconciled through the observation that the occupancies of these interactions are much less than that observed for hex. It is likely that, as the longer length of 3c9 results in the hydrophobic tail being more constrained within *cal2 & cal3*, the motions of the *H-bonded segment become amplified* by rigid body displacements and shifts in position are additionally abetted by low occupancy interactions with the dynamic Ω loop. Further

	Direct hydrogen bonding: 3Z-Olefins					
				(%)	**(ps)**	**(ps)**
	Acceptor	**Donor-H**	**Donor**	**Occup**	**Avg LT**	**Max LT**
3c6	3c6@O1	TYR120@HH	TYR120@OH	70.3	5.2	134
	ALA103@O	3c6@H7	3c6@O1	23.5	4.7	293
	Unique Partners under 4% occupancy = 8					
3c7	3c7@O1	TYR120@HH	TYR120@OH	64.1	4.6	113
	ALA103@O	3c7@H8	3c7@O1	51.7	5.2	121
	LEU40@O	3c7@H8	3c7@O1	5.0	11.2	299
	3c7@O1	THR21@HG1	THR21@OG1	4.1	7.9	111
	Unique Partners under 4% occupancy = 20					
3c8	3c8@O1	TYR120@HH	TYR120@OH	58.1	4.0	83
	ALA103@O	3c8@H9	3c8@O1	38.8	4.2	122
	PHE38@O	3c8@H9	3c8@O1	7.1	3.4	49
	LEU40@O	3c8@H9	3c8@O1	6.0	3.0	298
	Unique Partners under 4% occupancy = 19					
3c9	LEU40@O	3c9@H10	3c9@O1	15.7	7.7	258
	PHE38@O	3c9@H10	3c9@O1	13.8	3.1	50
	3c9@O1	TYR120@HH	TYR120@OH	13.2	3.8	144
	3c9@O1	THR21@HG1	THR21@OG1	9.2	6.0	129
	3c9@O1	LEU40@H	LEU40@N	4.1	1.5	20
	Unique Partners under 4% occupancy = 28					

Table.4.9: 3Z-olefin direct H-bond interactions tabulated using AMBER notation. H-bond lifetimes (LT) are measured in picoseconds and the maximum lifetime assesses the longest contiguous period a particular H-bond is present. The average H-bond LT is calculated by averaging all contiguous periods that an H-bond is present in the entire trajectory. See p271 of AMBER 14 manual for further details. For reference, H-bond lifetimes in pure water have been measured to be 1 to 20 ps [376].

evidence for this proposition is provided in §4.3.7. There is an obvious difference in H-bonding patterns between n-alkanols and 3Z-olefins, and it is extremely likely that the reason for the more diffuse COM distributions of 3Z-olefins (**Fig.4.13-14**) is due to presence of the double bond which facilitates:

1. Ligand interactions with a completely different part of the protein to n-alkanols.

2. The promotion of radically different modes of internal and external dynamics to n-alkanols (§4.3.7).

This Page left intentionally blank

4.3.6. Orientational entropy: The number of minima

At this point it has been established that all ligands tested possess significant residual motion when bound and the positional entropy has been quantified. Moreover, the structural environment that houses bound compounds has been described and correlations between differential H-bonding occupancies have been made with ligand structure and length. The correlation has allowed intermediate hypotheses regarding the root cause of increased COM positional displacements to be made (§4.3.5) and the next two sections examine the orientational entropy and how it aids our understanding of ligand dynamics within MUP (§4.3.6-7).

In order to do this, 1st order orientational entropies of every bond were calculated with the method described in §4.2.5.3 using two different approaches (**Fig.4.22-23** & **Table. A2.2.1-2**). The 1st approach averages the per-bond entropies obtained from six 200 ns trajectories (top 2 rows) and the 2nd approach applies the same calculation to a single concatenated 1.2 µs trajectory (bottom 2 rows). Hammer projections depicting the orientational movements of individual bonds from the concatenated 1.2 µs trajectories (2nd approach) are located in the appendix **Fig.A2.2.1-4.** They demonstrate that each bond can take up all possible orientations when free, while certain orientations are forbidden when the ligand is bound. Thus, the orientational entropy for all bonds are the same in the free state ($TS_{Or\bullet F}$) and the resulting trend line is flat, possessing zero variance and error (**Fig.4.22-23.a** & **Fig.A2.2.1-2**). There is a greater amount of variation in bound per-bond orientational entropies ($TS_{Or\bullet B}$) and this propagates to the resulting per-bond entropy differences ($T\Delta S_{Or}$). The disparities in resulting trend lines between the congeneric compounds constituting both panels reveal considerable detail on ligand dynamics.

The 1st approach yields per-bond trend lines for individual 200 ns repeats and $TS_{Or\bullet B}$ values for any given bond span a gamut of values in the bound state (See dotted lines in **Fig.4.22-23.a**). This variability must be the result of interactions with the particular subcavity the ligand is localised within and is consequently correlated with the COM distributions. The trend lines described by each repeat are generally consistent with one another, but some repeats exhibit marked deviations from the majority. For example, in the case of 3c9: The trend line with the lowest per-bond $TS_{Or\bullet B}$ is obtained from repeat-4, whilst repeat-6 possesses the greatest entropies. Additionally, the shape of the latter is less rugged than its counterparts. Analysis of the COM translational distribution of repeat-6 indicates that the ligand accesses a greater variety of minima than it does during repeat-4 which remains localised within a single minima for the majority of (the 200 ns) simulated time (**Fig.4.21**). Visual inspection of the trajectories indicate that repeat-4 favours a vertical pose which occupies *cal1*, *cal2 & cal3*, whilst repeat-6 fluctuates between vertical poses and horizontal poses localised in *cal1*. The bottom row in **Fig.4.21** depicts the orientational distribution of the C3-C4 bond for both the repeats and comparison to the COM distributions demonstrates that there is a correlation between the two and this will be utilised in the following sections to elucidate further detail on ligand dynamics. The relationship between bond orientation and location may be self-evident, but it is worth explicitly stating, because it highlights the fact that the various locales housed within the architecture of the calyx are capable of driving ligand dynamics and behaviour in distinct ways that can potentially be separated from one another by clustering. Hence, positional and orientational analyses provide a link between structure and dynamics; factors that must be taken into account when structurally modifying drug-like compounds.

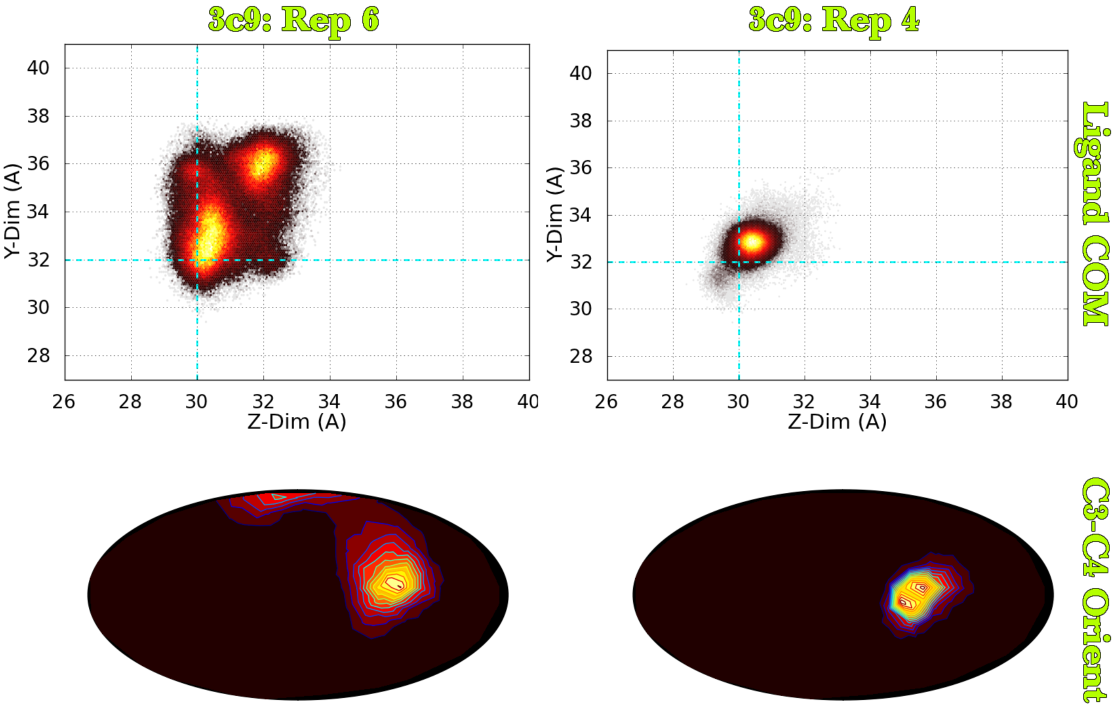


Fig.4.21. The concatenated 1.2 μs simulations are composed of 6x 200 ns independent simulations and the distributions from two of these are depicted here for comparative purposes. The top row shows the COM distributions of bound 3c9 viewed along the zy projection. As shown in **Fig.4.14**, The ligand is capable of much greater exploration of the calyx with additional sampling. However, in individual repeats the ligand can potentially remain trapped within favourable minima for periods of time that many researchers would consider sufficient for analyses e.g. the ligand is localised to a greater extent in repeat-4 compared to repeat-6 within a span of 200 ns. The bottom row uses hammer projections to depict the orientational distribution of the ligand's C3-C4 bond to reveal the relationship with the COM distribution.

In quantitative terms, entropies calculated from the single concatenated 1.2 μs trajectory (2nd approach) *are not averages* of the entropies calculated from its constituent 200 ns segments (1st approach). Entropies are dependent on the manner in which the total available dataset is evaluated and the calculation yields larger entropies for concatenated 1.2 μs trajectories compared to that generated by averaging entropies obtained from 6x

200 ns trajectories. As discussed in §3.1.1, any entropy calculation evaluates the density of states of a given observable and correct estimation is dependent upon total accessible phase space being fully represented. The number of orientational minima captured by discrete 200 ns trajectories is sufficient to capture patterns and trends in the motions of the bonds that constitute the ligand. However, as the entropy calculation is only performed on a subset of accessible phase space, the values returned are underestimated compared to entropies calculated from a PDF that better approximates the ergodic distribution by virtue of increased sampling i.e. the ergodic entropy cannot be recreated by averaging subsets of the data in a similar manner to simulation temperature or pressure because (the classical) entropy is dependent upon the density of states and all accessible minima have to be considered simultaneously. For example, the distribution of an initial 200 ns trajectory may only possess one or two orientational minima within the bound state. As additional 200 ns segments are added to the calculation, the population of the existing minima are expanded and new ones discovered. This is characterised by a large increase in $TS_{Or\bullet B}$ as the bond vector "jumps" between increasing numbers of minima. After a point, sampling is sufficiently large to identify all available minima and $TS_{Or\bullet B}$ oscillates around an increasingly stable value as the relative occupancies of bound minima fluctuate. Convergence is only achieved when the relative populations of existing minima have stabilised and no new minima remain to be uncovered; these conditions are hard to achieve due to the existence of rare events. Thus, the larger entropies obtained for 1.2 μs concatenated trajectories are the direct result of a greater number of minima being simultaneously evaluated by the calculation. The larger (1.2 μs) dataset yields an entropy increase of ~0.5% and ~19% for free and bound bond vectors respectively. There is greater difference in $TS_{Or\bullet B}$ because higher energy barriers separate the minima that make up its orientational distribution compared to that encountered when the ligand is free in solution. Identical starting coordinates were used as the genesis of each 200 ns trajectory and different random number seeds provided the impetus for the unique evolution of each discrete simulation. When processing power is limited, this approach is likely to yield much better sampling than a single simulation that is six times as long. This is because overall, an increased number of independent shorter simulations are stochastically less likely to all be trapped within identical minima (**Fig.4.21**). Thus the concatenated trajectory is very likely to possess a greater variety of minima than a single trajectory that is six times as long. Note that this advantage is likely to disappear in the case of extremely long simulations and these observations are made with reference to ligand dynamics i.e. multiple short simulations may not efficiently capture protein dynamics that occur on longer timescales [91].

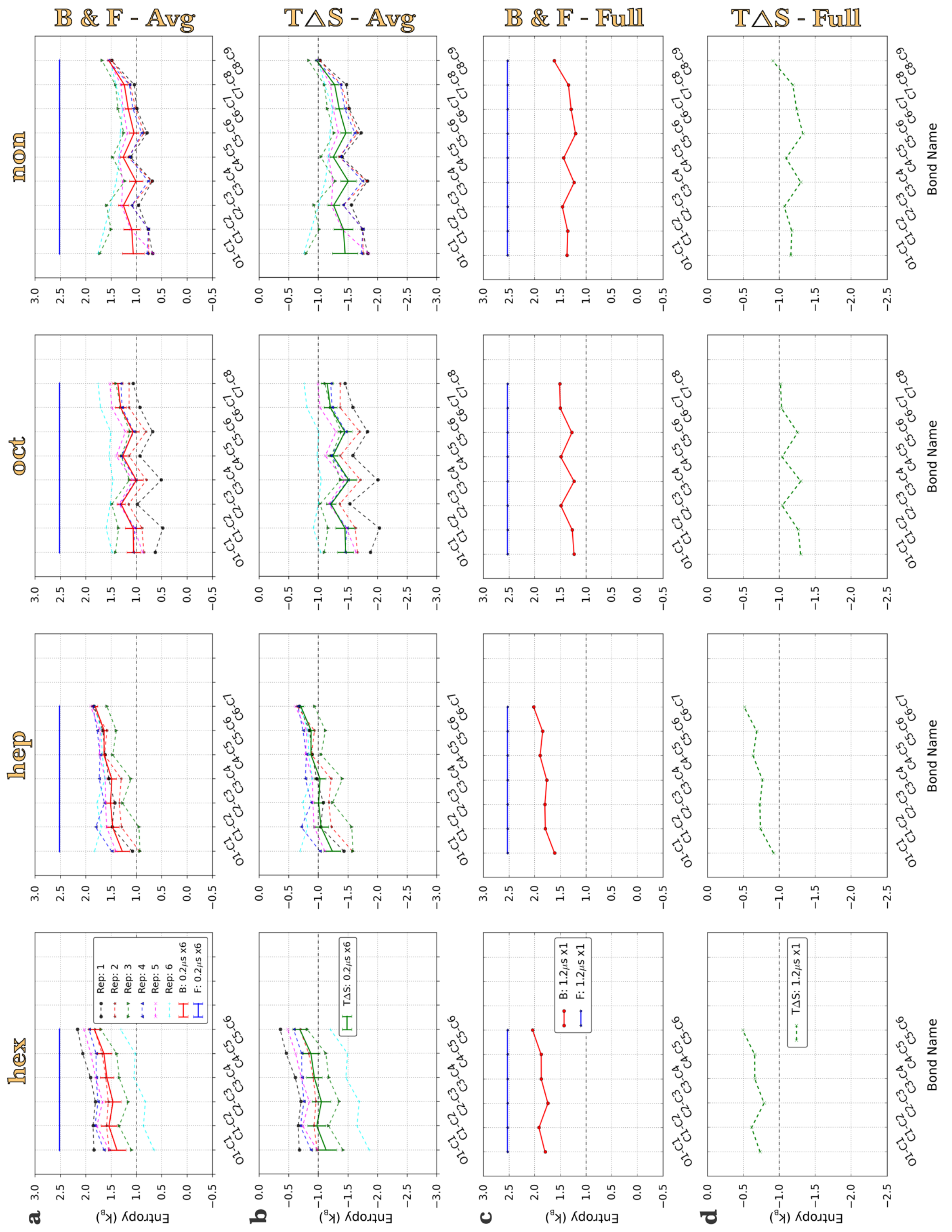


Fig.4.22. Per-bond rotational entropies calculated for 3Z-olefin ligands. **(a-b)** The top two rows depict averaged entropies calculated from six independent 200 ns simulations. Individual repeats are shown as coloured dashed lines. The average for the free species is a solid blue line, whilst bound state is represented by a solid red line and the relative entropy difference is a solid green line. Errors are shown as standard errors. **(c-d)** The bottom two rows display entropies calculated from a single concatenated 1.2 μs trajectory. The magnitude of the calculated values is greater than that obtained by averaging as the density of states that describe ligand phase space is better approximated (§4.3.6). n-alkanols have greater probabilities of being displaced into *cal2* & *cal3* as ligand length increases. This is accompanied by a characteristic alternating pattern **(d)** that arises from crankshaft motions (§4.3.7.1). Also note the difference in per-residue $T\Delta S_{0r}$ baseline separating the two shorter ligands from longer ones.

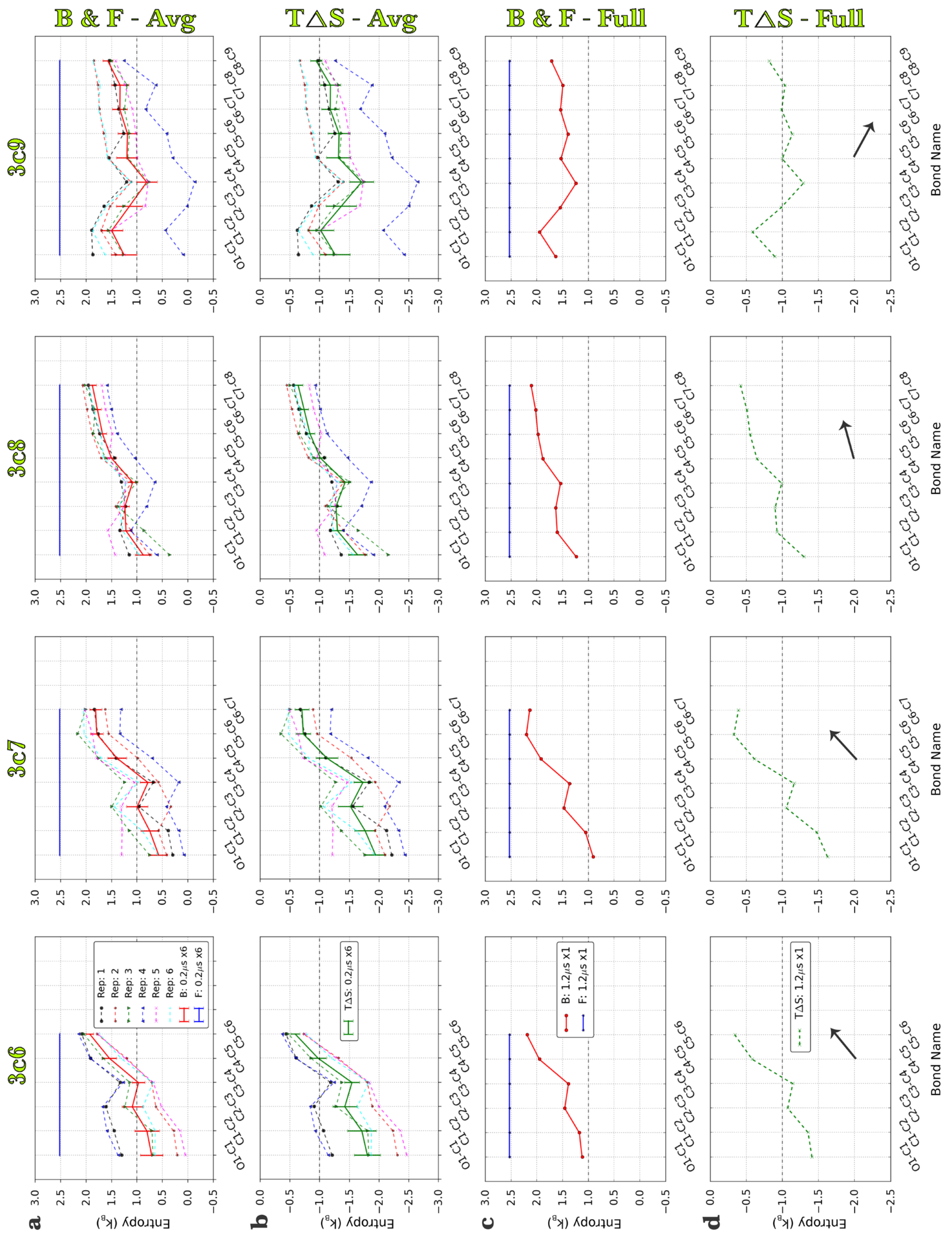


Fig.4.23. Per-bond rotational entropies calculated for 3Z-olefin ligands. **(a-b)** The top two rows depict averaged entropies calculated from six independent 200 ns simulations. Individual repeats are shown as coloured dashed lines. The average for the free species is a solid blue line, whilst bound state is represented by a solid red line and the relative entropy difference is a solid green line. Errors are shown as standard errors. **(c-d)** The bottom two rows display entropies calculated from a single concatenated 1.2 μs trajectory. The magnitude of the calculated values is greater than that obtained by averaging as the density of states that describe ligand phase space is better approximated (§4.3.6). The black arrow in the bottom panel indicates the increasing probability of the terminal segment becoming restrained within *cal2 & cal3* as ligand length increases. In the case of longest ligand, 3c9, crankshaft motions begin to occur within the terminal segment in a similar manner to that observed in longer n-alkanols (§4.3.6).

4.3.7. Trends in per-bond Orientational motion

The analysis of per-bond $T\Delta S_{Or}$ trends in this section allows conclusions to be made regarding ligand dynamics, and additional rationale for panel-specific differences in global entropic binding signatures are proposed. The following observations are based on the more ergodic per-bond orientational entropy differences ($T\Delta S_{Or}$) obtained via the 2nd approach because they are a better representation of accessible phase space (**Fig.4.22-23.d**).

4.3.7.1. n-alkanol orientational trends:

1. Ligands in the panel can broadly be divided into two categories based on the relative magnitude of measured per-bond TDSOr. Hex and hep possess larger entropies ranging between -1.0 to -0.5 kJ/mol, whilst oct and non lie between -1.5 and ~ -1.0 kJ/mol.

2. In the majority of compounds, the bond that possesses the highest value of $T\Delta S_{Or}$ belongs to the terminal methyl. This is because this apolar moiety is connected to a single neighbour and it experiences relatively little restriction compared to the polar head of the ligand which has a propensity to form H-bond interactions.

3. On increasing ligand length a characteristic “alternating” pattern near the centre of the molecule is observed. The pattern is similar to that obtained from per-dihedral analysis and reports on crankshaft motions that occur in long linear chains confined within small volumes (§3.3.6.2 & **Fig.3.13**). Increases in per-dihedral $T\Delta S_{In}$ (within C3-C4 and C5-C6) were associated with increased conformational transitions between trans and gauche states. However, comparable increases in $T\Delta S_{Or}$ are associated with increased movements in C2-C3 and C4-C5 bonds and the peaks and troughs reporting on crankshaft motions are inverted (**Fig.4.22.d**). The apparent anti-correlation is the result of differences in the method by which the two subsets of the entropy are measured (A2.1.3).

4. Adding methylene groups increases $T\Delta S_{Or}$ of some of the bonds near the centre of these linear saturated ligands. Hence, per-bond entropy-entropy compensation ameliorates summed $T\Delta S_{Or}$ losses in a manner that would *suggest* that they are not linearly correlated to the addition of a methylene group. Note that the conclusions obtained from PCA in §3.3.6.1-2 still hold true with regards to the correlated motions of n-alkanols (**Fig.3.15**). On binding, large coordinated wagging motions of the terminal segments of the molecule are replaced by alternating small amplitude motions throughout the body of the ligand. As COM distributions indicate that longer ligands occupy a reduced number of increasingly defined translational minima (**Fig.4.13**), it can be surmised from the averaged binding poses (**Fig.4.19**) that the pronounced alternating pattern represents crankshaft motions that arise as a result of protein-ligand interactions within the crowded confines of *cal2 & cal3*. Further evidence for this proposition is provided in §4.3.9.3.

4.3.7.2. 3Z-olefin orientational trends:

1. The trend lines described by per-bond $T\Delta S_{Or}$ in most ligands can be conceptually subdivided according to whether the bonds lie before or after the C3-C4 double bond. These will be referred to as the hydroxyl segment and terminal segment respectively. The hydroxyl group of the ligand is engaged in H-bond interactions that act to restrain the head of the ligand and thus, orientational motions of O1-C1, C1-C2 and C2-C3 bonds are curtailed. 3c9 is the exception to this rule for reasons that will shortly become apparent. The bonds constituting the terminal segment in most 3Z-olefin ligands have relatively greater orientational freedom than those in analogous n-alkanols for a several reasons. As the terminal segment grows from an ethyl through to a butyl chain, the additional methylene groups create a gradient of steric restraints that is at its maxima at the C3-C4 restriction. The two segments possess markedly different behaviours that are delimited by the position of the restriction which can be thought of as a fulcrum or tipping point. Extending the shortest terminal segment (possessed by 3c6) with additional methylenes, results in a general overall reduction to the orientational freedom of bonds constituting the terminal segment, whilst bonds composing the hydroxyl segment generally undergo entropic increases relative to the previous ligand in the series (**Fig.4.23.d**).

2. In addition to steric considerations which arise from internal interactions within the ligand itself, there are also external factors related to whether the terminal segment is located within *cal1,* or *cal2 & cal3*. If it is in the former subcavity, values for $T\Delta S_{Or}$ will be larger as the environment provided by this subcavity is less crowded. On the other hand, if it is positioned within the latter subcavity, the magnitude of $T\Delta S_{Or}$ values will be smaller because the surroundings are more crowded. Thus, the values depicted in **Fig.4.23.d** are an average of the disparate trends resulting from localisation within both of these zones (§4.3.6). As more methylenes are added after the double bond, the increasing size of the terminal segment invariably increases the probability of it being displaced into *cal2 & cal3,* and becoming more restricted within the pocket. This hypothesis is supported by the decreasing per-bond orientational entropies of the bonds constituting the terminal segment. This immobilisation results in a greater amount of compensatory orientational dynamics in the hydroxyl segment which is at its maximum in the case of bound 3c9 (**Fig.4.23.d**).

3. Before reading this point the reader is encouraged to view **video 4.1** that is included in the supporting materials. In chapter 3.0, the torsional data demonstrated that the C2-C3 and C4-C5 dihedrals predominantly occupy gauche(+) and gauche(-)

conformations similar to that observed in the free state and chemical knowledge tells us that these rotate, so as to minimise steric clashes with each other. Internal motions cannot propagate along the body of the ligand in a similar manner to that observed in n-alkanols because the rigid cis-3-4 double bond acts to decouple the segments that lie on either side of it and the *internal dynamics* of the terminal segment are to a large extent (mostly) *independent* of the hydroxyl segment. In the case of flexible n-alkanols the same C3-C4 saturated bond is in a state of conformational flux between gauche and trans states and the lack of internal rigidity predisposes longer n-alkanols to favour fewer, better defined translational and orientational minima than their unsaturated counterparts. Thus, the flexible C3-C4 bond acts as a conduit through which *internal motion can be propagated* from one termini of the ligand to the other because each unit in the chain is influenced by the steric requirements of its neighbours. However, the *rigid restriction within 3Z-olefins acts as a sink that disallows such propagation* and as the hydroxyl head of the ligand is (usually) restrained by H-bonds, the ligand's internal motions are transformed, and manifest in the form of external rigid body displacements of *the terminal segment* of the hydrocarbon tail which is unrestrained by any such polar interactions. If oct and 3c8 are taken as contrasting examples, it is clear that the terminal segment (C3-C4 to C7-C8) of 3c8 has greater ability to "switch" between orientations so that the ligand alternates between horizontal and vertical poses (**video 4.1**). It is likely that the switching mechanism is focussed on the C3-C4 bond and its fluctuations are chiefly responsible for a concomitant switch in orientations of the bonds constituting the terminal segment. Hammer projections comparing the orientational motions of bond vectors show that 3c8 possesses a greater spread of orientational minima compared to oct within the time simulated (**Fig.4.24**). This is further supported by the COM density maps whose intensities indicate that the 3Z-olefins possess a more diffuse spread of positional minima than n-alkanols (**Fig.4.13-14** & §4.3.3.1).

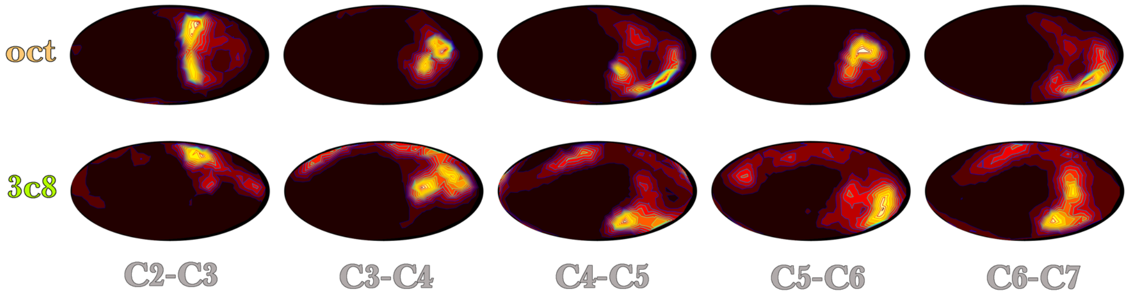


Fig.4.24. Hammer projections of the orientational distribution of the hydrophobic tail of oct and 3c8 obtained from 1.2 μs trajectories. The projection is a flattened 3D density distribution calculated from the motions made by the terminus of the bond vector as it moves across the surface of a sphere after translation has been removed. Some orientations are disallowed. The C3-C4 bond and the other bonds that constitute the terminal segment of 3c8 possess a greater variety of allowed orientations than the corresponding bonds in oct. These additional states account for reduced $T\Delta S_{Or}$ penalties in the case of unsaturated compounds. Larger images can be found in §A2.2.3-4.

As these distributions do not account for the frequency of bond switching, it is not possible to make a conclusion as to whether oct remains in stable states for longer periods of time than 3c8. In order to get a measure of this, the fluctuations in distance of the C3-C4 bond's COM from the origin (first frame) were calculated and plotted with respect to time (**Fig.4.25**). This bond is implicated in the switching of the ligand from vertical poses that locate the compound within *cal2 & cal3* to horizontal poses those that position it within *cal1*. Note that the change in distance baseline is a good marker for the frequency of (bond) orientational changes. The starting structure for both the oct and 3c8 complexes, position the ligand in the vertical pose so that it is occupies *cal1*, *cal2 & cal3*. As the simulation of bound oct evolves, the distance of the C3-C4 bond from its starting position oscillates around a stable average (~1.0 Å) and the ligand undergoes localised conformational changes within a defined minima. On occasion (~50 ns), oct escapes the grasp of the residues situated in *cal2 & cal3* and adopts horizontal poses that occupy *cal1*. This is reflected in the jump in the C3-C4 baseline which again undergoes localised distance fluctuations around ~2.0 to 3.0 Å. In contrast, the rigid C3-C4 double bond in 3c8 fluctuates wildly in terms of both localised and baseline variations. This proves that the C3-C4 bond in 3c8 switches between minima with higher frequencies than oct which tends to dwell within stable minima for longer periods of time. This phenomenon is further explored in §4.3.9.2-3.

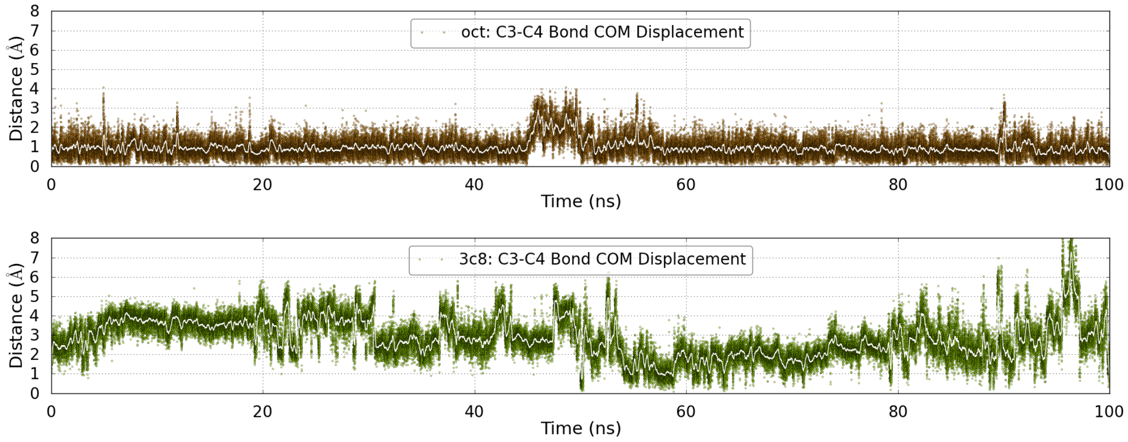


Fig.4.25. Frequency of C3-C4 bond switching measured via COM Displacements (Å) of the bond for oct (top) and 3c8 (bottom) over a representative period of 100 ns. A rolling average (white line) was calculated using a 2 ns window.The result shows that the C3-C4 bond in 3c8 switches between minima with higher frequencies than oct which tends to dwell within stable minima for longer periods of time.

4. Another factor that is likely to contribute to increased dynamics of 3Z-olefins is the shape and structure of the ligand. The double bond introduces a kink into unsaturated ligands that allows them to occupy states lower down along the y axis as the end-to-end length is shortened (**Fig.3.8** & **Fig.4.10**). This also accounts for the differences H-bond occupancies (§4.3.5) observed in 3Z-olefins, compared to n-alkanols. As the double bond structurally predisposes the centre of the ligand to be offset from its termini, a greater variety of compact conformations can be adopted. Moreover, when the hydroxyl segment is restrained by H-bonds, this "curved" shape is likely to exaggerate rigid body displacements of the terminal segment that are transversal and out of plane to the longitudinal axis of the ligand i.e. the methylene groups (C2 and C5) flanking the C3-C4 bond adjust their conformations to avoid steric clashes and the satisfaction of these internal constraints result in relatively small torsional adjustments promoting the external displacement of the entire terminal segment which is unrestrained by any polar interactions. Hence, the "curved" shape and cis-3-4 structure facilitate the "switching" of C3-C4 bond between the two main orientational minima that correspond to horizontal and vertical poses.

5. The factors described in the preceding four points result in a different mode of entropy-entropy compensation to that seen in n-alkanols. Whilst crankshaft type motions in the longer ligands of the latter panel result in C3-C4 and C5-C6 bonds undergoing reductions in $T\Delta S_{Or}$, adjacent bonds compensate this with relatively higher $T\Delta S_{Or}$ values. *Per contra*, 3Z-olefin bonds in the terminal segment possess a greater spread of states than n-alkanols as the intrinsic difference in ligand structure results in very different dynamics. A combination of high frequencies of orientational and translational displacements result in 3Z-olefins rapidly switching occupation from the larger *cal1* subcavity to the more crowded confines of *cal2 & cal3* and then back again. Despite high occupancy, locale-specific H-bonding interactions with the hydroxyl moiety that penalise per-bond $T\Delta S_{Or}$ in the hydroxyl segment, the greater variety of minima accessed by bonds in the terminal segment result in ameliorated *summed* $T\Delta S_{Or}$ losses in the majority of 3Z-olefins (**Fig.4.23.d**). In the case of 3c9, the larger size of the ligand results in the terminal segment being more tightly constrained, whilst the hydroxyl segment sees a concurrent increase in the number of available minima. These observations indicate that the proposed cumulative (group) effect of adding a methylene is likely to be an over-simplification as the concomitant change in ligand dynamics is not straightforward.

4.3.8. Trends in summed ligand $T\Delta S_{In}$ & $T\Delta S_{Or}$ values

4.3.8.1. The orientational entropy possesses internal and external components

Computational methods have to necessarily obtain thermodynamic values of interest by building up the total from smaller component parts. In order to estimate the total ligand $T\Delta S_{Or}$ orientational entropy, summed $TS_{Or\bullet B}$ values were subtracted from $TS_{Or\bullet F}$ (**Fig.4.26** & **Table.4.10**). The resulting trends are a 1st order estimate as correlations between orientational DOF have not been taken into account. Thus, the values are higher than expected. Second or higher order estimates are expected to reduce the magnitude of the final sum. The accuracy of the summed $T\Delta S_{Or}$ values will be fully discussed in §4.4.1 as it should be considered within the context of the total ligand entropy loss on binding. For now, trends in the data are examined.

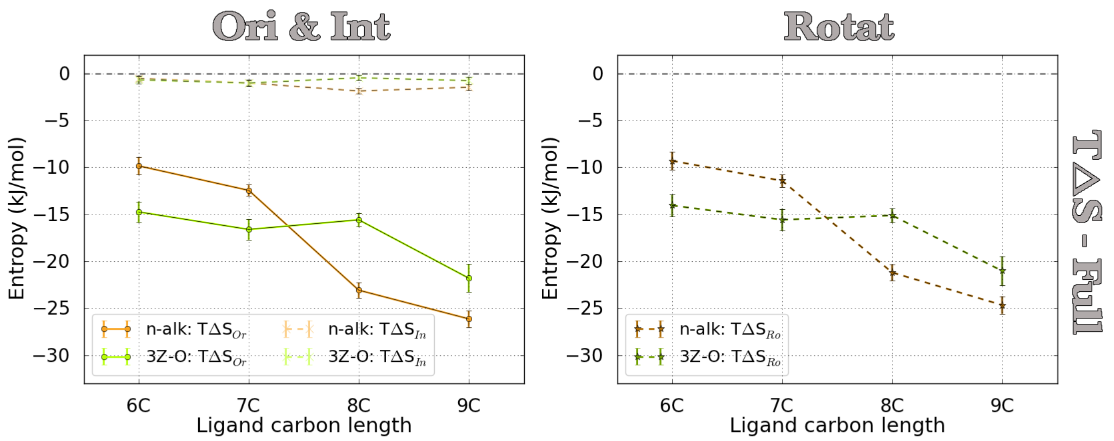


Fig.4.26. Total $T\Delta S_{Or}$ values are made up of contributions from $T\Delta S_{In}$ and $T\Delta S_{Ro}$ and in the case of flexible molecules, these cannot be calculated separately with the method used in this chapter. Hence, $T\Delta S_{In}$ is calculated independently and subtracted from $T\Delta S_{Or}$ to yield an estimate for $T\Delta S_{Ro}$. The results were generated from 1.2 μs simulations and demonstrate that the rotational contribution is much larger than the internal contribution, which is negligible. **(a)** Total $T\Delta S_{Or}$ (solid lines) and $T\Delta S_{In}$ (dashed lines) obtained by summing per-bond entropies for ligands in n-alkanol and 3Z-olefin panels. **(b)** Estimated $T\Delta S_{Ro}$ values for both ligand panels (**eqn.4.16**).

As previously discussed, the orientational component (as defined in this work) contains an implicit contribution from internal DOF and the method utilised in this chapter cannot deconvolute the two in the first instance (§4.1.3). However, it should be theoretically possible to obtain a retroactive estimate of the "pure" rotational entropy ($T\Delta S_{Ro}$) via **eqn.4.16**.

In order to accomplish this, $T\Delta S_{In}$ was recalculated with the method described in §3.2.4 and data from the aggregate 1.2 μs trajectories was utilised to ensure that exploration of phase space was identical to that of the orientational component (**Table.4.11**). Doubling the amount of data points does not significantly perturb the values obtained in chapter 3.0 and the internal contribution remains small (**Table.3.7**). As previously demonstrated, only relative differences between free and bound states have physical significance and differences in (bin) calibration between orientational and internal DOF would introduce errors into the subtraction of the end state values e.g. $TS_{Ro\bullet B =} TS_{Or\bullet B}$ - $TS_{In\bullet B}$. The partitioning of global values such as the entropy into internal and external DOF is arbitrary and only exists for ease of calculation [303,304]. It is always preferable to obtain the best estimate of the total entropy, and the orientational entropy (as defined in this work) could theoretically provide a much better approximation for the loss of a flexible ligand's DOF compared to the conformational entropy. This is because $T\Delta S_{Or}$ implicitly accounts for both internal and external factors when assessing the motions of bond rotors. Of course, this claim would be dependent upon accounting for second order or higher correlations. Fortunately, at the time of writing, this feature is currently in the final stages of development. Values for $T\Delta S_{Ro}$ are presented here only to illustrate the relationship between $T\Delta S_{Or}$ and $T\Delta S_{In}$ (**Table.4.10-11** & **Fig.4.26**). In this particular instance, the ligand possesses multiple bound minima and the internal contribution is negligible when compared to the rotational contribution. It is likely that a small contribution from internal DOF would also be observed in other promiscuous protein-ligand interactions due to the flexibility of ligands and protein binding sites (§4.1.1). Thus the established practice of merely accounting for the ligand conformational entropy could well introduce errors into the total binding entropy - a fundamental thermodynamic driving force.

		Orientational Entropy (kJ/mol)		
		$TS_{Or\text{-}B}$	**$TS_{Or\text{-}F}$**	**$T\Delta S_{Or}$**
Orientational: 1x 1.2 µs	**hex**	27.98 ± 0.92	37.84 ± 0.00	-9.85 ± 0.92
	hep	31.70 ± 0.60	44.15 ± 0.00	-12.45 ± 0.60
	oct	27.39 ± 0.82	50.46 ± 0.00	-23.07 ± 0.82
	non	30.63 ± 0.87	56.74 ± 0.00	-26.11 ± 0.87
		$TS_{Or\text{-}B}$	**$TS_{Or\text{-}F}$**	**$T\Delta S_{Or}$**
	3c6	23.10 ± 1.10	37.84 ± 0.00	-14.74 ± 1.10
	3c7	27.54 ± 1.10	44.15 ± 0.00	-16.61± 1.10
	3c8	34.87 ± 0.70	50.46 ± 0.00	-15.59 ± 0.70
	3c9	34.97 ± 1.50	56.74 ± 0.00	-21.80 ± 1.50

Table.4.10. Total orientational entropies obtained by summing per-bond entropies. Errors shown as standard errors.

		Entropy (kJ/mol)	
		$T\Delta S_{In}$	**$T\Delta S_{Ro}$**
Int & Rota: 1x 1.2 µs	**hex**	-0.54 ± 0.31	-9.30 ± 0.97
	hep	-1.02 ± 0.25	-11.43 ± 0.65
	oct	-1.88 ± 0.31	-21.19 ± 0.88
	non	-1.47 ± 0.33	-24.66 ± 0.93
		$T\Delta S_{In}$	**$T\Delta S_{Ro}$**
	3c6	-0.69 ± 0.39	-14.05 ± 1.17
	3c7	-1.03 ± 0.36	-15.58 ± 1.16
	3c8	-0.47 ± 0.27	-15.12 ± 0.75
	3c9	-0.77 ± 0.41	-21.03 ± 1.56

Table.4.11. Total internal and rotational contributions obtained from 1.2 µs simulations. Rotational entropy calculated via **eqn.4.16.** Errors are shown as standard errors and have been propagated for the rotational contribution.

4.3.8.2. Summed $T\Delta S_{Or}$ inter-panel trends

The shape of the n-alkanol trend line for summed $T\Delta S_{Or}$ is similar to that obtained for $T\Delta S^{\circ}_{Po}$ (**Fig.4.16** & **Fig.4.26**) and this reflects the relationship between ligand location within the cavity and the correlated effect on orientational DOF (§4.3.6). Surprisingly, the first 2 ligands in the n-alkanol panel suffer reduced $T\Delta S_{Or}$ losses compared to their unsaturated analogues (**Fig.4.26**). This is an unexpected result because the initial hypothesis stated that the offset (linear) entropic signatures obtained from the ITC data (**Fig.3.1**) were principally derived from the loss of ligand DOF. As $T\Delta S°_{Po}$ values for all the compounds tested were relatively similar (**Fig.4.16),** losses in orientational DOF were expected to yield the correct entropy differential that separated the trend lines of the two panels. As this is not the case, summed $T\Delta S_{Or}$ trends suggest that the ligand contribution cannot be the sole reason for the global ITC signatures.

In terms of the simulation data, the underlying rationale for this difference is partially derived from disparities within the H-bond network of the various complexes. Both shorter n-alkanol ligands preferentially H-bond to residues within the Ω loop with relatively *low occupancies* and as a result, most of their constituent bonds occupy the less crowded environment provided by *cal1* for large proportions of time (**Fig.4.20**). *Per contra*, their unsaturated analogues (3c6 and 3c7) possess *high occupancy* H-bonds to two residues situated near the middle of the calyx: TYR120 and ALA103 (**Fig.4.10**). Thus, a comparatively greater proportion of time is spent in *cal2 & cal3*. This disparity results in the two shorter n-alkanol compounds being able to access a greater variety of orientational minima than their unsaturated analogues (**Fig.A2.3-4**).

4.3.8.3. Summed $T\Delta S_{Or}$ intra-panel trends

The first 3 compounds in the 3Z-olefin panel have identical orientational entropy contributions within error and the resulting trend is almost flat. As the ligand's terminal segment can switch between tsubcavities with high frequencies, the bonds within this segment enjoy relatively elevated values of $T\Delta S_{Or}$ compared to bonds within the restrained hydroxyl segment (**Fig.4.23.d**). As the length of the ligand is increased (3c6 --> 3c8), the terminal segment becomes more restrained, whilst the hydroxyl segment undergoes a compensatory increase in mobility. There are more bonds in the terminal segment than in the hydroxyl segment, and the imbalance in bond number results in the first 3 ligands within this panel roughly "breaking-even" in terms of summed $T\Delta S_{Or}$ values (**Fig.4.23.d** and **Fig.4.26**). As bound 3c9 is more likely to be displaced and fixed within *cal2 & cal3*, most of the bonds within the terminal segment become restrained and the increased dynamics of the (fewer) bonds that make up the hydroxyl segment cannot quite ameliorate the overall loss in orientational entropy.

The trend line describing n-alkanol ligands comes closest to approximating the linear trend produced by global ITC data. Yet, the result indicates that the two shorter ligands are in a different subset to the longer ligands because there is a ~10 kJ/mol difference between hep and oct, whilst a mere ~2.5 to 3.0 kJ/mol difference separates ligands within each subset. This disparity could be rationalised by recalling that ligands within each subset occupy broadly similar translational states that have been shown to have a correlated effect on bond orientations (**Fig.4.13** & **Fig.4.21**), but it would be unwise to make this proposition without a thorough assessment of convergence: the Achilles heel of MD.

4.3.9. The implications of non-conforming non-convergence

4.3.9.1. The difficulties in assessing convergence

It is incredibly difficult to gauge whether a simulation has converged or not. Typical visual metrics may indicate convergence, but could actually be reporting a quasi-stable distribution and measured values could potentially change again upon the discovery of regions of phase space previously inaccessible due to high energy barriers (§4.3.6 & §4.3.9.2). Thus, simulation software cannot *a priori* deliver the correct Boltzmann weighted ratio of stable states that make up the ergodic distribution, because this is usually unknown. The best remedy against this is to run multiple long (> 1.0 μs) independent simulations (ideally N > 20) to maximise the probability of discovering new states and ensuring that their relative populations have stabilised [219]. The application of this measure to moderate sized (~150 residues) biological systems of interest is a computationally expensive proposition and few examples in the literature have achieved this level of sampling. To recognise when simulations match reality and avoid erroneous conclusions, the following checks should be implemented:

1. Monitoring the change in measurements with timescale.

2. Inter-methodological validation e.g. comparison to NMR spectroscopy, ITC, etc.

3. Intra-methodological validation e.g. comparison to other *in silico* methods.

4. Self-consistency across multiple independent perturbations e.g. constructing trends from ligand panels.

In addition to difficulties in finding the resources to extend simulated time, there are often issues with inter and intra-methodological validation as other studies often do not provide data that allows "like-for-like" comparisons. An expanded discussion of these issues is provided in the appendix (§A2.1.4). In conjunction with the first 3 types of check, it is feasible, and indeed advisable to assess the internal consistency of multiple MD simulations describing a series of perturbations to the same protein. The correlations between multiple analyses (such as H-bonding, positional and orientational entropies) obtained from a *single* protein-ligand binding interaction could be dismissed as self-fulfilling because correlations between the different results could be the consequence of the rules governing the simulation. However, the inter-woven correlations between trends obtained from a panel of multiple independent binding interactions are more reliable, because any disagreement in the trends rapidly highlights inadequacies in hypothesis, method and sampling. Thus, if non-convergence is suspected, outliers can easily be determined by non-conformance and data from other members in the panel used to identify the correct trend.

4.3.9.2. Bound orientational entropy ($TS_{Or•B}$) convergence

To assess the convergence of the orientational entropies, $TS_{Or•B}$ values are assessed in preference to $T\Delta S_{Or}$, because the main source of error is most likely to arise from difficulties in sampling the more complex energy landscape of the former contribution. As $TS_{Or•F}$ is well converged, $T\Delta S_{Or}$ assessment would contain a graded offset that adds an unnecessary factor into the trends (**Fig.4.27**). Also note that because this is a 1st order calculation, correlations have not been taken into account. Therefore differences between ligands have the propensity to become bigger as the number of bonds in the compound increases.

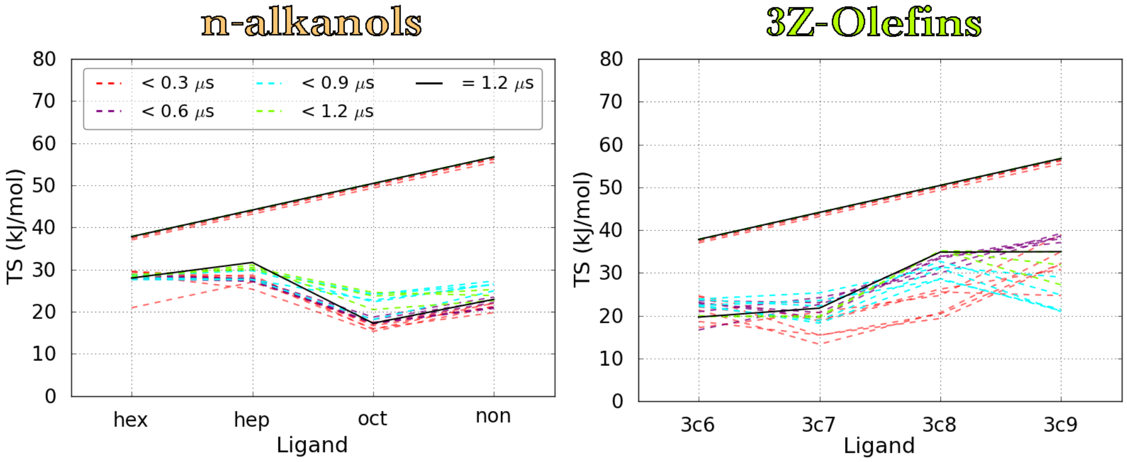


Fig.4.27. Convergence for summed TS_{OrB} and TS_{OrF} values were assessed over 1.2 µs of aggregate simulation time for both ligand panels. The number of points made available to the entropy calculation were iteratively increased by 6 x 10^4 (60 ns) each round. Solid black lines show the value calculated for the full 1.2 µs, whilst coloured dashed lines indicate how the entropy total for a given ligand changes with better sampling. Some of these lines cannot be discerned as they are overlaid and ligands that possess small separations in-between different dashed coloured lines indicate good convergence. Greater visibility of cyan and green dashed, coloured lines are discernable near points that have convergence issues.

It would appear that whilst hex and hep seems relatively well converged, the other n-alkanols do not possess fully equilibrated distributions. The problem is most prominent for oct as its $TS_{Or•B}$ value severely disrupts the panel trend. Oct also encountered more convergence problems compared to non when $TS°_{Po•B}$ values were considered (§4.3.3.4). As previously discussed, the rationale for this is likely to be because oct has the *potential* to move between different positional minima but tends to remain trapped within a given minima within the simulated time. Obviously, this has a correlated effect on the orientational distributions (**Fig.4.21**). On the other hand, 3Z-olefins possess greater facility to transition between bound positional minima and this concurrently increases the availability of orientational minima. Though 3Z-olefins can discover new minima more rapidly, there are issues with establishing the stable Boltzmann ratio of states within the simulated time. Thus, there are large swings in $TS_{Or\cdot B}$ as more data points are added to the calculation. Again, larger ligands in this panel have greater difficulties in transitioning between minima. In order to probe the effect of increased sampling, an additional 1.2 μs simulation of bound oct was run.

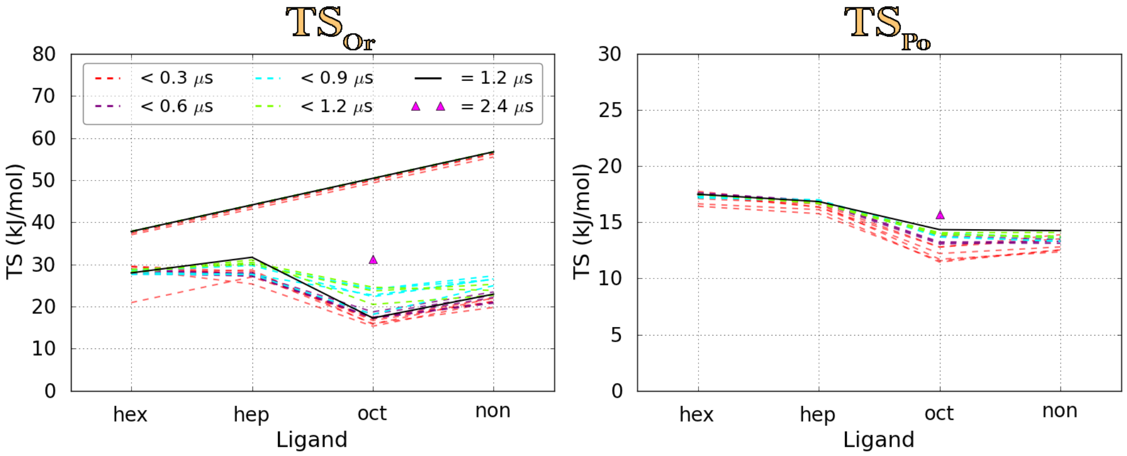


Fig.4.28. The convergence for total TS_{OrB} and $TS°_{Po\cdot B}$ was assessed over 1.2 μs of aggregate simulation time for the n-alkanol ligand panel. The magenta triangle marks the value returned by the entropy calculation on extending oct to 2.4 μs.

Values for $TS°_{Po\cdot B}$ and $TS_{Or\cdot B}$ were recalculated for the concatenated 2.4 μs oct trajectory, and entropic increases of 1.36 kJ/mol and 3.97 kJ/mol were obtained respectively. The new values modify the trend lines so that $TS°_{Po\cdot B}$ values across the panel of ligands become more linear and the increase in $TS_{Or\cdot B}$ brings oct more into line with the trend set by its neighbours (**Fig.4.28**). As there are more DOF involved in the calculation of summed orientational entropies compared to positional entropies, it is expected that there would be greater difficulties in convergence of the former contribution. Moreover, it is apparent that full convergence still has not been attained after 2.4 μs. The increase in the bound entropy narrows the difference between bound and free states, and if sampling time for all ligands were extended, it is expected that linear increases in line with ligand length would be observed. These effects are expected to be less pronounced for both smaller ligands compared to oct and non. This is because larger ligands have more difficulties in accessing different positional minima. Interestingly short (10 ns) simulations of n-alkanes binding to MUP performed by Wang et al. (2011) obtained identical free energy values for oct and non using the WaterMap program [281]. In response to the failure of this solvent-focussed approach, the group modified their program with an additional corrective factor to account for an additional methylene group possessed by non occupying *cal3*. Whilst this approach had the effect of linearising the intra-panel free energy trend-line, the modification is based on a static perception of binding. It does not deal with convergence by explicitly accounting for dynamic ligand switching between subcavities; a sampling dependent factor that affects the final results in terms of the enthalpies and entropies; and consequently the free energies.

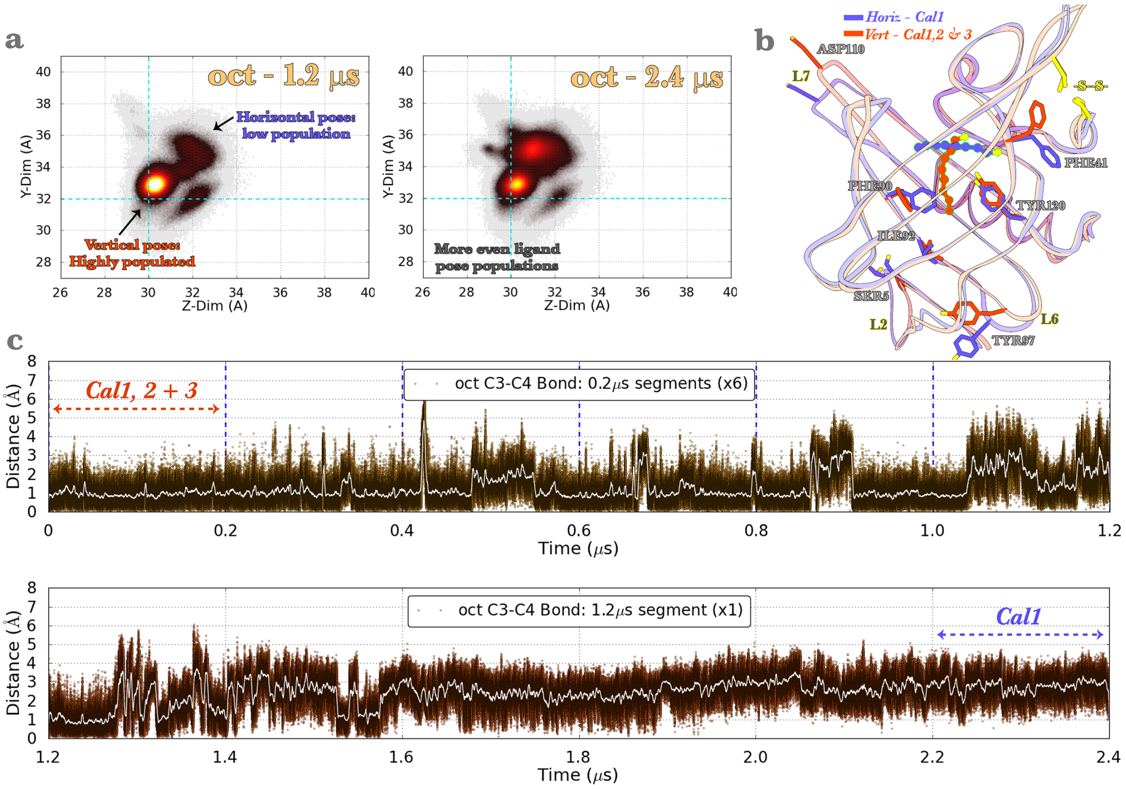


Fig.4.29. (a) 2D Projections comparing the COM distributions of bound oct obtained from 1.2 µs with 2.4 µs worth of sampling. It can be seen that there is increased occupation of *cal1* with increased sampling. The highest density minima correspond to localities within the calyx where oct can adopt vertical and horizontal poses. **(b)** Isolated 0.2 µs segments from the beginning and end of the 2.4 µs trajectory correspond to vertical and horizontal poses respectively. Averaged protein and ligand coordinates were generated from all segment frames to illustrate the difference between poses. Note the significant differences in ligand pose and protein residues: PHE41, TYR97 and PHE90. **(c)** The frequency of C3-C4 bond switching measured via distance displacements of that bond over the 2.4 µs concatenated trajectory. A rolling average (white line) was calculated using a 2 ns window. Dotted blue lines mark the boundaries of the independent, component 0.2 µs trajectories that were aggregated to yield the initial 1.2 µs dataset (top). The second 1.2 µs dataset (bottom) was obtained from a single, long trajectory. This demonstrates that the C3-C4 bond in oct tends to dwell within stable minima for long periods of time and the ligand has difficulty switching from vertical to horizontal poses. The baseline for ligand vertical poses occupying *cal1, cal2 & cal3* is ~1.0 Å, whilst horizontal poses adopted in *cal1* are marked by a baseline of ~2.0 to 3.0 Å. Segments at the beginning and end of the concatenated trajectory were isolated to perform cluster analysis on the two key ligand poses and are labelled in the figure (§4.3.9.3).

X-ray crystallography structures suggest that oct and non are more likely to adopt a vertical pose, but other factors relating to the crystallisation process (such as crystal packing) cannot be ruled out [36]. Nonetheless, the increase in measured entropies must be the result of substantial occupation of new minima, and analysis of the COM displacements indicates that bound oct increases its adoption of horizontal poses within *cal1* in the 2.4 μs trajectory, (**Fig.4.29.a**). On plotting the COM displacements of the C3-C4 bond versus time it can be ascertained that the additional 1.2 μs extension begins with oct occupying a vertical pose in *cal2* & *cal3* (**Fig.4.29.c**). Around 1.3 μs it switches orientation to adopt horizontal poses and the body of the ligand begins a long lived occupation of *cal1*. This is a stochastic phenomenon and additional long simulations are likely to see stable occupation of either pose in their respective positional minima. For this reason it would take a great many repeats to build up the correct stable ratio of minima that best approximates the ergodic distribution. Due to its larger size, bound non has a lower probability of switching between poses and this problem would be greatly exacerbated.

The freedom of the ligand to move within the pocket is likely to be affected by the particular region of the calyx it is situated in, and the next section assesses this relationship.

4.3.9.3. The principal regimes governing ligand dynamics (clustering II)

The discussion so far has highlighted the importance of the C3-C4 bond as the crux around which ligand dynamics revolves. The disparate pieces of evidence gleaned from multiple sources strongly suggest that ligand dynamics in MUP is governed by two distinct regimes. The first reigns within the larger enclave formed by *cal1* and the second in the smaller territories demarcated by *cal2 & cal3*. Shorter ligands are better able to fully occupy *cal1* by adopting a horizontal pose. However, longer molecules have a much greater *probability* of adopting a vertical pose on being displaced down into *cal2 & cal3*. In this locality, the intrusion of protein sidechains limits space and consequently, bonds are more restrained. To recapitulate the findings of §4.3.6, the flexible C3-C4 bond disfavours rigid body COM displacements and longer saturated ligands tend to occupy localised positional states for longer periods of time compared to their unsaturated analogues. Conversely, the rigid double bond introduces a curvature into 3Z-olefin ligands that facilitates the extraction of the hydrophobic tail from *cal2 & cal3* because the C3-C4 bond is able to rapidly switch its orientation. This results in greater positional and orientational displacements. As most 3Z-olefins can switch between both regimes with higher frequencies than n-alkanols, improved per-bond $T\Delta S_{Or}$ in the bonds constituting the terminal segment are the results of two key factors:

1. The confinement of the terminal segment of 3Z-olefins within *cal2 & cal3* is associated with unfavourable reductions in positional and orientational entropy. However, this is compensated by favourable entropic contributions arising from increased displacements into *cal1*, the more spacious upper chamber. 3c9 is the exception within this panel as it spends a greater proportion of time with its hydrophobic tail displaced into *cal3* due to its larger size. Moreover, the number of contiguous methylenes in unsaturated ligands is disrupted by the double bond and crankshaft motions are unlikely to occur, as at least 4 successive methylenes are required (§3.3.6.2). This requirement is only satisfied in the longest ligand, 3c9 (**Fig.4.23.d**). In contrast, analysis of the 1.2 μs trajectories indicated that longer n-alkanols tended to remain trapped within *cal2 & cal3*. Hence, crankshaft motions observed in per-bond $T\Delta S_{Or}$ trend lines are more pronounced (**Fig.4.22.d**).

2. In 3Z-olefins, the high frequency of switching between vertical and horizontal poses result in a greater spread of orientational minima than that observed for n-alkanols within 1.2 μs of simulated time. Thus, even if ligand size is increased, the greater number of accessible minima significantly ameliorates ligand entropy losses. The benefit can clearly be observed in the elevated $T\Delta S_{Or}$ entropies measured for bonds in the terminal segment of most 3Z-olefins, because this portion of the ligand is not restrained by polar interactions and can easily “flip” between subcavities (**Fig.4.23.d**).

The veracity of these statements can be established by analysing subsets of the concatenated 2.4 μs trajectories constituting the clusters that separate horizontal and vertical binding poses (**Fig.4.29.c**). Bound oct is a good candidate for such an analysis because the extended 2.4 μs trajectory contains two well-populated clusters that correspond to both poses (**Fig.4.29.a,b**). Fortuitously a complex cluster analysis is not required because oct remains within positional minima for long periods of simulated time and baseline shifts of the C3-C4 bond provide a excellent metric by which poses can be separated. Hence, a 0.2 μs segment was taken from the beginning and end of the 2.4 μs trajectory. These correspond to clustered ligand vertical and horizontal poses respectively. Visual inspection of these isolated trajectory subsections and averaged ligand binding orientations indicate that there is little (if any) cross-contamination between the two poses (**Fig.4.29.b**).

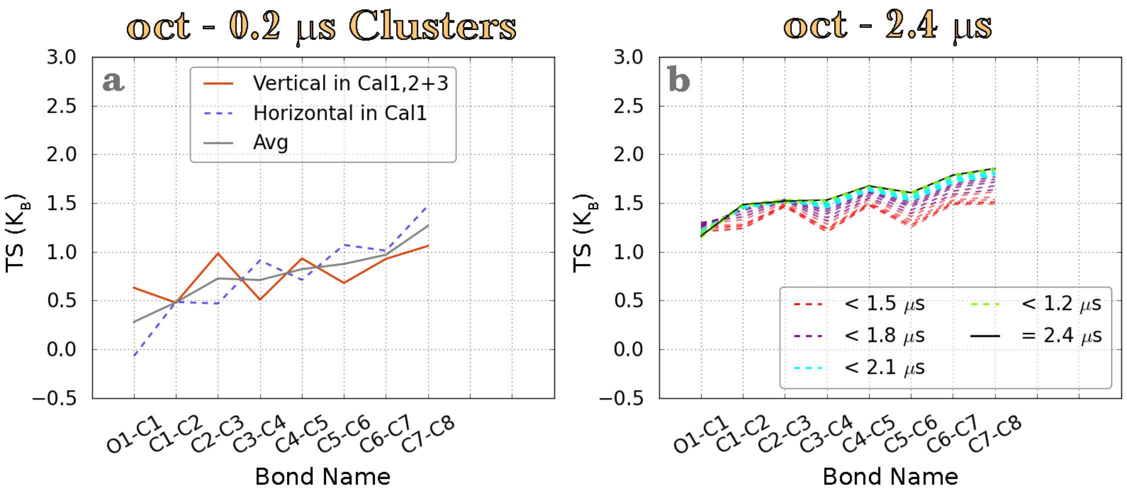


Fig.4.30. (a) The different per-bond TS_{OrB} trend lines obtained for vertical (solid red) and horizontal (dashed blue) poses for bound oct. 0.2 μs clusters representing each pose were isolated from the beginning and end of the extended 2.4 μs trajectory (**Fig.4.29.c**). The solid grey line plots the arithmetic mean of bonds in each cluster. **(b)** Per-bond TS_{OrB} convergence calculated from subsets of the full 2.4 μs trajectory. In the first 1.2 μs the ligand principally occupies the vertical pose, whilst the remaining time is overwhelmingly spent within the horizontal pose (**Fig.4.29.b**). Thus, on providing additional data points to the calculation, there is a transformation between the pronounced alternating pattern (due to crankshaft motions) arising from localisation to *cal1, cal2 & cal3* into a more linear trend line. This is the result of greater localisation in the spacious environment offered by *cal1* and the opposing per-bond TS_{OrB} trend generated by the horizontal pose.

The isolated clusters were used to recalculate the per-bond $TS_{Or•B}$ entropies for each pose (**Fig.4.30.a**). The results indicate that summed $TS_{Or•B}$ values (6.2 k_B) obtained from the horizontal pose are *on average*, equivalent to those obtained from the vertical pose (6.1 k_B). As fewer data points are used in the calculation, the overall magnitude of summed values obtained for each 0.2 μs cluster (**Fig.4.30.a**) is less than that obtained for the full 2.4 μs trajectory (**Fig.4.30.b**). More importantly, the minima described by the two very different poses are not simultaneously factored into the calculation (§4.3.6). To illustrate, the ligand initially adopts a vertical pose which generates a characteristic (alternating) pattern in the per-bond trend line. As more points are added to the calculation, increased occupation of the horizontal pose results in the pattern becoming less pronounced (**Fig.4.30.b**). This is because localisation of the ligand to the two different environments within the calyx generates compensating per-bond $TS_{Or•B}$ values which has a linearising effect on the trend line upon accounting for the entire dataset i.e. crankshaft motions in the horizontal pose are staggered in a manner that directly opposes $T\Delta S_{Or}$ troughs and peaks produced by the vertical pose.

The ligand's hydroxyl moiety is (generally) restrained by hydrogen bonds in both the poses isolated, but the O1-C1 bond is less restricted in vertical poses compared to horizontal poses (**Fig.4.30.a**). This is because bonds in the terminal segment are more tightly constrained within *cal2 & cal3* and when the ligand adopts vertical poses, the crowded environment promotes pronounced crankshaft motions that effectively redistribute per-bond $TS_{Or\bullet B}$ values throughout the body of the ligand. *Per contra*, when adopting horizontal poses, the terminal segment possesses greater freedom in *cal1*, and $TS_{Or\bullet B}$ baseline values show a graded increase as bond distance from the ligand's H-bonded hydroxyl group increases. However, the persistence of crankshaft motions in this pose indicates that ligand H-bonded aided localisation of one end of the compound plays an important role in the promotion of this phenomenon. This is probably because this strong polar interaction assists in maintaining the co-linearity of the terminal bonds (§3.3.6.2).

Despite increasing ligand size, 3Z-olefins roughly break even in terms of summed positional and orientational entropy penalties due to their facility to switch between subcavities with ease. This is further discussed in §4.4.1.

4.4.0. Conclusion

The results provided in this chapter indicate that the hypotheses formulated in chapter 3.0 requires modification as the internal view of dynamics did not sufficiently account for the size of the binding cavity and alternate ligand binding poses. A more comprehensive hypothesis can now be formulated subsequent to obtaining the positional and orientational entropies, because bound compounds that retain significant residual motion possess convoluted dynamics that are the result of interwoven contributions from both internal and external sources. This necessitates a holistic treatment that combines analyses from as many entropic contributions as possible.

4.4.1. Total external and internal ligand entropies

In order to establish whether the global ITC entropic signature ($T\Delta S^\circ_{Glo}$) chiefly arises from the ligand, orientational and positional entropy differences ($T\Delta S^\circ_{Lig} = T\Delta S^\circ_{Po} + T\Delta S_{Or}$) were summed to yield total ligand entropy contributions (Fig.4.31 & Table.4.12). It is clear that summed entropies are too high for two primary reasons:

1. This is a 1st order calculation and correlations between orientational DOF have not been taken into account. Hence, the true magnitude of $T\Delta S^\circ_{Lig}$ should be much lower than that calculated. This could be more pertinent for 3Z-olefins as the analysis performed in §4.3.7.2 indicated that bonds within the terminal segment undergo correlated shifts in orientation as the ligand switches between horizontal and vertical poses.

2. Larger ligands such as oct, non and 3c9 disrupt the trends of their respective panels. As discussed in §4.3.9, this is primarily due to the difficulty of obtaining simulation data that fully represents the correct Boltzmann ratio of states. In the case of oct, TΔS°$_{Lig}$ losses are ameliorated on increasing representation of undersampled horizontal poses, and the n-alkanol trend line becomes more linear (magenta triangle in Fig.4.31). Likewise, further sampling for the other ligands is required to correct the slope and offset of both panel trend lines. In an ITC aliquot containing 3.4 x 10^{-5} Moles of protein, there are approximately hundred quintillion protein molecules that exist in bound, unbound or intermediate states. Due to technological limitations, it is currently impossible to obtain an equivalent amount of protein-ligand conformational sampling via MD. However, methods such as metadynamics and accelerated molecular dynamics offer an avenue through which the correct Boltzmann ratio of stable states can be sampled [134,135,377–382].

As discussed in §4.1.2.5, it is very difficult to experimentally decompose entropies into component contributions due to broad nature of underlying assumptions. The ultimate determination of *in silico* accuracy would be achieved by the calculation of the protein, ligand and solvent entropy, so as to yield the total system decomposition. However, this is a challenging and ambitious goal which will require 2nd to 3rd order calculations to achieve pinpoint accuracy. This is beyond the reach of this work, but the 1st order entropic trends generated from multiple independent simulations afford the ability to assess the veracity of calculated values and make conclusions regarding the disparate contributions that compose the global entropy.

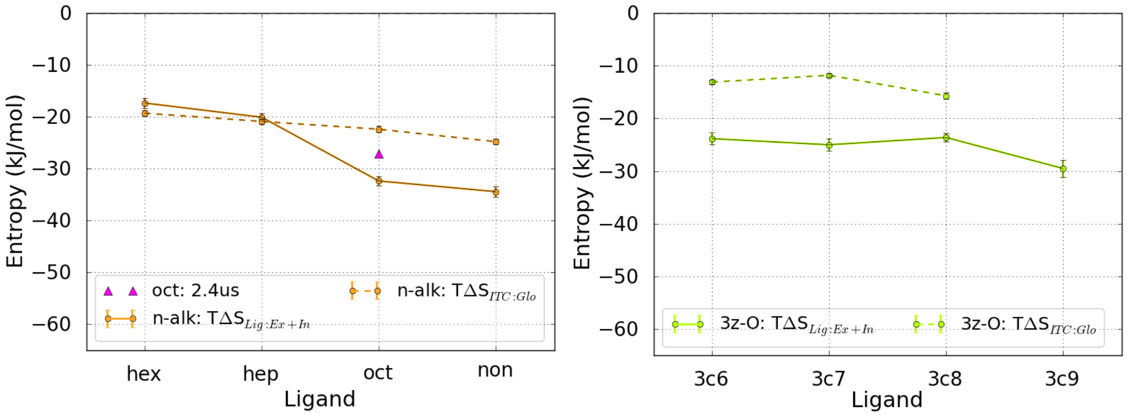


Fig.4.31. Total ligand contributions (TΔS°$_{Lig}$) for ligands in both panels were calculated by summing total positional (TΔS°$_{Po}$) and orientational entropy (TΔS$_{Or}$) contributions (solid lines). These values also include internal contributions. The magenta triangle marks the value returned by the entropy calculation on extending oct to 2.4 ms. Global ITC values are depicted as dashed lines and the scale used for the y axis is the same as **Fig.1.a** in Malham et al. (2005) [36]. No ITC data is available for 3c9.

In the case of the n-alkanol panel, hex and hep are very close to the experimental values, and it is likely that the data points for both of the longer ligands will also come very close

Total Ligand Entropies vs. Global ITC Data (kJ/mol)						
		$T\Delta S_{Po}$	$T\Delta S_{Or}$	$T\Delta S_{Lig}$	$T\Delta S_{ITC}$	$T\Delta\Delta S$ (Expt - Calc)
n-alkanols	**hex**	-7.50 ± 0.28	-9.85 ± 0.92	-17.35 ± 0.96	-19.3 ± 0.6	-1.95
	hep	-7.64 ± 0.42	-12.45 ± 0.60	-20.09 ± 0.73	-20.9 ± 0.4	-0.81
	oct (1.2ms)	-9.31 ± 0.48	-23.07 ± 0.82	-32.38 ± 0.95	-22.4 ± 0.6	9.98
	oct (2.4ms)	-7.95 ± 0.40	-19.10 ± 0.75	-27.06 ± 0.85	-22.4 ± 0.6	4.66
	non	-8.32 ± 0.51	-26.11 ± 0.87	-34.43 ± 1.01	-24.8 ± 0.5	9.63
		$T\Delta S_{Po}$	$T\Delta S_{Or}$	$T\Delta S_{Lig}$	$T\Delta S_{ITC}$	$T\Delta\Delta S$ (Expt - Calc)
3Z-olefins	**3c6**	-9.10 ± 0.31	-14.74 ± 1.10	-23.84 ± 1.14	-13.1 ± 0.4	10.74
	3c7	-8.40 ± 0.41	-16.61± 1.10	-25.01 ± 1.17	-11.8 ± 0.3	13.21
	3c8	-8.00 ± 0.35	-15.59 ± 0.70	-23.59 ± 0.78	-15.7 ± 0.6	7.89
	3c9	-7.71 ± 0.60	-21.80 ± 1.50	-29.51 ± 1.62	na	na

Table.4.12. Comparing globally measured ITC values to computationally generated entropy values.

to the ITC values on increasing sampling. However, it is doubtful whether restriction of ligand DOF is the sole reason for measured $T\Delta S^\circ_{Glo}$ values because computed values for ligands in the 3Z-olefin panel are too far away from the experimental values. These unsaturated ligands are already very dynamic and additional sampling is unlikely to fully close the gap between $T\Delta S^\circ_{Glo}$ and $T\Delta S^\circ_{Lig}$. On considering the magnitude of the shift in the $T\Delta S^\circ_{Lig}$ data point for oct on increasing sampling, it is likely that $T\Delta S^\circ_{Lig}$ will range from ~ -15 to -23 kJ/mol for all the ligands tested here on increasing the number of simulation repeats and accounting for higher order corrections. We are currently in the process of confirming this and if this is indeed the case, computed values for 3Zolefins would be higher than the experimental values, whilst n-alkanols should be lower. Regardless, a pinpoint accurate calculation of $T\Delta S^\circ_{Lig}$ is not required to establish that the observed $T\Delta S^\circ_{Glo}$ trends for both ligand panels are not be principally derived from the loss of ligand DOF. The argument for this will be outlined in §4.4.2 and proven in chapters 5.0 to 6.0.

In order to address the key hypotheses posited in §4.1.4.1, $T\Delta S^\circ_{Lig}$ trends will be described on a per-panel basis and compared to global ITC values. These hypotheses are restated below:

> ***Hypothesis-1:*** *Extending ligand length by a single methylene group yields an enthalpic gain due to increased protein-ligand van der Waals contacts. However, a compensating entropic penalty of 5.4 kJ/mol is paid due to the addition of a rotor which inevitably becomes restrained on binding*[36].

Hypothesis-2: *Disabling a rotor via introduction of a double bond avoids the entropic penalty on binding as this debt has been paid during the process of chemical synthesis* [36,178].

Hypothesis-3: *Pre-organisation of the ligand (i.e. 3Z-olefins) so that the structure complements the shape of the binding site ameliorates entropic penalties by virtue of less strain being imposed on the ligand* [174,177–179].

Strictly speaking, *hypotheses-1-2* specifically refers to the restriction of internal DOF. However, this contribution was found to be negligible and it is of interest to test these hypotheses against the loss of all ligand DOF to see if they have any merit. In the case of the n-alkanol panel, the loss of ligand internal and external DOF are correlated with the addition of a rotor and this partially validates *hypothesis-1* in terms of increasing ligand size and greater amounts of restriction. However, $T\Delta\Delta S_{Lig}$ values obtained by calculating the difference between the two shortest members of the panel (~2.5 kJ/mol) do not fulfil the theorised prediction of a 5.4 kJ/mol penalty paid on extending the ligand by a rotor. This is because bound compounds retain considerable residual motion and the proposed value of 5.4 kJ/mol is an overestimate based upon the cyclisation of saturated hydrocarbons (§3.3.4) [171]. Unfortunately the values returned for oct and non are not converged, so additional evidence from this panel remains to be obtained. On the other hand, $T\Delta S°_{Lig}$ values for the first three members of the 3Z-olefin panel are identical within error and $T\Delta\Delta S_{Lig}$ is < 1.5 kJ/mol. So, even after allowing for the loss of all ligand DOF, a systematic group penalty for each additional methylene group is not realised.

The 3Z-olefin panel constitutes a particularly interesting perturbation because the dynamic behaviour that arises from a seemingly small ligand modification was impossible to predict *a priori*. The change in molecular shape and rules governing its motion resulted in complex interactions with the other entities within the system. Rather counter-intuitively, $T\Delta S°_{Po}$ values marginally improved with increasing ligand length whilst summed $T\Delta S_{Or}$ values for the first 3 ligands in the series were identical within error (**Fig.4.16, Fig.4.26** & **Table.4.12**). Due to the dynamic complexity of these unsaturated compounds, summed inter-panel $T\Delta S°_{Po}$ and $T\Delta S_{Or}$ contributions are identical within error and the final $T\Delta S°_{Lig}$ trend line remains relatively constant (~ -24.1 kJ/mol). As 3c9 is unlikely to be fully sampled for similar reasons to oct and non, no special physical significance is attached to the lower value obtained for this ligand. $T\Delta S°_{Lig}$ values that compose the trend line are (too) far away from $T\Delta S°_{Glo}$ and it is likely that other (non-ligand) contributory factors will further modulate the magnitude and slope of the final trend. The results indicate that *hypothesis-2* is too simplistic a rationale for the observed behaviour as introduction of a double bond does not result in a graded trend line with a systematic differential betwixt panel members.

Indeed, $T\Delta S^{\circ}_{Lig}$ values remain relatively constant. As discussed in §4.3.5-7, 3Z-olefins avoid $T\Delta S^{\circ}_{Lig}$ losses on increasing ligand length due to their shape, structure, and ability to adopt differential H-bonding patterns. The combination of these three factors allows them to dynamically switch between orientations and positions with much higher frequencies than n-alkanols and thus access a greater variety of states. For this reason, *hypothesis-3* is partially correct as the mechanism described does result in less "strain" in 3Z-olefins due to shape complementarity. However, as *hypothesis-3* is rooted within a static concept of binding, it does not account for the dynamic switching of these unsaturated ligands between subcavities, and is thus too vague to be universally applicable to all protein-ligand binding systems.

All three of these hypotheses are the simplest explanations for the difference in the global ITC trend of n-alkanols versus 3Z-olefins. Testing them has ensured that the principles embodied by Occam's razor have been followed. This law advises that problems should be solved via the simplest explanations possible, whilst ensuring that the number of explanations is not multiplied needlessly (parsimony). As none of the tested hypotheses are universally applicable to the trends exhibited by all the ligands, it is useful to be reminded of Hickam's dictum which states that "patients can have as many diseases as they damn well please". This is the medical counterbalance to the parsimonious approach advocated by Occam's razor and reminds the doctor that a particular diagnosis should never be excluded on the grounds that a greater amount of diagnoses violates Occam's razor. This is because the parsimony principle refers to new assumptions, not additional diagnoses which are entirely possible, and indeed likely with increasing patient age [383,384]. In a similar vein, the description of the relationships between position, orientation and H-bonds offer a more comprehensive, albeit complex hypothesis regarding the main factors that modulate ligand entropy. The proposals have been refined by assessing the trends exhibited by these distinct factors and account for the relationships between them. Because this fully explains the data obtained from 8 different complexes, they are considered to be more reliable than the simpler, initial hypotheses which do not account for all the observations.

Quantification of $T\Delta S^{\circ}_{Lig}$ does not unequivocally indicate whether the restriction of ligand DOF is the sole reason for global entropy losses on binding. However, it is very likely that linear trends will captured from *multiple independent simulations* on accounting for the required level of sampling. Having analysed the subset of the DOF that constitute the ligand, the apportioned thermodynamic values should be further tested within the context of the superset of components that make up the entire system. So, the next section scrutinises the falsifiability of the ligand-centric theories posited in this chapter by gauging how calculated $T\Delta S^{\circ}_{Lig}$ values fit within a total decomposition of the entire system.

4.4.2. Predictions for the total system thermodynamic decomposition

As thermodynamic data describing the binding of various ligands to MUP were uncovered, several thermodynamic decompositions based on the experimental data to-hand were proposed along with a single *in silico* based decomposition [34,89,91,100,163]. The focus of these analyses was primarily on the binding of IPMP and IBMP to MUP. Consequently, whilst there are differences in the thermodynamic characteristics of those hydrophobic ligands compared to the n-alkanols studied in this work, many of the core principles are still applicable. In terms of the binding of the alcohols, a key assumption underlying the thermodynamic decomposition that led the initial hypotheses is that the protein contributes minimally to the global entropic signature [36]. Thus, restriction of ligand DOF and ligand desolvation were identified as the principal contributors. However, the assumption is likely to be invalid as the protein contribution differs between IPMP and IBMP binding to MUP (**Table.1.3**). Moreover, relaxation experiments performed by Zidek et al. (1999) report *increased* backbone dynamics on binding SBT [100]. It is somewhat counterintuitive to expect protein DOF to become more dynamic on ligand binding but studies on other protein-ligand systems have also seen similar increases [385–387]. These factors point to a paradigm within which there is a differential protein response that is dependent on the identity of the bound ligand.

It seems very likely that, though the ligand contribution to the total entropy loss on binding is relatively large, it cannot account for the offset linear trends captured by ITC. The total computed system entropy ($T\Delta S°_{Sys} = T\Delta S°_{Lig} + T\Delta S_{Prot+} T\Delta S_{Solv}$) is made of various contributions that have the potential to accumulate or (partially or totally) cancel when summed. As there are more DOF available in the protein and solvent components, their contributions are likely to have a substantial impact on $T\Delta S°_{Sys}$ and thus, the magnitude and gradient of the final $T\Delta S°_{Sys}$ trend lines are likely to be strongly affected.

The increased localisation of n-alkanols on increasing ligand length suggests greater interfacial interactions that scale with size i.e. this promotes increasingly favourable enthalpic interactions. Thus, it is also likely that the binding of n-alkanols is accompanied by graded reductions in $T\Delta S_{Prot}$ because stronger protein-ligand interactions would decrease the dynamics of protein residues. *Per contra*, 3Z-olefins are likely to possess reduced interfacial contacts due to substantial residual motion and an unfavourable $T\Delta S_{Prot}$ contribution with a shallower slope should be expected. Obviously, this would result in an enthalpy that was more unfavourable than that of n-alkanols. If this were the case, summed $T\Delta S°_{Lig}$ and $T\Delta S_{Prot}$ values (for both panels) would *inevitably* be far more *negative* and unfavourable than reported by $T\Delta S°_{Glo}$. As MUP is suboptimally hydrated and bound waters are relatively disordered, a large entropic contribution should not be expected from protein desolvation. However, as explicated in the literature, the ligand desolvation penalty on binding primary alcohols to MUP affords a *positive* entropic contribution ($T\Delta S_{Solv}$) that linearly increases with ligand surface area [36,180,388]. In the case of the n-alkanols tested here, ligand desolvation entropies range from +57.7 to +68.8 kJ/mol at 300 K [36]. Thus, the favourable desolvation contribution would compensate the overwhelmingly unfavourable protein and ligand contributions to yield the values measured by global ITC. To avoid criticism that ligand desolvation and protein contributions are invoked as *deus ex machina* devices that function to smooth the narrative, the next two mini-chapters will briefly examine each in turn.

4.4.3. Methodological developments & thermodynamic "rules of thumb"

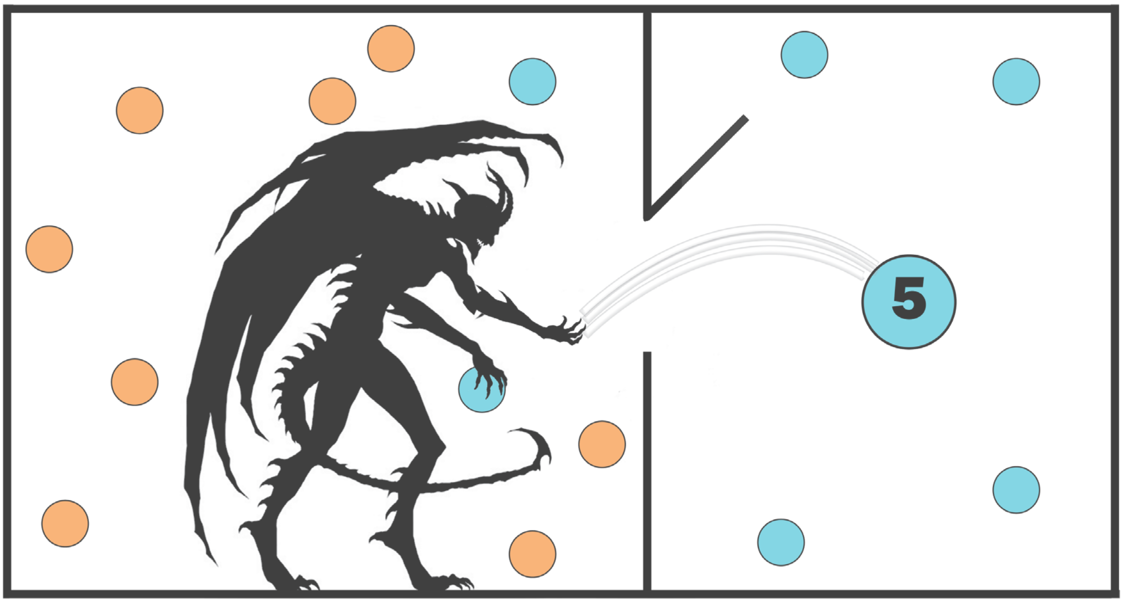
Ultimately, it is of great importance to benchmark methods to ensure that they attain the highest level of accuracy possible and the methods presented here will be further developed so as to account for second and higher order correlations. However, an argument can be made that pinpoint accuracy in drug design is unnecessary when using proven methods because this is expensive and time consuming. It would be more efficient to broadly characterise the most important components within the system via analysis of the trends exhibited across panels of compounds possessing incremental structural differences. This approach easily identifies outliers, and the insights offered could rapidly guide the rational modification of ligands during the early stages of drug development.

In this chapter, the total ligand entropy has been "built up" from the smallest tractable units applicable to its positional and orientational subcomponents. The unique features afforded by the methods developed in this chapter are:

1. Ligand positional entropy: This 3D histogramming method does not assume that the underlying COM distribution possesses any functional form e.g. the harmonic approximation. Thus, the detailed shapes of distributions are better described and $T\Delta S°_{Po}$ of ligands possessing multiple bound minima are more accurately quantified than methods currently in use. Additionally, the method's simplicity allows rapid implementation and execution.

2. Ligand orientational entropy: This method also does not assume any underlying functional form to the distribution and the chief benefit of the approach is the calculation of the entropic penalty paid on hindrance of the principal rotations of ultra-flexible ligands. This is accomplished by calculating $T\Delta S_{Or}$ on a per-bond basis and analysis of the resulting trends affords greater insights into the "how and the why" regarding both ligand dynamics and the resulting entropic binding signatures. A benefit of this approach is that both the conformational entropy ($T\Delta S_{In}$) and the rotational entropy ($T\Delta S_{Ro}$) are captured in a single calculation (§4.3.8.1). To obtain accurate summed results, 2nd order or higher correlations are required. However, the same problem exists when calculating the conformational entropies and this feature is in currently the final stages of development.

The reductionist partitioning of the entropy into various internal and external DOF is artificial and arbitrary, its chief purpose being to facilitate the description of motion; a complex phenomenon that arises from the interaction of its constituent parts and the relationships between them. It could be said that motion is an emergent property of the definitions of its constituents and understanding this does not afford *a priori* predictions of dynamic behaviour when placed within the context of disparate environments i.e. the motion of a body in different environments cannot be qualitatively predicted by a theoretical understanding of it's parts as it is always more than the portrait painted by its elemental definitions. This is because the relationships between the various DOF have to be taken into account, and inherent uncertainty in the predicted result restricts the broad application of what has been deduced in one system to others. This is the principal reason why proposed group effects for entropic penalties (e.g. the introduction of a double bond) yields inconsistent results [177–179]. The entropy has to be captured in totality or as close to that as possible. If this is not done, the gap in information regarding the system is likely to obfuscate data and vitiate conclusions because entropic components have the potential to accumulate and cancel. Even within an ostensibly simple system like MUP, complex interwoven dynamics result from seemingly small unit modifications to the structure of the ligand. This situation is likely to be applicable to many other protein-ligand systems due to the prevalence of promiscuous binding and the results demonstrate that contributions from positional and orientational components are likely to dwarf that of the commonly calculated conformational entropy. Hence, it is necessary to expand the repertoire of methods that accurately quantify the entropy and shore the barriers against a rising tide of failed drug-like compounds. This chapter has contributed to this goal by reductively characterising three primary factors (H-bonds, position & orientation) and holistically integrating them to provide a better understanding of the nature of ligand dynamics within MUP.



Maxwell’s Daemon

Chapter 5.0: Ligand Internalisation & Desolvation

5.1.0. Introduction

This chapter complements the structural/dynamic analysis presented in chapter 4.0 and further maps the architecture of MUP. Researchers have hypothesised the postulated mechanism of ligand entry into the calyx of various lipocalins [72,389–392] and others have probed different aspects of cavity and ligand desolvation [90,91,180,393–399]. However, with regard to MUP, there are still many unanswered questions and this chapter attempts to address these by developing a protocol to study the ligand internalisation process. This provides insights into some of the underlying mechanisms associated with ligand internalisation and also reports on the behaviour of water molecules during that process.

5.1.1. Protein-Ligand encounter complexes

According to collision theory, the formation of a bimolecular complex is affected by three principal terms relating to collision frequency, optimal structural alignment of the interacting entities, and the activation energy. Assuming perfect alignment and an activation energy of zero, the rate of association of two equally-sized molecules is diffusion controlled and is estimated to be ~10^9 s^{-1} M^{-1}. In the case of protein-ligand binding, the rate is higher (10^9 to 10^{12} s^{-1} M^{-1}) due to a combination of high ligand mobility and the large surface area presented by the protein [400]. First contact is unlikely to yield optimal protein-ligand contacts and this loose association is known as an encounter complex. The overall process can be modelled as a generalised two-step model where P and L are the free species, PL* the encounter complex, and PL the final bound complex (**eqn.5.1**) [400–403].

$$P + L \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} PL^* \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} PL \qquad \textbf{(eqn.5.1)}$$

The initial formation of the encounter complex is governed by Brownian motion and long-range electrostatic interactions between the separated entities in solution. In contrast, the lifetime of the encounter complex is characterised by short-range protein-ligand interactions as the ligand explores the protein surface. As the initial collision is unlikely to be near the final binding site, this mechanism acts to reduce the search space from three dimensions to two and a classic theoretical study on the binding of transcription factors to DNA quantifies this process [401,404]. Similar to models describing protein folding, this conformational search mechanism is said to shape the binding free energy landscape into the form of a funnel that directs the ligand to the final complexed state [364,401–403,405–408]. Various intermediate transition states within the funnel possess differential binding energies due to variations in intermolecular interactions. This is postulated to form a free energy gradient that allows the ligand to adopt the final bound pose [407,408]. Structural features on the surface of the protein such as polar/charged residues, proximal and distal to the binding site assist the capture of the ligand and its subsequent translocation. Computational mutation of these key residues has been shown to affect the association rate and the pathways ligands use to reach their targets [403,409]. But why is the study of protein-ligand encounter complexes important?

Swinney (2004) criticised current prevalent attitudes in drug development because the focus is on delivering thermodynamically tight binding compounds. But kinetic factors associated with the open systems typically found within *in vivo* environments are typically neglected [410]. The majority of inhibitors are geared towards competing with the target's native substrate, and thus act to increase its binding affinity (K_m) but fail to affect its maximum activity (V_{max}). This can result in unprocessed substrate concentration increasing. Thus, higher compensatory doses are required to maintain effective inhibition. A study of the most successful drugs indicated that they could be characterised by a two-step model in which an initial equilibrium stage was followed by a transition to a non-equilibrium state e.g. irreversible, slow disassociation, allosteric interactions via corepressors, etc. Additionally, the mechanisms by which these compounds acted were well known in the early stages of development [410]. Thus, designing drugs that targeted non-equilibrium states allowed V_{max} to be more effectively targeted and this was preferential to optimising K_m. This is a compelling argument to study the events that precede the final bound state because understanding these association pathways facilitate the design of better small molecule inhibitors. The next section provides a brief overview of *in silico* techniques that assess ligand internalisation.

5.1.2. Ligand internalisation via free diffusion & enhanced sampling

As protein-ligand affinity is described by more than what can be gleaned from equilibrium thermodynamics, greater gains can be reaped by studying non-equilibrium association processes which reveal interactions crucial to molecular recognition. A popular method that assists such investigations is high throughput docking. It allows the mapping of putative binding sites and associated ligand binding modes. Speed is prioritised by reducing the DOF in the system but this impedes accuracy. Nonetheless, scoring functions provide estimates of the binding affinity of multiple candidate compounds to a desired target and their rank order can be rapidly determined (§1.4.2). But nowadays, there is the requirement for methods that offer greater accuracy, even if this is at the expense of speed because computational technology has evolved, and these two factors can now be rebalanced.

In silico calculations can be broadly categorised into endpoint and pathway methods [411]. The first encompasses techniques such as MM-(P/G)BSA, linear interaction energy (LIE) and various methods that capture the entropy differential between free and bound states (e.g. Chapter 3.0 to 4.0). Pathway methods were greatly influenced by the work of Kirkwood, De Donder and Zwanzig. They include techniques such as FEP, TI, double-decoupling and double-annihilation [412–415]. As detailed in §2.1.3-4, these methods facilitate optimal navigation through the tangled skein of n-dimensional phase space via the construction of a discretised path linking target and reference states. In order to avoid the computational expense accompanying thorough Boltzmann sampling, discrete waypoints along the route are sampled with a biased or unbiased Boltzmann factor. Thus, less expensive estimates of the total free energy difference are obtained by summing ΔG values calculated from the overlapping zones centred around each waypoint [379,411]. Whilst these approaches usually deal with *in-situ* transformations, there also exist several techniques that spatially separate end points e.g. umbrella sampling (US), steered molecular dynamics (SMD) and targeted molecular dynamics (TMD) [411,416–418]. In the context of protein-ligand binding, this allows characterisation of the free energy profile associated with the route the ligand takes from the unbound state (within bulk solvent), to the final bound state. Unfortunately, this approach requires extensive sampling and prior knowledge of the path (or reaction coordinate). Nevertheless, the recent advent of accelerated GPU compute capabilities has enabled the development of new methods that overcome these limitations. These technological advancements also lower the computational cost of using explicit solvent, and this contributes to more rigorous results than that offered by docking [411,419].

Pathway methods can fully characterise the ligand binding mechanism both thermodynamically and kinetically. For instance, free diffusion (FD) simulations allow the ligand to freely diffuse from bulk solvent to its protein target and this totally unbiased technique affords mechanistic information on the binding process and facilitates the discovery of alternative binding sites. However, it is largely stochastic and limited by difficulties in obtaining binding events. In terms of discovering new ligand modes and binding sites, efficiency can be increased by flooding the simulation with excess ligand and several studies have reported better results with this approach than that obtained by docking [420–422]. Unfortunately, recovering rigorous kinetic data from FD simulations typically requires much more sampling than necessary for the thermodynamic evaluation of equilibrium simulations. This is because the complete free energy landscape associated with the ligand's movement from bulk solvent to its final bound position has to be characterised. Nonetheless, this approach allows better understanding of binding pathways via the identification of various metastable states on the surface of the protein that act to hinder or abet binding [403,405]. Early efforts captured valuable mechanistic details with relatively little computational expense (100's of nanoseconds). However, limited sampling meant that these attempts fell short of obtaining quantitative data on the energetics and kinetics of such processes [423–425]. Shan et al. (2011) and Buch et al. (2011) reset the bar in charting the free energy topography associated with protein-ligand binding by performing simulations that reached microsecond lengths when aggregated [419,426]. The latter study utilised 495 (100 ns) simulations of the binding of benzamidine to trypsin (187 binding events) to construct Markov state models (MSMs) that yielded a free energy estimate < 1.0 kCal/mol of experimental measurements. Furthermore, they also obtained the associated k_{on} and k_{off} rates. The binding pathway could be energetically categorised into a series of metastable states representing the ligand's position in bulk, first-recognition contacts, surface sites that offered greater protein-ligand stabilisation, and the final bound state. Benzamidine did not move directly into the binding site from bulk because binding was dependent on key interactions with these intermediate metastable states. The transitions between states defined rate limiting steps and understanding the kinetics associated with internalisation allows better inhibitor design. In order to minimise the expense associated with FD experiments, the protein was harmonically restrained in the centre of a volume demarcated by an additional flat-bottomed harmonic restraint; both of which respectively functioned to prevent protein and ligand diffusing away. Later developments in this area saw the creation of techniques such as funnel metadynamics which enhances sampling by increasing the sophistication of restraints. This acts to further decrease the large free volume and limit sampling to functionally relevant areas [411,427].

5.1.3. Objectives

Having just described the state of the art, it is recognised that the same level of sampling cannot be achieved with the resources available to this project. As hardware resources were limited, a protocol was developed whereby ligand internalisation could be studied in a computationally cost-effective manner. Despite this, the quality of the presented results are in line with the achievements of published work within the literature [423–425]. Simulating ligand internalisation allowed the following questions to be addressed:

1. How does the ligand find the entrance of the occluded cavity?

2. What is the mechanism by which the protein admits ligands into the occluded cavity? Can any detail be added to existing atomistic studies done on other lipocalins? [391,392]

3. What is the process by which the ligand rearranges itself to adopt poses similar to that observed in crystal structures?

4. How many waters are expelled from the pocket on ligand binding?

5. How many waters enter with the ligand on binding?

6. To what extent is the ligand desolvated upon ligand internalisation?

5.2.0. Methods

In order to allow simulation data to be generated, I built a custom-built quad-GPU machine to generate all the simulation data. Special thanks go to my industrial sponsor Janssen for supplementing my personal stipend with precisely the correct amount to achieve this.

5.2.1. Ligand internalisation protocol

Unlike the strategies used by Limongelli et al. (2013) and Buch et al. (2011), no restraints were placed on the protein to hold it in place during any production stage. This was done to minimise perturbation of the protein's internal dynamics [411,419].

5.2.1.1. Stage-1 (Free Diffusion):

Some of the equilibrium simulations utilised in the previous chapters had to be rerun as the ligand became unbound for stochastic reasons (**Table.4.2**). Multiple unbinding events allowed the location of the ligand exit to be identified as that proposed by Timm et al. (2001) [72]. To begin understanding the ligand entry mechanism, four FD simulations were set up using the AMBER ff03 force field. The likelihood of success was maximised by randomly modelling the 12 different ligands from the three panels tested in chapter 3.0 approximately 10 Å away from the protein. The simulations were unbiased in every way apart from the relatively short distance between protein and ligands. The solute was solvated with TIP3P waters using a box size of 10.0 Å. Each simulation was run for 100 ns using NPT conditions as described in §3.2.3. At the end of this process a single ligand (3c9) from one simulation entered the pocket, whilst the others drifted away or were involved in non-specific interactions with the protein. Examination of this sole internalisation event revealed that the ligand lingered near the L3 and Ω loops within a small depression near the entrance that acted to corral the ligand in place. On the basis of these observations, twelve additional simulations were run using the protocol detailed below.

5.2.1.2. Stage-2 (Targeted to Free Exploration MD):

Concepts from TMD and FD simulations were amalgamated to create a protocol referred to as Targeted to Free Exploration (T2FE) MD. A structure was isolated from the stage-1 simulation in which the 3c9 ligand was located within the cupped depression formed by the close association of the *d* strand, L5, L3 and Ω loops (henceforth referred to as the mouth). Bulk waters, ions and other ligands were deleted. The ligand was repositioned ~18 Å away from the alpha carbon of TYR84 and four TIP4P-EW waters were modelled into the cavity (**Fig.5.1**). Subsequently all entities were solvated with TIP4P-EW waters using a box size measuring 10 Å. All simulations were parameterised with the ff99SB-NMR and GAFF force fields.

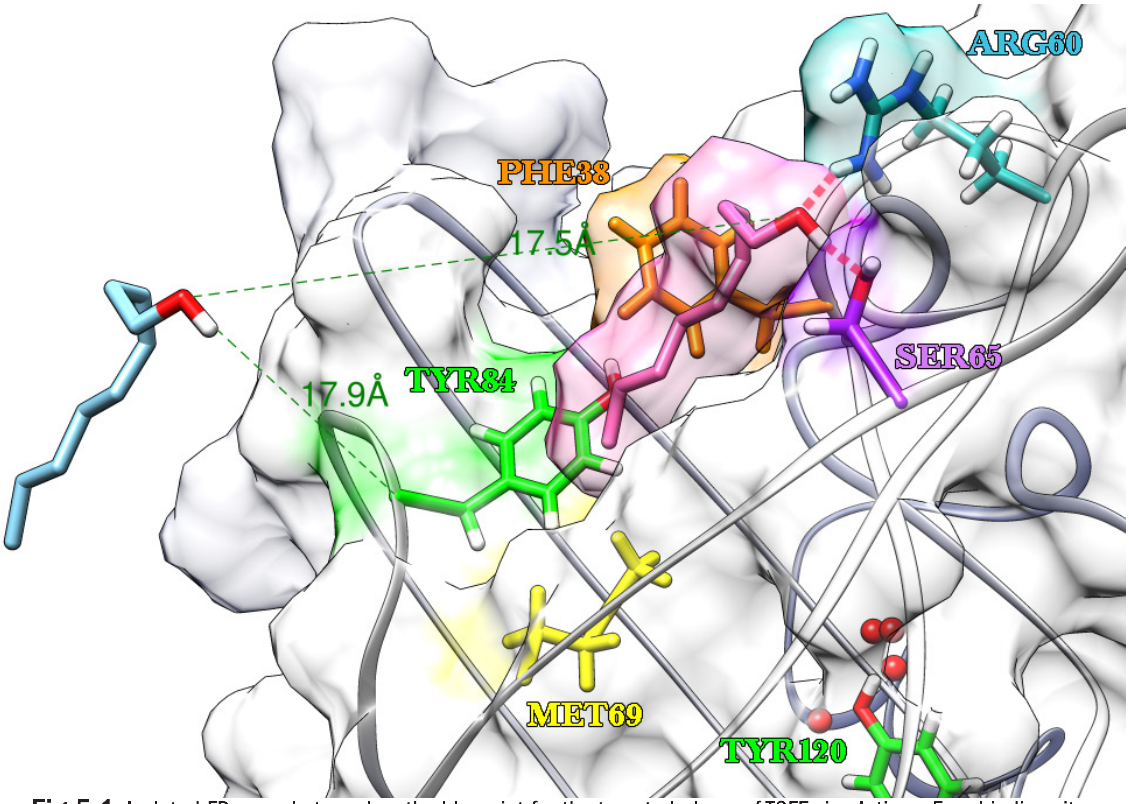


Fig 5.1. Isolated FD snapshot used as the blueprint for the targeted phase of T2FE simulations. Four binding site waters are represented as red spheres. Blue coloured ligand is positioned ~18 Å away from protein in bulk solvent. The pink coloured ligand depicts the target pose. Note the transient H-bonds between the ARG60, SER65 and the ligand.

Two minimisation stages were run for a 1,000 steps each. The first 500 steps used steepest descent, whilst the remainder utilised the conjugate gradient method. A non-bonded cut-off of 12 Å was set. The first minimisation stage held protein and ligand fixed in position using a harmonic restraint of 20.0 kcal/mol Å $^{-2}$, and the second stage released the restraint. This was followed by two short 40 ps NVT stages to equilibrate the volume and temperature to 300 K with a Berendsen thermostat. A time step of 2 fs was used for all subsequent stages. Once again, only the first stage placed a harmonic restraint on protein and ligand in order to allow the crystalline-like solvent lattice to "melt". The final equilibration stage was an unrestrained short 500 ps NPT simulation which, in conjunction with the previous NVT stage allowed both protein and ligand to adjust their positions and reorientate. This was followed by a short 8 ps pre-production stage in which four NMR style restraints were utilised to move the ligand from its location in bulk to the mouth of the calyx. O1, C4, C6 and C9 atoms were respectively targeted to ~3.5 to 5 Å of ARG60, ASN37, LEU40 and MET69 with a strong force constant of 20 kCal/mol (**Fig.5.1**). The restraints were then released and a totally unbiased production stage (140 ns) commenced under NPT conditions (§3.2.3).

5.2.3. Ligand and protein desolvation

Two simple methods were used to estimate the number of water molecules stripped from the ligand upon internalisation, and bound waters displaced from the cavity.

1. The cpptraj command watershell was used to estimate the number of waters in the ligand's first solvation shell (< 4.0 Å)

2. After RMS fitting the protein to a reference frame as described in §4.2.2, the trajectory was analysed with cpptraj's mask command to discover the names of all waters within 4.0 Å of internal protein residues (TYR120, PHE56, LEU101) and the ligand. Subsequently, the distance from each water's oxygen atom to the hydroxyl group of TYR120 was measured. Finally a python script plotted the distances after applying a filter that assessed whether each water molecule remained within 10 Å of TYR120 for more than 20 ps. Waters failing to meet this criterion were not plotted as they were judged to be in bulk. Accepted water molecules were confirmed via visual inspection.

5.2.4. Entropic cost of bound water

The positional entropy contributions for bound waters were calculated using the 3Dh method described in §4.2.4.2 using 60 bins along each axis. The entropy of a free TP4 water molecule was calculated from simulations run using the protocol detailed in §4.2.1.3.

5.2.5. Other analysis

Images and videos of simulation data were made with UCSF Chimera [227]. Graphs and other data analyses utilised Python [226]. The method used for 2D COM density maps is detailed in §4.2.2.

5.3.0. Results & Discussion

5.3.1. Surface encounter complexes

Though the initial FD simulations were a means to an end, they offer some insights into potential metastable states that could be later targeted via US or MSMs for more quantitative analysis. Whilst the solvent-facing surface of MUP is largely covered by extended hydrophilic residues (purple), the hydrophobic trenches (white/purple) that criss-cross its exterior can easily mould themselves around the shape of the ligand (**video.5.1-2**). The predominantly hydrophobic ligands naturally localise to these zones and the protein engages in sporadic H-bonding interactions with their polar moieties so that they are more effectively localised to the surface. These pathways move in an almost peristaltic-like manner (particularly at the apex of the protein near the Ω loop) as the protein changes conformation and these motions are likely to guide ligands towards the mouth of the calyx. **Video.5.1** shows ligands clustered on hydrophobic patches located near the base of the protein. Additionally, note how unattached compounds located in solvent close to the protein become attracted to these hydrophobic patches. On internalisation of 3c9, other ligands approach the mouth from the top and bottom, and one competitor (ThP) attempts to follow 3c9. The high concentration of ligands in such a small volume allows the visualisation of multiple encounter complexes that sometimes separate due to frustrated contacts, only to reattach some time later.

The recapture of ligands by hydrophilic residues near the apex of the cavity is of particular interest (**video 5.2**). The association rate constant of proteins such as MUP will necessarily be smaller than the diffusion controlled limit due to the occluded nature of the cavity. However mobile loops can assist with internalisation by enclosing and holding the ligand. Such a loop is known as a molecular trapdoor and can greatly speed up the process of internalisation [400,428]. Some examples of other proteins with this facility are tyorosyl tRNA synthetase [429], triosephosphate isomerase [430–433] and lactate dehydrogenase [433–435]. From the videos, it would seem that hydrophilic residues at the front of the Ω loop and the L3 loop of MUP seem to fulfil a similar function and this feature is further explored in this chapter.

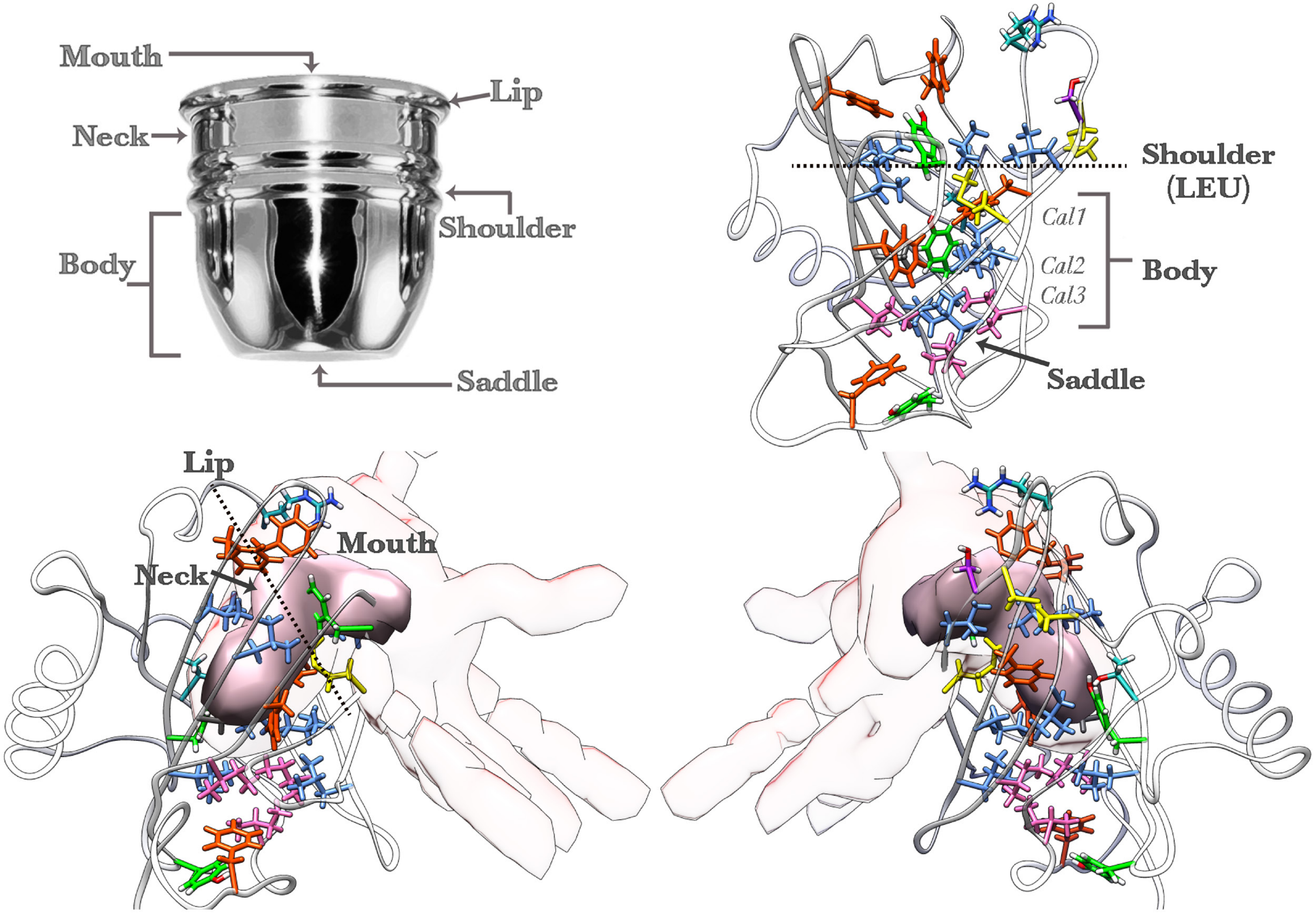
Even after flooding the system with ligand molecules, FD simulations yielded a relatively low rate of initial binding and successful internalisation events. The rate of productive binding events can be greatly enhanced by employing a targeting phase. However, it was desirable to obtain spontaneous internalisation events that were unaffected by the possible bias associated with restraints. Thus, a hybrid protocol was developed whereby the ligand was targeted to the surface of the protein, after which, the restraints were removed. This allowed the ligand to spontaneously internalise or diffuse away. If the position on the surface was chosen correctly, this would greatly enhance the success rate of internalisation. In this case, clues from the FD simulations identified the choice for the target as being one of the surface poses the ligand adopted just prior to the commencement of internalisation (§5.2.1.2). This strategy yielded seven successful internalisation events from twelve discrete simulations. The failed simulations were discarded owing to space limitations.

5.3.2. Overview of ligand internalisation

In order to describe the internalisation process, the gross architectural features of MUP are further labelled in **Fig.5.2** for ease of discussion. As the etymology of the term ‘lipocalin’ is derived from the conjunction of the words “lipo” and “calyx” to describe the hydrophobic cup-like structure formed by the protein fold, relevant terms describing the anatomy of a cup have been appropriated. The “gatekeeper” residues (TYR84, PHE38, LEU40 and MET69) forming the lip of the calyx act to separate the body from the mouth and are situated on the L5 & Ω loops, and *b* & *d* strands. It takes only a relatively small coordinated separation of these secondary structure elements to allow the ligand access to the interior. The neck of the calyx lies between the lip and the shoulder, and the latter is principally composed of aliphatic leucines that provide a hydrophobic environment that assists the internalisation of apolar ligands. PHE56 forms a buttress that supports and partially prevents the inward motion of the gatekeeper residues, LEU40 and MET69. A similar supporting function is provided by LEU105 and LEU116 which act to reinforce PHE38 and PHE114. Both phenylalanines are situated within the large Ω loop and push against TYR84. The very bottom of the calyx is termed the saddle and is principally composed of several aliphatic residues along with TYR80 and the conserved tryptophan, TRP19. These are situated at the base of most of the secondary structure elements composing the body which has already been described and subdivided into the zones termed *cal1*, *cal2 & cal3* (chapter 4.0).

Seven out of twelve stage-2 T2FE simulations successfully underwent ligand internalisation. Differences in ligand pathways to the targeted pose revealed by the 2D distributions in **Fig.5.3** and 3D density maps in **Fig.5.2** demonstrate that reorientation of both protein and ligand during the unrestrained equilibration stages afford a measure of inter-replicate simulation variability (§5.2.1.1-2). After the initial 8 ps targeting phase (displayed as incoming grey tendrils), the ligand explored the area around the mouth of the calyx from between ~15 to 30 ns in three of the seven simulations (**Table.5.1**). The remaining four simulations saw almost instant internalisation for reasons that will shortly become apparent. Also note that the five failed simulations depicted extensive surface exploration but the ligand eventually diffused away due to frustrated contacts. If the targeting phase had localised the ligand further down (along the *d* strand) towards the N-terminus of the protein, it is likely that the failure rate would have been much higher and this approach would only be suitable with the support of substantial

Fig.5.2. Labelled anatomy of MUP. Bottom two images show 3D isosurface that charts the movement of the ligand from bulk solvent into the cavity. Highest density areas of 7 successful internalisation events are within the calyx and around the mouth. See Fig.5.3 for 2D representation.



computational resources. It is clear that this approach will be sensitive to the choice of target location. Without data from the FD simulations selecting an appropriate target would have been much more difficult.

	Ligand Internalisation Time (nanoseconds)				
rep	**Surface Duration**	**Initial Entry**	**Full entry**	**Entry Duration**	**1st *cal3* Entry**
01	0.0	0.0	30.1	30.1	na
02	0.1	0.1	30.8	30.8	23.5
03	14.2	14.2	18.2	4.0	11.4
04	29.3	29.3	38.9	9.6	69.9
05	0.1	0.1	0.8	0.8	2.1
06	16.8	16.8	39.5	22.7	32.3
07	0.5	0.5	51.1	50.5	37.3

Table.5.1. Entry times taken for internalisation events subsequent to the targeting phase. Entry duration is the time the ligand spends on the threshold, whilst full entry is classified as the point when all ligand atoms are within the calyx. Note that the ligand sometimes makes escape attempts, but in all cases internalisation was completed. Rapid internalisation was observed in repeat one because the cavity happened to be wide open at the end of the targeting phase. 1st cal3 entry marks the ligand's first adoption of a vertical pose.

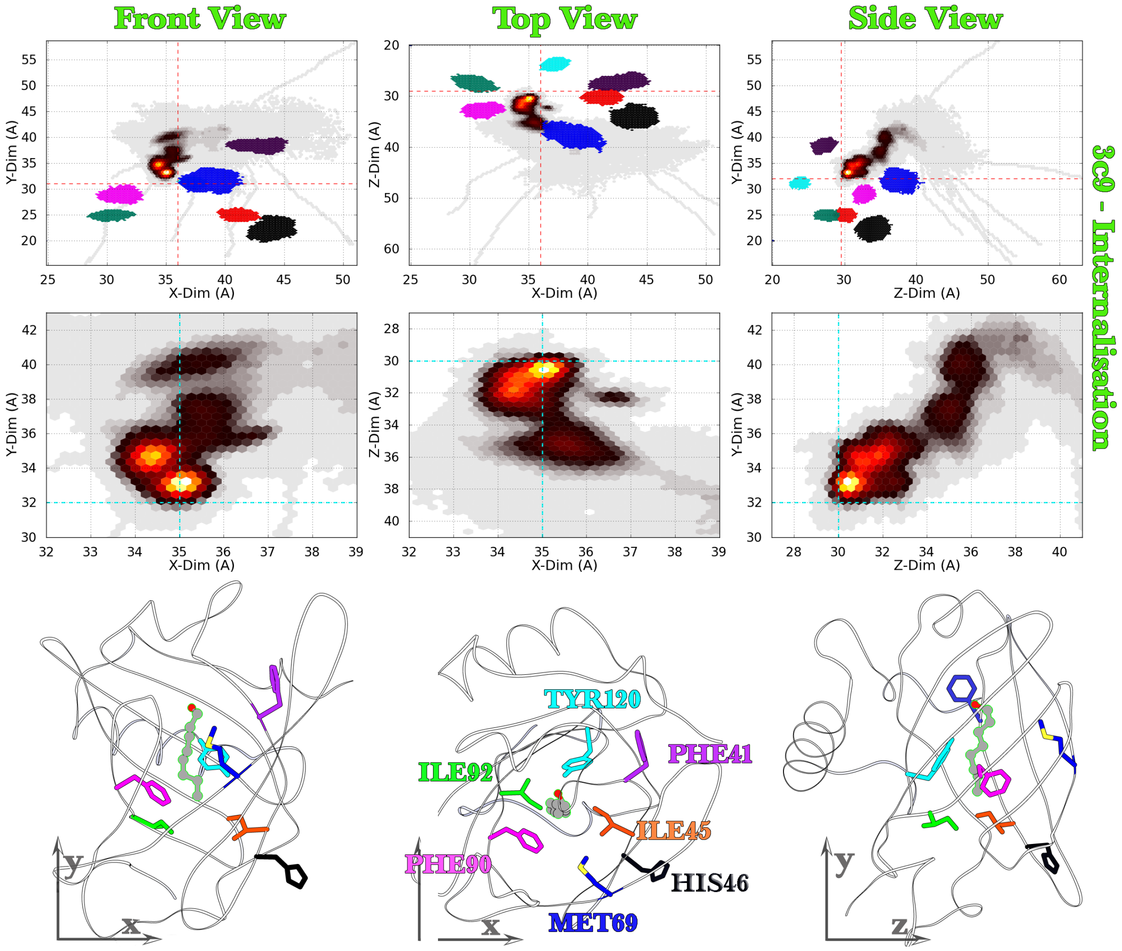


Fig.5.3. 2D density map of the ligand's COM displacements. Data compiled from the concatenated 7 successful 140 ns simulations. Middle panel in bottom row contains key for residue coding. Cyan and red dashed lines are identical to those used in chapter 4.0 (**Fig.4.11**-14).

The process of entry was deemed complete once the ligand ceased protruding through the lip of the calyx, and the duration of time taken to fully occupy *cal1* varied from ~1 to 51 ns (**Table.5.1**). The zy projection in **Fig.5.3** shows two zones of moderate density at (35,37) and (36,40) that are offset along the y axis. The latter corresponds to the volume explored around the mouth of the cavity, whilst the former represents time spent within the neck. Both are favourable metastable states that must be visited before the ligand fully enters and forms H-bonding interactions with residues at the back of the cavity. Usually, the success of internalisation events is measured by the RMSD of the final simulated poses to that of the complexed crystal structure [411,419]. However, this metric is not suitable for ligands that transition between multiple bound minima, so instead, COM distances between the ligand and the hydroxyl group of TYR120 were used to establish the extent of internalisation (§5.3.4.2 & §5.3.4.5). After successful entry and passage through the neck, the ligand adopted horizontal poses in *cal1*. In all of the replicates bar one (rep 01) the ligand then moved from the horizontal pose to a vertical pose that spanned *cal1*, *cal2 & cal3* in times that ranged from ~2 to 70 ns. During the internalisation process some of the repeats (02, 04 and 06) feature brief excursions to escape poses due to ligand conformational changes. Generally, the ligand spends a lot of simulation time settling into the cavity and only adopts vertical poses towards the end. Thus, there is not as much occupancy of *cal3* compared to the equilibrium simulations. (**Fig.5.3** & **Fig.4.14**).

5.3.3. Ligand entry into the occluded calyx

As discussed in chapter 4.0, the H-bond network has a dynamic component as the populations of active H-bonds shift and participants are likely to have multiple alternative partners. Thus, (de)stabilisation of such interactions often cannot be reduced to a binary on/off paradigm and is better represented by the probability of occupancy over longer timescales. The analysis of non-equilibrium simulations of complex systems is hampered by the complexity of the task. It is very difficult to accrue sufficient sampling because the mechanisms studied are usually transient. Furthermore, once the data has been accumulated, sophisticated analyses are required to deconvolute the network of interactions and produce statistically significant results. Whilst some things are clear on viewing crystal structures or simulation data, there are a host of subtle coupled effects that propagate throughout the network of interactions within the protein. Analysis of this is further complicated by the possible existence of degenerate solutions that confound expectations of neat digestible biological mechanisms. Ligand internalisation in MUP is a multi-stage process that is driven by many small coordinated motions in secondary structure elements that can span the entire protein. As this is difficult to analyse, only the initial mechanism that triggers the opening of the calyx and several associated mechanisms will be discussed. However, the detailed causality and interdependence of these will be uncovered by more sophisticated analyses at a later date using techniques such as graph theory.

5.3.3.1. Initial unlocking trigger

Comparing T2FE simulations to the equilibrium simulations obtained in chapter 4.0 reveals that when the protein is unbound, hydrophilic residues situated on loops at the top of the protein participate in an H-bonding network that tighten the protein's structure and assist in keeping the lip of the calyx sealed. The principle participants in this flickering network are termed anchor residues and act to tether L5 and L7, to the front and middle of the Ω loop respectively (**Fig.5.4**). Members include ASP27, LYS28, LYS31, ASN37, ASP85, LYS109 and GLU112. At the front of the protein, additional supporting H-bonds are made between ASN37 to TYR84, and PHE38 to TYR84. All successful internalisation events saw the ligand parting gatekeeper residues in the lip of the calyx and entering hydrophobic tail first. During ligand exploration, the alcohol group was often observed near gatekeeper residues, but close proximity did not trigger internalisation, and it is likely that the hydrophilic head acts to orientate the ligand via *transient* H-bonds to key residues such as ASN37, PHE38, ARG60. On occasion, it is likely that *protein-ligand H-bonds* might assist internalisation via *direct* disruption of protein-protein H-bonds, but a systematic pattern was not observed over seven repeats.

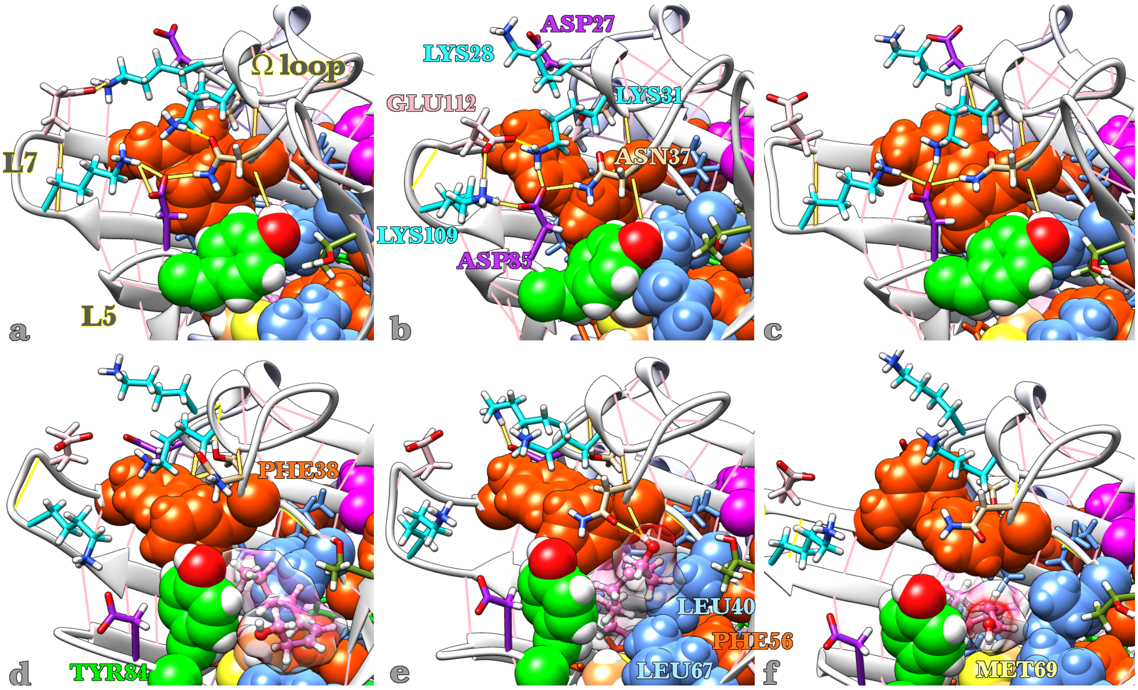


Fig 5.4. (a-c) A few interesting examples of the many permutations of H-bond patterns between anchor residues that act to tether various secondary structure elements to one another. **(d-f)** Ligand mediated disruption of hydrophilic tethers and internalisation events as described in main text. Note that the protein structure is more relaxed than that depicted in the top row.

An examination of the chain of events triggered by protein-ligand steric interactions revealed that when the ligand explores the mouth of the cavity, several important *protein-*

protein H-bond interactions between secondary structure elements were perturbed *indirectly* by the presence of the ligand. The ligand is held within the mouth of the calyx by LEU67 which acts to press it against gateway residues. As the ligand transitions through various conformations, the hydrophobic tail wedges its bulk within the intersections formed by LEU67, PHE56, LEU40 and TYR84. Here, H-bonds locking TYR84 (on the L5 loop) with PHE38 (in the Ω loop) are broken and more importantly, H-bonds that tether anchor residues on the various loops are destabilised. The most immediate repercussions are threefold. Firstly, the loss of these interactions releases the H-bond tethers which limit the motion of the Ω loop. Secondly, relaxation of the inter-secondary structure H-bond constraints imposed by anchor residues allows greater independent movement at the tops of L5 and L7 loops. Thirdly, the ligand is better positioned to exploit the fissure between the L5 and Ω loops and move its tail further into the widening rift (**video.5.3**).

5.3.3.2. Modulation of the Ω loop by arginine residues

The Ω loop functions to block access to the interior of the protein and is characterised by a short ordered 3_{10}-helix that lies to the opposite side of the mouth. Three of the residues (ILE32, PHE38 and LEU40) within this large capping loop are hydrophobic and inward facing, whilst the remainder are hydrophilic and are capable of extensive H-bonding interactions with one another and other polar moieties situated upon various other secondary structure elements. On internalisation, the dynamics of the Ω loop and the other parts of the protein distal to the mouth are modulated by the H-bond interactions of several solvent-exposed arginine residues (**Fig.5.5**). Brief explanations of their putative roles follow. Consult §1.3.1 and **Fig.5.5** for more detail on residue names and canonical secondary structure labels. **videos.5.4-7** are essential to the explanation.

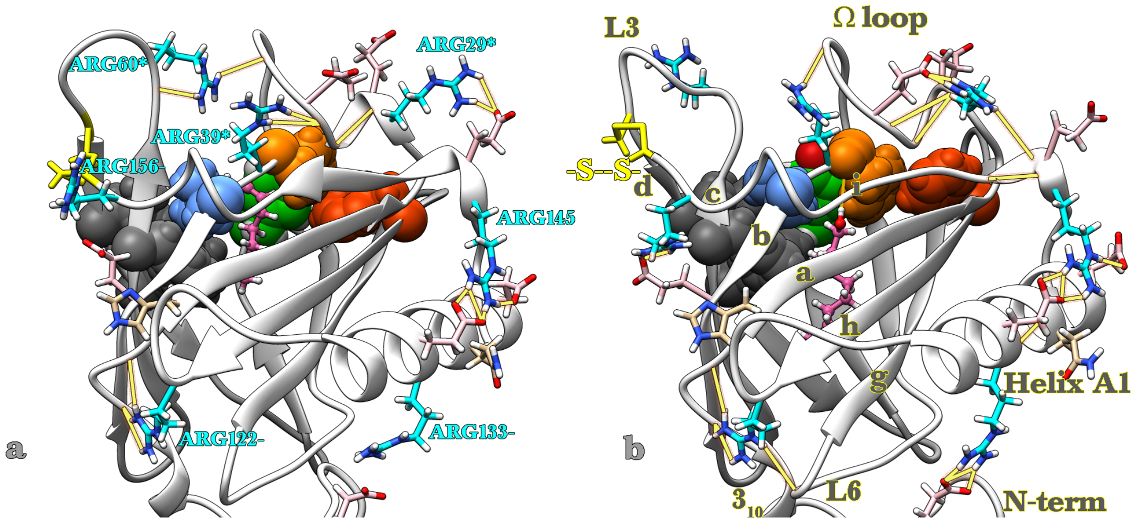


Fig 5.5. Position of arginine residues implicated in internalisation. **(a)** Snapshot taken midway during internalisation. Arginine residues marked with a minus sign are destabilised relative to the holo complex, whilst those marked with an asterisk display complex H-bond patterns. **(b)** Snapshot taken of the fully bound complex.

Note that panels a & b depict ARG29 in upright and flat conformations respectively.

1. ARG39: This residue assists in the immobilisation of the Ω loop in the apo state via the formation of two H-bonds to ILE32 and GLY36. The latter H-bond is disrupted by the ligand's presence in the mouth of the cavity, and when this occurs, this residue can then function as a fulcrum by virtue of its remaining H-bond. However, its function is degenerate or "fuzzy" in that this residue can also act by pulling the L3 loop to the Ω loop via the formation of directional electrostatic interactions.

2. ARG60: This residue is attached to the L3 loop when the protein is in the apo state and is usually H-bonded to this secondary structure element or exploring the mouth of the cavity. In the apo state, this residue rarely makes contact with the Ω loop. But during the process of ligand internalisation it primarily acts in conjunction with ARG39 and THR58 to pull the destabilised Ω loop towards the L3 loop. The effect of the simultaneous anchoring and pulling results in the upwards and lateral displacements of PHE38 (**Fig.5.6** and **videos.5.4-5**). Note that other residues such as TYR84 and PHE114 also move away as a result of global protein relaxation. Thus, function is birthed by the union of structure and dynamics.

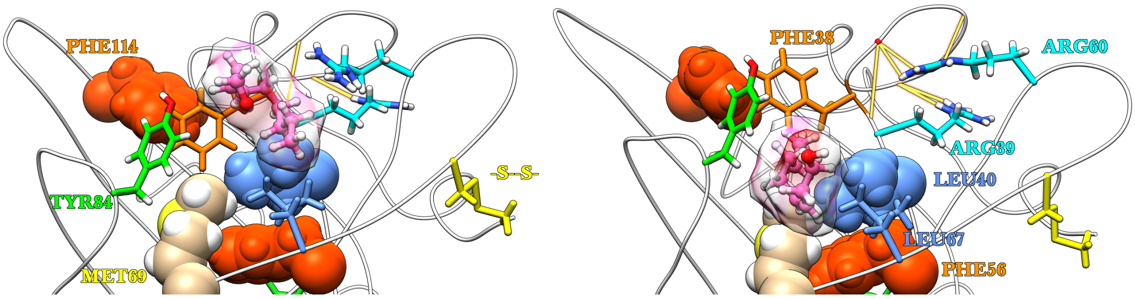


Fig.5.6. H-bonding interactions of ARG60 and ARG39 abet protein conformational changes on ligand entry. Note the large movement of gateway residues such as PHE38 and TYR84. Also observe the general relaxation of the upper part of the protein and conformational changes of other residues such as PHE56 and PHE114. Also see videos.5.4-5.

3. ARG156: This residue is situated in the mostly unstructured loop between the MUP's only disulphide bond and strand *i*. In the apo state, it makes extensive H-bond interactions with residues situated on the side of the protein and its possible function is to stabilise the closure of the gateway residues by reducing the dynamics of the loop connecting the L3 loop to the *i* strand and A1 helix. During internalisation, it ceases to make any H-bonds at all, or preferentially H-bonds to adjacent residues within the loop e.g. CYS157 or ALA154. This is likely to assist relaxation of the protein's structure and facilitate internalisation.

4. ARG29: The precise manner in which this residue acts is difficult to characterise, but it is notable in that it can make H-bond interactions to adjacent glutamates (GLU30 and GLU33) on the Ω loop and residues (such as GLU146) on the structured turn immediately preceding strand *i*. Furthermore ARG29 exhibits two distinct conformations. The first is extended and upright, whilst the second lies flat and is orthogonal to the first. In the apo state, the latter conformation might stabilise and assist in the locking down of the Ω loop. *Per contra*, during ligand internalisation, the former conformation may increase the dynamics of the Ω loop by H-bond mediated modulation of residues to either side. PCA analysis reveals a certain amount of motional synchronicity with ARG60 during internalisation (**videos.4.4-5**). It should be noted that its precise mechanism is not completely understood and other unidentified elements may also play a part in its function. Nonetheless, the interactions mediated by this residue are part of a linked network that is likely to propagate dynamics from the Ω loop to the N-terminus. The next intermediary along this particular chain of H-bond mediated interactions is ARG145.

5. ARG145: This residue is situated on the structured turn immediately preceding strand *i* and it forms relatively stable H-bond interactions with hydrophilic residues situated on the A1 helix such as GLU132, GLU139 and GLN136. Differences in H-bond patterns occur as this residue tilts to either side and this is likely to help shift the position of the large α-helix. This possibly acts as a steering mechanism that assists the modulation of the opening and closing of gateway residues and the global relaxation and tightening of the protein's structure.

6. ARG133: The A1 helix is positioned roughly near the middle of the protein and is well positioned to link the top of the protein to the bottom. The second arginine, ARG133 forms the link between the A1 helix and the N-terminus as it sometimes binds to residues such as GLU2. Prolonged H-bond interactions are likely to draw the N-terminus up into the body of the protein and are observed more often in holo simulations compared to apo. This interaction is likely to tighten the structure of the protein.

7. ARG8: Situated in the N-terminus and is likely assist stabilisation of interactions there. Not assessed in this chapter.

8. ARG122: In the Calycin family, this residue is generally an arginine or a lysine and lies over a conserved tryptophan. It is interesting in that, it is capable of forming extensive H-bond interactions that link TYR87 in the L6 loop to the L4 loop in the bound state. This interaction is associated with greater stabilisation of the protein's structure. In the apo state, and during the process of ligand binding, this residue ceases to make H-bond interactions with the L6 loop and sporadically makes directional H-bonds with the 3_{10} helix located within the N-terminus. It is likely that the loss of interactions with the back of the protein assists the relaxation of the protein's structure.

A series of 4 videos have been created to highlight the role of these arginine residues in binding (**videos.5.6-9**). Please consult **Table.5.2** for a synopsis of arginine behaviour. Note that the internalisation videos display a mix of features from endpoint apo and holo states. Therefore the behaviour of the protein should be assessed with regard to the position of the ligand. Furthermore, note that the global structure of the apo protein tends to be more relaxed than that of the holo complex and is therefore likely to visit a broader range of conformations.

	MUP Apo	**MUP Holo 3c9**	**Internalisation Rep-05**	**Internalisation Rep-06**	**Notes**
ARG60	Docked to L3	Docked to L3	Pulls Ω loop	Pulls Ω loop	Degenerate Function
ARG39	Stabilising Ω loop	Stabilising Ω loop	Pulls Ω loop (fulcrum)	Pulls L3 loop	
ARG156	Docked to side	Docked to side	H-bonds destabilised	H-bonds destabilised	-
ARG29	Subtle H-bond patterns	Subtle H-bond patterns	Subtle H-bond patterns	Subtle H-bond patterns	Needs statistical analysis
ARG145	Subtle H-bond patterns	Subtle H-bond patterns	Subtle H-bond patterns	Subtle H-bond patterns	Needs statistical analysis
ARG133	Bound to A1 Helix H-bonds	Prefers N-terminus H-bonds	Endpoint H-bond transition	Endpoint H-bond transition	
ARG122	H-bonds destabilised	H-bonds stabilised	Endpoint H-bond transition	Destabilised. Too short for transition	
N-Terminus	Pulled up	Dropped down	Dropped down	Dropped down	Observe TYR97 & PHE10
Video	5.6	5.7	5.8	5.9	

Table.5.2. Brief synopsis of behaviour of arginine residues in videos.5.6-9.

In addition to the arginine residues discussed, it is possible that surface-exposed histidines also play a part in ligand internalisation. These pH sensitive residues are strategically placed near the junctions of several key secondary structure elements. Changes in protonation could well assist in relaxation and tightening of the protein's structure so as to aid or inhibit conformational rearrangements. This would be consistent with the putative role of MUP-I as an agent that delays the release of VOCs. As it is expressed in areas associated with pheromone production such as the liver and kidneys prior to being excreted in urine, these different regions are likely to have different pHs. Furthermore, excreted urine is likely to change pH due to external factors and aging. Other studies have uncovered pH-dependent ligand binding in other lipocalins such as human tear lipocalin [389,436], β-lactoglobulin [437–441] and retinol binding protein [442,443] amongst others.

5.3.4. Protein cavity & ligand desolvation on binding

5.3.4.1. Methods that characterise desolvation on binding

The global entropy of binding can be greatly affected by solvent reorganisation. This is a difficult contribution to evaluate because water is diffusive by nature and there is a large number of water molecules present in typical biological systems. The most commonly calculated subcomponents are ligand and cavity (de)solvation. Intuitively, further contributions could also arise from disparities in receptor and complex solvation at the protein's surface. However, these are usually assumed to cancel, and are thus overlooked in discussions of protein-ligand binding. The remaining sections of this chapter examine the first two desolvation-related subcomponents.

Ligand desolvation: As discussed in §1.1.5, ligand solvation entropy can be experimentally estimated via vapour-solvent partition equilibrium experiments [35,444]. As ligand binding in MUP is characterised by the ligand moving from bulk solvent into a suboptimally hydrated cavity, ligand desolvation entropy is obtained by inverting the sign of the measured solvation entropy. When these partition experiments cannot be carried out, group contribution methods such as that proposed by Plyasunov and Shock (2000) [94] have been used to obtain theoretical estimates [36,180].

One of the first estimates of the impact of ligand desolvation in MUP was obtained by Sharrow et al. (2003) who examined the transfer of several small n-alkanols and SBT analogues from water to cyclohexane [388]. They established that the entropy of desolvation provides a favourable contribution that increases linearly with increasing surface area. This type of partition experiment mimics the transfer of a hydrophobic compound to a hydrophobic pocket and is thus smaller than that obtained from vapour-solvent experiments. Later partition equilibria studies on n-alkanols and 3Z-olefins demonstrated that the desolvation contributions of these compounds would display strongly linear thermodynamic binding signatures [36,180]. However, such experiments only provide a measure of the maximal possible ligand desolvation entropy *as this contribution is critically dependent on the number of waters stripped from the inhibitor* upon entry into the cavity. The computational results presented in §5.3.4.2 provide an estimate of this number.

Protein cavity desolvation: It is very difficult to experimentally ascertain the number of waters expelled from the cavity during ligand binding and those that remain subsequent to the event. X-ray crystallography is capable of capturing ordered waters, but it is not an effective method for the quantification of the number of disordered water molecules present in the cavity. This is because these waters are not localised to a single defined minimum and their electron densities are poorly defined [445]. NMR spectroscopy is a more promising approach, and in the most favourable cases it identifies more dynamic waters than X-ray crystallography. Whilst X-ray diffraction is limited to a detection rate of 10 to 50%, NMR spectroscopy methods can report on water present at levels as low as 10% [320]. Moreover, details on the dynamic characteristics of water can be recovered e.g. orientational fluctuations and the rate of exchange with bulk [320]. Other notable, yet oft neglected methods that allow quantification of the displacement of bound waters are ultrasonic densimetry and quartz crystal microbalance experiments [446–448]. However, experimental methods provide results that are averaged over long time scales and do not tend to yield information on brief excursions of macromolecules from the equilibrium ground state to functionally important excited states. The structural and dynamic details of these rare events can be reproduced by MD. However, results have traditionally been limited by the relatively short timescales covered by atomistic simulations and the paucity of corroborating experimental techniques. Recently, rapid improvements in processing power and technologies have overcome some of these limitations. For example, millisecond length molecular dynamics simulations of bovine pancreatic trypsin inhibitor (BPT1) have been used to successfully probe the accuracy of modern day force fields via comparison to relaxation dispersion experiments [449].

In the case of MUP, the precise number of waters expelled upon the entry of different ligands is unknown. Moreover, a rigorous thermodynamic assessment of the entropic benefits associated with the ejection of such waters is currently unavailable. The computational results presented in §5.3.4.5 lay the groundwork for the accurate estimation of these values.

5.3.4.2. The extent of ligand desolvation on calyx entry

T2FE simulations allow an atomistic view of the binding of 3c9 to MUP. Whilst the flexible ligand is free in solution, the number of water molecules in the first hydration shell fluctuates as the ligand transitions between compact and extended configurations (**Fig.5.7**). It then rolls across the surface of the protein and enters the cavity via a narrow passageway located beneath the shroud of the Ω loop (**Fig.5.7**). During this process, the ligand is divested of its waters, and it finally enters the calyx wherein it adopts poses akin to that observed in the crystal structure.

Each T2FE simulation began with 4 waters in the binding cavity and finishes with 0 to 1 waters (**Table.5.3**). This indicates that ligand internalisation is associated with the net egress of waters and is not counterbalanced by the entry of waters from bulk. On plotting the number of waters in the ligand's first solvation shell versus time, it is apparent that the compound's passage through the narrow neck of the calyx results in *total ligand desolvation* (**Fig.5.8.a-b**). These observations were further corroborated by visual inspection of the trajectories.

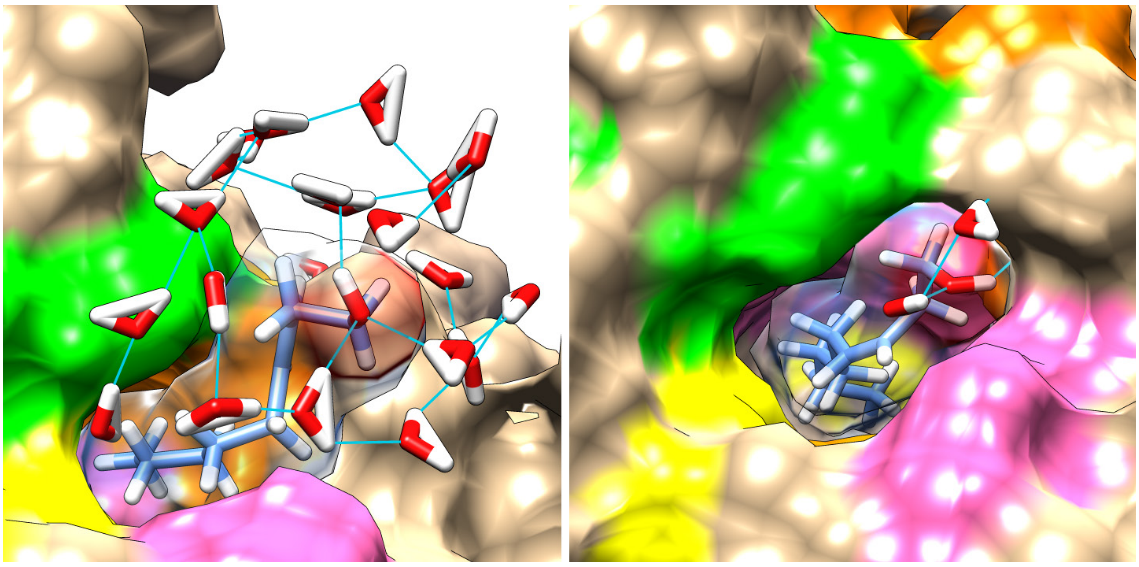


Fig.5.7. 3c9 is coloured blue and is depicted as a stick model surrounded by a translucent surface. Waters within 4Å of ligand are displayed as a stick model and hydrogen bonds are displayed as blue lines. The ligand is entering MUP's occluded binding cavity whilst waters preferentially bind to each other or the ligand's hydroxyl group. Right image was taken 150ps subsequent to left.

	No of waters at end of T2FE simulations						
Repeat	**01**	**02**	**03**	**04**	**05**	**06**	**07**
Final Waters	0	1	1	1	1	1	0

Table.5.3. Number of waters in calyx at the end of each 140 ns simulation.

Even though the evidence has been obtained from the binding of only one ligand over 7 T2FE simulations, the structural characteristics of the narrow passageway into the calyx mean that it is likely that other ligands considered in this work would be desolvated to a similar extent. Prior to this analysis the true extent of ligand desolvation was not quantified and the desolvation entropy calculated from vapour-solvent partition equilibrium experiments only offered an upper bound. This result indicates that the majority of the experimentally calculated desolvation entropy would be available to the system upon binding.

In the absence of partition equilibria data for 3Z-olefins, Malham (2012) used the group contribution method proposed by Plyasunov and Shock (2000) to calculate theoretical desolvation contributions [94,180]. Such an approach is extremely useful because it allows the fast evaluation of large panels of ligands. Furthermore, it allows the assessment of compounds that are not amenable to partition-equilibria experiments. This particular method indicates that n-alkanols enjoy a larger desolvation entropy contribution compared to ligands with a double bond. However, it was surprising that the method did not distinguish between double bond stereochemistry and consequently 3Z-olefins and 2E-olefins have identical desolvation contributions (**Table.5.4**).

	Desolvation entropy (kJ/mol)		
Carbon no	**n-alkanols**	**3Z-olefins**	**2E-olefins**
6C	48.0	43.4	43.4
7C	52.4	47.9	47.9
8C	56.9	52.4	52.4

Table.5.4. Ligand desolvation entropies calculated at 298 K using the method proposed by Plyasunov and Shock (2000) [94]. Values taken from Malham (2012) [180].

To assess whether MD simulations could give an indication of the relative solvation of the different ligand panels, the numbers of waters in the ligand's first solvation shell were calculated. These results indicate that these small molecule inhibitors are solvated to similar extents (**Fig.5.8.c-d**). Relatively small variations in the distributions arise because structural differences result in differences in conformational exploration. As discussed in §3.3.2, 3Z-olefins adopt more compact conformations than n-alkanols and the decrease in exposed surface area results in a small reduction in solvation which is in line with the decreased desolvation entropy reported in **Table.5.4**. A comparable ligand with a trans double bond should therefore be associated with a small increase in the number of waters in its first solvation shell because the ligand will be predisposed to adopt extended conformations. Thus, the desolvation entropy should be slightly more favourable compared to an analogous saturated compound.

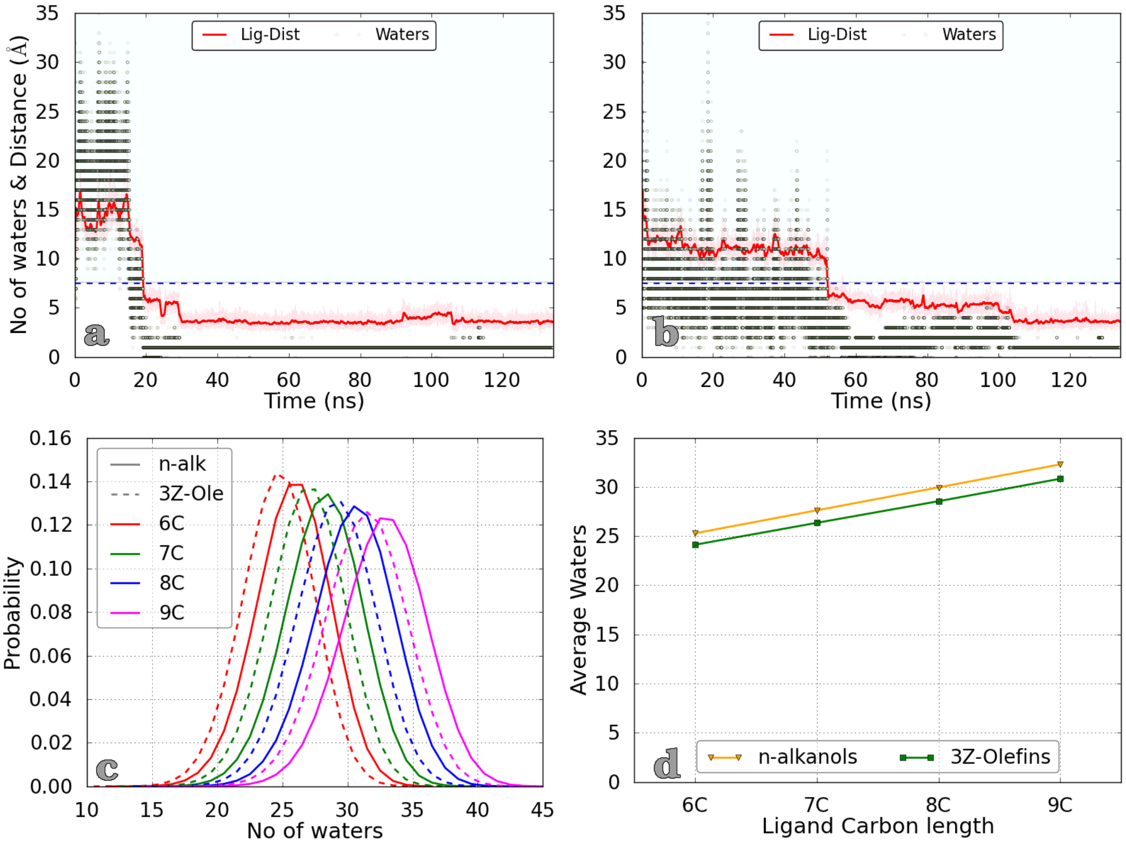


Fig.5.8. (a-b) Ligand desolvation as a function of distance from the binding site for two discrete repeats. Distance between ligand's COM to TYR120 is depicted as a red line. Green dots represent the number of waters within the ligands first hydration shell (< 3.4 Å). Dashed line represents boundary between the cavity and bulk. Increases in the number of water molecules subsequent to binding are the result of the ligand coming close to the lip of the

cavity. The y axis scale is appropriate for both the distance (Å) and number of waters. **(c)** Distributions of the range of waters in first solvation shell for n-alkanol and 3Z-olefin ligands in the equilibrated free state. **(d)** The average number of waters calculated from c.

As the average number of waters plotted between analogous ligands in **Fig.5.8.d** only differ by a single water molecule, it is unlikely that this would be associated with any significant error when assessing small compounds. However, solvation differences between larger, more complex compounds could potentially be much greater.

It would be useful to examine how ligand desolvation entropy fits into the global entropy of binding, and such an analysis will be carried out in the conclusion (§5.4.1.2) after the expulsion of cavity waters is considered in §5.3.4.6. But first, let us consider how transient H-bonds assist ligand internalisation and the nature of water mediated H-bonds within the cavity.

5.3.4.3. H-bond reorganisation within the calyx subsequent to ligand entry

Multiple T2FE simulations captured the binding of 3c9 to MUP and data obtained from ligand internalisation demonstrates how transient H-bonds assist in ligand translocation from an unbound pose to one akin to that found in published crystal structures. The data also supports the hypothesised ligand entry pathway put forward by Timm et al. (2001) [72]. The transient nature of the H-bond network in the large cavity means that during internalisation, there are multiple combinations of ligand conformational changes coupled with positional shifts within the cavity. The following description depicts one of these, and is chosen on the basis that it illustrates the architecture of the cavity along with dynamic elements pertaining to protein, ligand and water. **video.5.10** charts the internalisation of a different repeat (05) with 4 waters and viewers will note the inter-replicate similarities. A short time after that covered by this video all waters but one is expelled. This phenomenon is further discussed in §5.3.4.5.

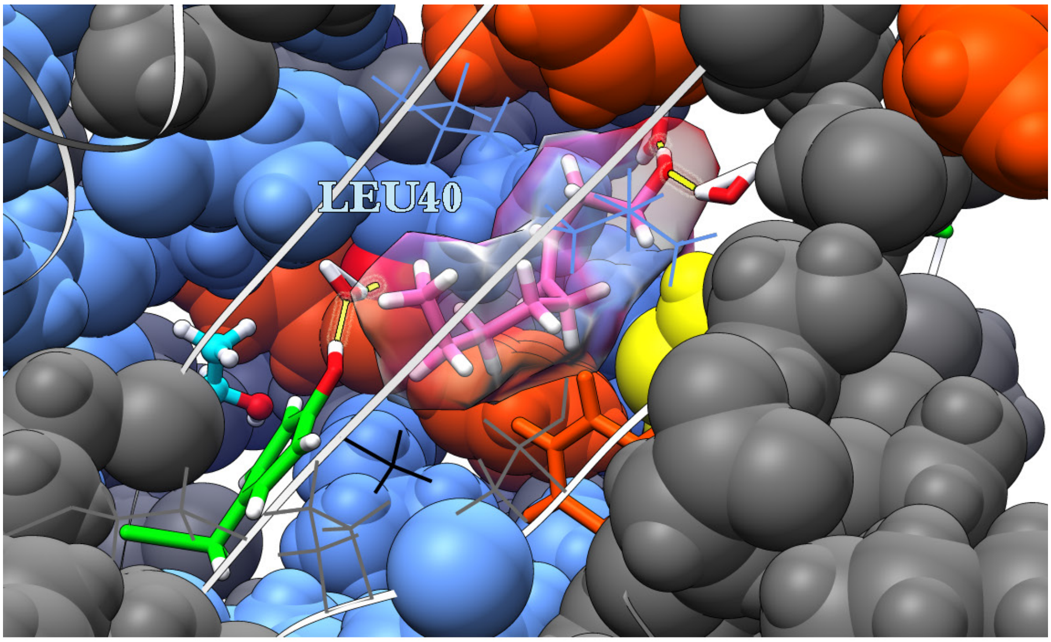
As the ligand enters the cavity, hydrating water molecules are stripped off due to the narrow confines of the entranceway (**Fig.5.7**). The three panels in **Fig.5.9** show events subsequent to entry.

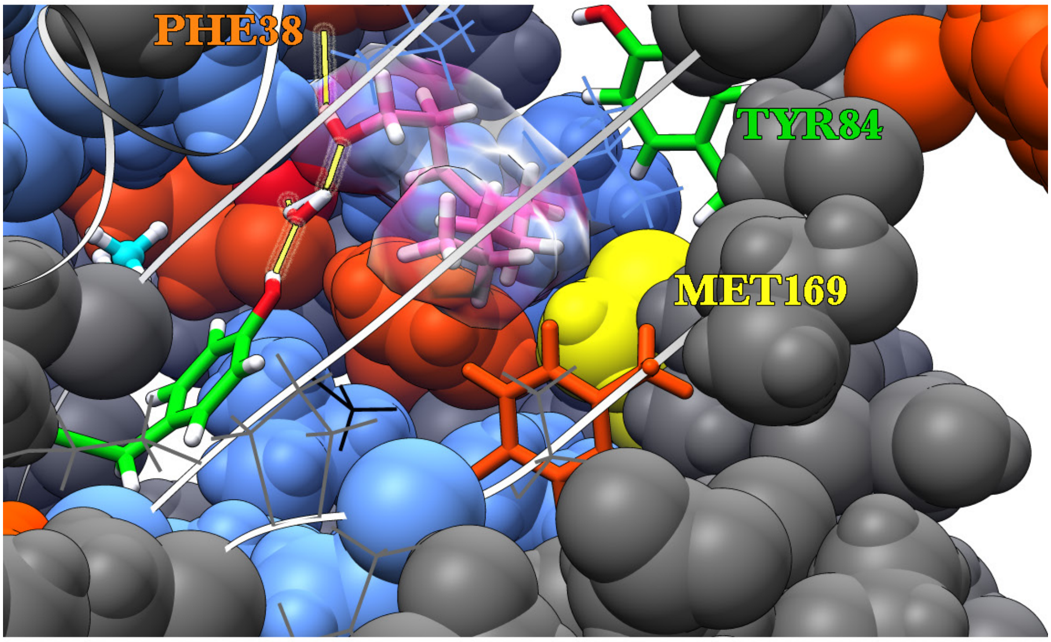
Fig.5.9. The triptych shows key ligand conformational changes that occur subsequent to internalisation. The images were generated from a free diffusion simulation that captured the binding of 3c9 to MUP. The images show a cutaway section of *cal1* (and the top portion of *cal2*) from the side. The Ligand is represented as a pink stick model and is covered by a translucent surface to demarcate the extent of its van der Waals surface area. Water molecules are depicted as triangles and H-bonds are coloured yellow. All amino acid residues are represented by space filling molecules unless otherwise stated. Phenylalanines are coloured orange-red and PHE90 is detailed as a stick model. All tyrosines are shown as stick models with green carbon atoms - TYR84 lies near the entrance of the cavity, whereas TYR120 adopts a more buried position. THR21 is shown as a ball and stick model and its carbon atoms are coloured cyan. ALA103 is depicted as a black wire and the MET169 residue that forms part of the gateway is coloured yellow. Leucines are coloured light blue and LEU105 and LEU116 are shown as wires to maintain a clear view. All other residues are coloured grey. See main text for further details.

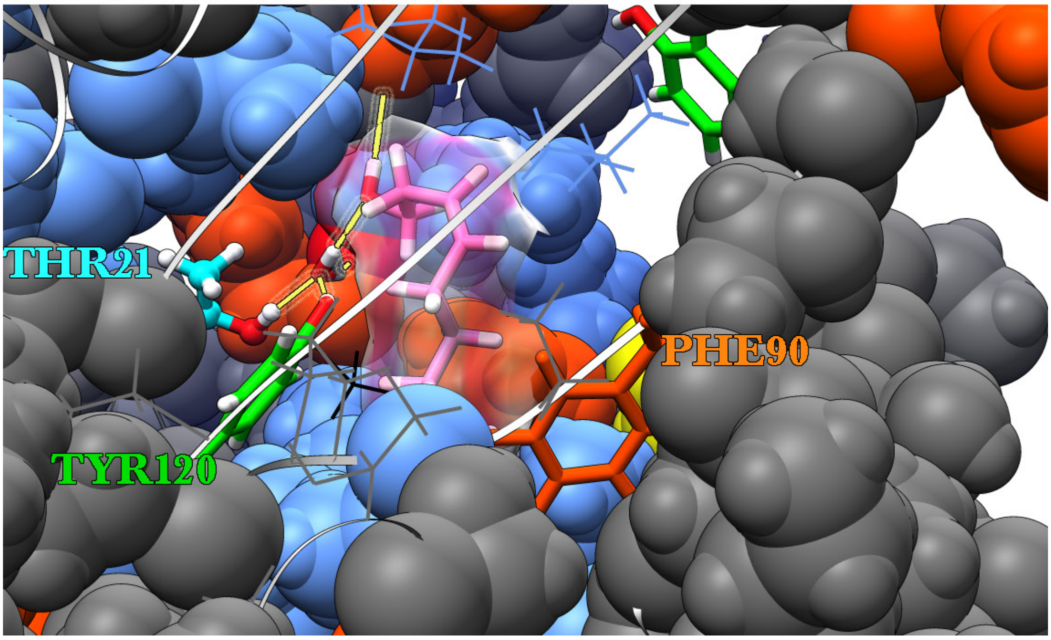
Top panel: The side view focuses on the largest *cal1* chamber. The top portion of *cal2* is shown, whilst *cal3* is hidden below. 3c9 enters the calyx hydrocarbon tail-first. Its alcohol group is positioned and orientated far away from TYR120 and it still maintains H-bonds to a couple of waters located within bulk solvent. A single binding site water is engaged in H-bond interactions with TYR120 and the oxygen (helpfully coloured red) within the amino acid backbone of LEU40.

Centre panel: The ligand has fully entered the calyx and is no longer capable of making H-bond interactions with bulk solvent. The entrance way into *cal1* is now blocked via the closure of TYR84. The alcohol moiety can form a number of transient H-bond interactions with several amino acid backbone atoms. In this snapshot, the backbone oxygen of PHE38 helps steady the head of the ligand with an H-bond interaction. Additional stabilisation is provided by a water molecule that is itself anchored by polar interactions to TYR120 and LEU40. Stabilisation of the hydroxyl group allows the ligand's flexible hydrocarbon tail to bend back on itself and reorientate the polar portion of the polymer so that it is closer to TYR120. Note that the interactions formed by this water molecule with these amino acids can also be observed in MUP holo crystal structures (1ZND, 1ZNE, 1ZNG, 1ZNH & 1ZNK).

Bottom panel: The binding site water forms an additional H-bond interaction with THR21, and in this snapshot, PHE90 undergoes a conformational change that allows the ligand's hydrocarbon tail to be displaced down into *cal2* and *cal3*. Though there are several amino acids to which H-bonds can be created, it is important to remember that any strong protein-ligand interaction will (directly or indirectly) stabilise the hydrophilic head of the ligand. The longer hydrophobic tail is relatively free, and is only limited by the confines of the calyx. The dynamic motion of the flexible tail is thus responsible for the large range of displacements seen in the COM of the ligand and is likely to exert an opposing destabilising effect on the head (§4.3.3.1). Moreover, the availability of H-bond donors and acceptors (apart from TYR120) is inconsistent because backbone


LEU40


PHE38
TYR84
MET169


THR21
PHE90
TYR120

atoms are likely to be occluded as the protein changes conformation (§4.3.5). This inconsistency is likely to introduce an additional degree of instability to the head of the ligand which should be affected by fluctuations in H-bond availability.

5.3.4.4. Water mediated H-bonds: equilibrium simulations revisited

At this juncture, the equilibrium simulation analysed in Chapter 4.0 are revisited to better understand the nature of bound waters in MUP. A thorough understanding of the H-bonding network is complicated by the capability of bound water molecules to mediate bridged interactions between bound ligand and protein (**Table.5.6**). A complete characterisation of the thermodynamics and dynamics of bound water in MUP is beyond the scope of this work as the methods required to accomplish this are non-trivial. Even the design of a simple metric that counts the number of bound waters during a simulation is challenging due to difficulties in accurately demarcating the binding site volume to exclude bulk solvent. Moreover, the subsequent processing of the millions of snapshots required for converged statistics would double the storage requirement at a minimum. As the equilibrium simulations contain 9.6 x 10^6 snapshots, a rough indication of the number of bound waters is provided in **Table.5.5** using the first method described in §5.2.3.

		Waters per aggregate Simulation				
Ligand	**Starting Waters**	**Mode**	**Mode Occupancy (%)**		**Mean**	**Std Dev**
hex	3	0	19.8		1.35	1.26
hep	3	0	34.2		0.83	0.96
oct	1	1	66.5		1.02	0.50
non	1	1	55.8		0.96	0.60
3c6	2	1	61.9		1.27	0.79
3c7	2	0	42.9		1.02	1.20
3c8	2	1	36.2		0.91	0.87
3c9	2	1	29.4		1.75	1.00

Table.5.5: Statistics detailing the number of waters in n-alkanol & 3Z-olefin equilibrium simulations.

Even though the number of bound waters is small (typically < ~3), the classification of water mediated H-bonds and their occupancies is limited by the cost of calculating the manifold configurations of interacting partners. Hence, the data in **Table.5.6** focuses on bridged interactions i.e. water molecules that are H-bonded to two or more solute residues simultaneously. The analysis splits larger linked H-bonded networks into smaller parts that fulfil the definition given in the previous sentence (e.g. **Fig.5.10.b**). Thus, the synchronicity of bridged interactions is not detailed. The trends correlating ligand size and structure with direct H-bond occupancies (§4.3.5) are also echoed within the water mediated H-bond network. However, the correlation is somewhat weaker

Water Bridged hydrogen bonds: n-alkanols					**Water Bridged hydrogen bonds: 3Z-Olefins**			
Residue 1	**Residue 2**	**Ligand**	**Occupancy (%)**		**Residue 1**	**Residue 2**	**Ligand**	**Occupancy (%)**
LEU40	TYR120	hex	14.1		ALA103	LEU116	3c6	47.8
LEU40	-	hex	12.9		LEU40	-	3c6	18.1
TYR120	-	hex	8.6		ALA103	-	3c6	12.9
PHE38	-	hex	6.8		LEU116	-	3c6	11.5
THR21	LEU40	hex	5.4		TYR120	-	3c6	8.2
LYS31	-	hex	5.2		*Unique Combinations under 4% occupancy = 54*			
Unique Combinations under 4% occupancy = 70								
					ALA103	LEU116	3c7	16.8
LEU40	TYR120	hep	29.4		TYR120	-	3c7	4.8
LEU40	-	hep	12.3		LEU116	-	3c7	4.8
PHE38	-	hep	9.4		*Unique Combinations under 4% occupancy = 47*			
TYR120	-	hep	6.2					
THR21	LEU40	hep	5.0					
Unique Combinations under 4% occupancy = 89					ALA103	LEU116	3c8	17.1
					LEU40	TYR120	3c8	7.5
					THR21	LEU40	3c8	6.4
LEU40	TYR120	oct	59.9		LEU40	-	3c8	5.6
LEU40	-	oct	25.8		TYR120	-	3c8	4.6
TYR120	-	oct	9.5		LEU116	-	3c8	3.9
THR21	LEU40	oct	9.1		*Unique Combinations under 4% occupancy = 56*			
Unique Combinations under 4% occupancy = 61								
					LEU40	-	3c9	21.7
LEU40	TYR120	non	42.6		LEU40	TYR120	3c9	18.7
LEU40	-	non	22.1		THR21	LEU40	3c9	15.7
THR21	LEU40	non	9.0		TYR120	-	3c9	12.7
TYR120	-	non	7.3		ALA103	LEU116	3c9	10.9
Unique Combinations under 4% occupancy = 100					PHE38	-	3c9	5.5
					PHE38	LEU40	3c9	4.8
					LYS31	-	3c9	4.4
					Unique Combinations under 4% occupancy = 73			

Table.5.6. n-alkanol & 3Z-olefin bridged H-bond interactions.

due to fluctuations and differences in the number of bound waters throughout the time course of the simulation. For example, as 3c7 simulations have zero bound water molecules (mode 0, for 43% of simulated time) bridged H-bonds have relatively lower occupancies than that observed for the other simulations. Despite these caveats, it is of interest to consider the dataset so as to expand our understanding about the nature of H-bonding within the cavity.

If H-bond interactions for all ligands are considered, it is clear that TYR120 is spatially positioned at the nexus of a cluster of residues that constitute an outer frame with which the most common polar interactions are made (**Fig.5.10**). As demonstrated by Barratt et al. (2005), it is likely that this hydrophilic sidechain is important in terms of creating a microenvironment that localises the polar groups of bound ligands and water molecules to its immediate vicinity [90]. THR21 is immediately adjacent to it and contributes to a small hydrophilic microenvironment in the predominantly hydrophobic calyx. As this residue is often occluded by the motions of other sidechains it does not participate in direct H-bonds with the ligand to the same extent as TYR120. However, this does not prevent it from interacting with water molecules because they are small and much more mobile.

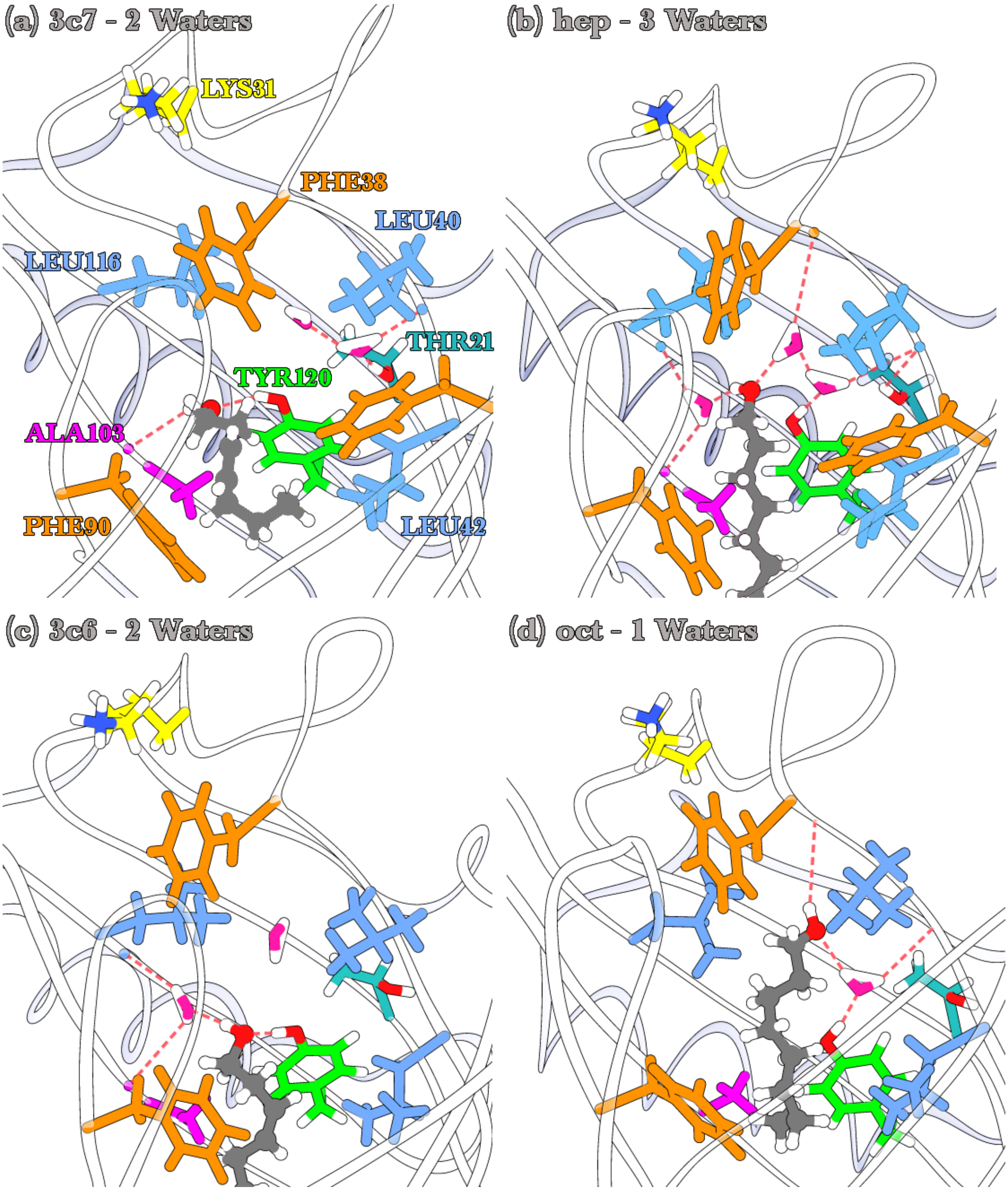


Fig.5.10. Examples of different h-bonding patterns observed in 3Z-olefins and n-alkanol holo simulations. The figures provide a flavour of various possible water mediated H-bonds in the calyx. See main text for details.

When n-alkanols are bound, the main site of water occupation is situated near TYR120, THR21 and Ω loop residues such as LEU40 (**Fig.5.10.d**). However, the hydroxyl head of bound ligands also contributes to the hydrophilic microenvironment within the calyx and its motions further perturbs the water mediated H-bond network. Differences in 3Z-olefin ligand dynamics see the occupation of an additional hydration site that possesses a greater probability of existence than that observed for n-alkanols. This involves ALA103 and LEU116 and is situated to the opposite side of the calyx from the first site (**Fig.5.10.c**). On rare occasions, bound waters and compounds find themselves favourably positioned so that the H-bonding network spans *cal1* and bridges both hydration sites (**Fig.5.10.b**). The examples of H-bond networks in **Fig.5.10** provide a flavour of the interactions between bound entities and the structural features of the calyx. It is important to realise that while the examples might give the illusion of H-bonding stability, the predominantly apolar environment of the cavity and suboptimal hydration results in these polar interactions possessing a mercurial nature. Thus, the positions of some bound waters will be in a constant state of flux within the dewetted cavity. Simply put, there are not enough synchronously available H-bond partners for *all the waters* to adopt fixed positions. Hence, webs of transiently stable H-bonds are destabilised by the creation of new internal networks, comprised of competing H-bond donor and acceptors whose availability is constantly modulated as the Brownian motions of bulk solvent drive the protein to adjust its internal conformation. The relative order of bound water molecules is further detailed in the next section.

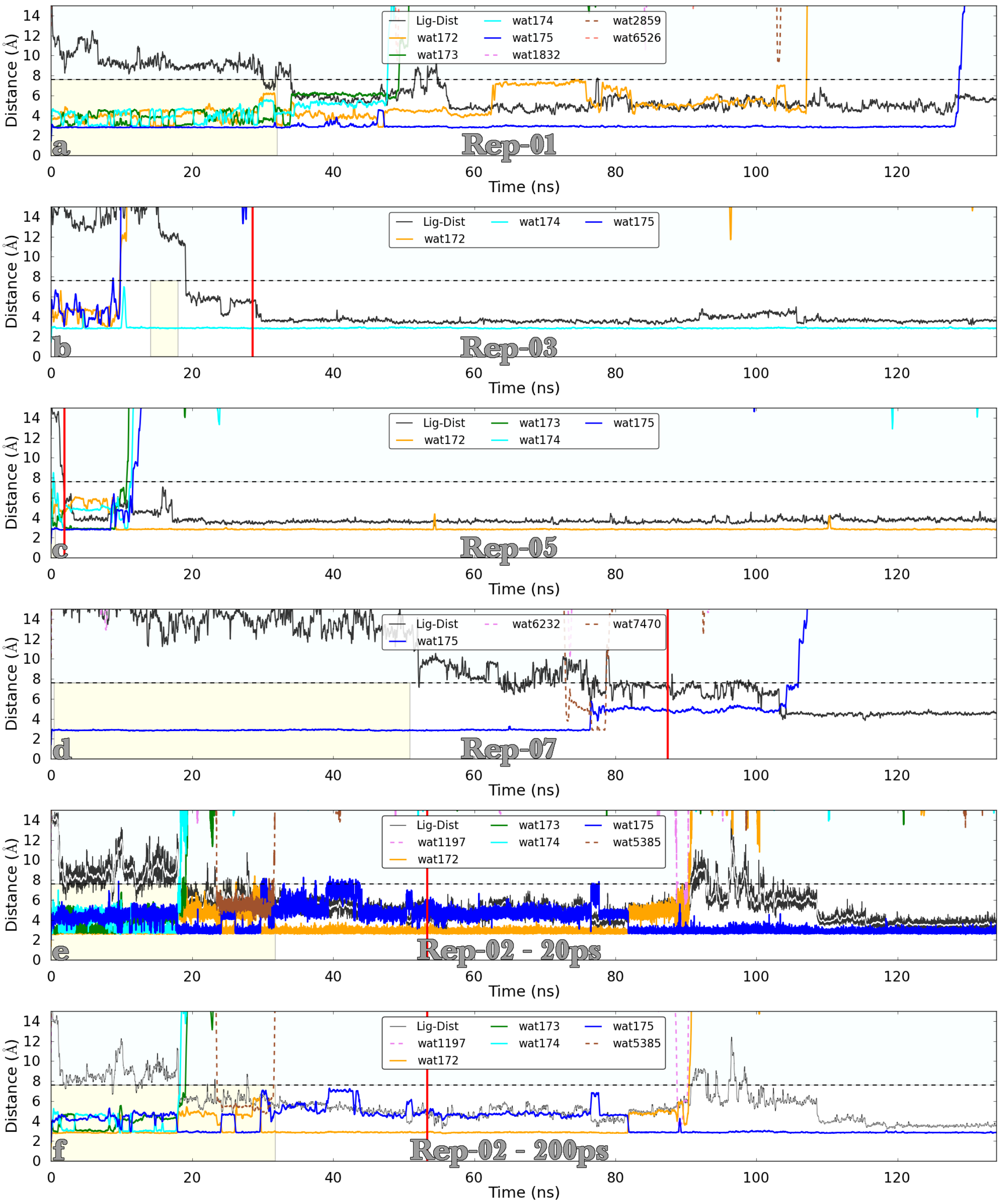


Fig.5.11. Running averages of distances for 4 simulation repeats (01, 02, 03, 05 & 07). The distance of the ligand COM to TYR120@OH in binding site is plotted as a grey dotted line. Bound waters modelled at the start of the simulation are solid coloured lines, whilst bulk waters are marked by dashed lines and have residue numbers higher than :175. Solid vertical red line is first adoption of a vertical pose similar to that seen in the crystal structure. As the ligand retains considerable residual motion, there is no necessity for stable maintenance of this pose. Shaded yellow area indicates the span of time that the ligand spends during internalisation i.e. the time spent in the threshold of being in or out of the cavity. Distances plotted in panels (**a-d**) are depicted as moving averages calculated with a 200 ps window. This has a smoothing effect and panels (**e-f**) demonstrate the difference when using a smaller 20 ps window for repeat-02. The dashed horizontal line at 8 Å represents the boundary between bulk water and the interior of the calyx

5.3.4.5. T2FE uncovers bound water response to ligand internalisation

To continue the discussion, the response of bound waters to 3c9 internalisation is analysed below. As these simulations were shorter and more amenable to visual inspection, the slightly more sophisticated method described in §5.2.3 was used to monitor water dynamics in preference to the first.

Four bound waters (molecules 172 to 175) were initially modelled within the calyx, but their numbers fluctuate throughout the time course of the simulation due to stochastic exchange with bulk. In the T2FE simulations, bound waters are displaced at disparate points during the internalisation process (**Fig.5.11**). Solvent expulsion does not occur at clearly defined points for the following reasons. The time period for complete ligand internalisation is fairly variable (~1 to 51 ns) and there is a small possibility that waters can escape while the ligand is midway through the lip of the calyx. Moreover, just prior to internalisation, gateway residues are likely to open and close with greater frequency (than observed during equilibrium simulations) because polar interactions between anchor residues are disrupted by the body of the ligand. At this stage the ligand is often still in the process of exploring the mouth of the protein and does not fully block the passageway into the cavity. This phenomenon accounts for the expulsions of waters prior to ligand internalisation (**Fig.5.11.d**). Also note that during internalisation bulk waters sometimes make brief forays into the cavity but do not stay for long (**Fig.5.11.d,f**).

Subsequent to entry, various electrostatic forces drive the ligand to make H-bond interactions with residues at the rear of the cavity and during this transitional period, it is likely that increased ligand dynamics facilitates the further expulsion of water molecules. Within 4 to 6 ns after ligand internalisation in repeat-05 (**video.5.10**), increased ligand dynamics displaces all bound waters apart from Wat172 which occupies a very stable position referred to as the "pole position". The increase in ligand instability is reflected as a change in ligand distance in from ~4 to 5 Å in **Fig.5.11.c**. In this time, **video.5.11** shows water molecules "walking" out of the occluded cavity with the assistance of hydrophilic protein residues such as TYR84 and the backbone oxygen of PHE38. As the ligand has just entered, gateway residues are still parted. And even though bound waters can H-bond to each another, the polar environment offered by bulk is preferable to that of the hydrophobic cavity. As indicated by **Fig.5.11,** the time taken for water expulsion is somewhat variable and the precise number of waters expelled across the different repeats is stochastic. As thermodynamic analyses typically evaluate the difference between endpoint states, the definition of the exact moment the ligand is considered fully bound is an open question. Is it immediately upon entry or x ns later? A possible answer is that the ligand should be considered fully bound when the protein begins to sample conformations corresponding to the equilibrium state relevant to that particular holo complex. Nonetheless, if the simulations were extended, it is probable that bulk waters would re-enter the cavity and visual inspection of the equilibration simulations demonstrated that over longer timescales, this is indeed the case.

Having described the response of bound water to ligand internalisation, the next section begins to quantify the relative order of more mobile waters compared to the pole position water.

5.3.4.6. All waters are not equal under thermodynamic laws

Dunitz (1994) theorised that the entropic cost of near total immobilisation of a single water molecule was ~9 kJ/mol by contrasting the difference between the entropy of ice/hydrated inorganic salts and liquid water [309]. Later, Li and Lazaridis (2003) quantified the entropic contribution of a single (long residency) water molecule within HIV-PR using inhomogeneous solvation theory and MD [330]. This particular water is very ordered and makes key H-bond interactions between the bound inhibitor and the protein's two flap domains. They obtained a value of ~12 kJ/mol and suggested that the value was higher than Dunitz's theoretical limit because this water was extremely ordered. Huggins (2015) also suggests that entropies higher than the theoretical limit are possible in cavities containing charged residues [450].

The relative (in)stability of bound waters to one another and bulk solvent is of particular interest and **Fig.5.11** indicates that the water molecule closest to TYR120 occupies a "pole position" which is more ordered than other waters within the cavity. This is due to the greater availability of potential H-bond partners located in the zone ~3 Å around TYR120's hydroxyl group (§5.3.4.4). In contrast, waters positioned further away experience greater positional fluctuations. *This suggests that the energetic cost of ejecting a water molecule from the calyx is not the same for all waters.* The favourable electrostatic interactions offered by the primary hydrophilic site promotes competition amongst bound waters which vie for the pole position. And at certain junctures, positions are swapped. For example, observed how orange and blue solid lines exchange positions in **Fig.5.11.a,f**. In MUP, the ejection of the pole position water to bulk is expected to yield the most favourable entropic contribution. However, displacement of the other waters should also be accompanied by smaller entropy gains. Quantification of these graded contributions would be a useful addition to the literature and the preliminary analysis is presented below.

The positional entropy contributions of bound and free TP4 water were assessed with the 3Dh method described in Chapter 4.0. Whilst free TP4 water was well sampled at 1.5 μs, bound waters were suboptimally sampled at 40 ns per water molecule. This is because the number of bound waters fluctuates and it is difficult to obtain long T2FE simulation segments with a constant number of waters. Consequently, at a later date, the equilibrium simulations will be clustered to yield trajectories with different numbers of bound waters. The results presented in **Table.5.7** are a rough approximation of positional entropy differences ($T\Delta S^\circ_{\text{Wat}\cdot\text{Po}}$), and are useful in that they can be compared relative to one another. The differences are calculated by subtracting bound from free, and report on the entropic benefit of expelling bound water. As the hydrophobic nature of the pocket precludes the free movement of water molecules throughout the entirety of the cavity, the resulting volume occupied by bound waters is very small [90]. Consequently, $T\Delta S^\circ_{\text{Wat}\cdot\text{Po}}$ provides a large contribution to the total entropic cost of expelling a water molecule ($T\Delta S^\circ_{\text{Wat}}$) because it is derived from the difference between bound and free standard state volumes (**eqn.4.12**).

	Water Positional Entropy (kJ/mol)							
	Bound Water Entropies			**Free**		**$T\Delta S^\circ_{\text{Po}}$**		
Water	**Rep 01**	**Rep 03**	**Rep 07**	-		**Rep 01**	**Rep 03**	**Rep 07**
172	17.0	-	-	26.5		9.5	-	-
173	17.0	-	-	26.5		9.5	-	-
174	17.0	10.5	-	26.5		9.4	16.0	-
175	12.23	-	13.3	26.5		14.3	-	13.2
All	17.83	-	-	26.5		8.7	-	-

Table.5.7. Water molecule's positional entropies calculated with 60 bins along each axis. The entropy difference was obtained by subtracting bound from free as water molecules are being expelled from the cavity to bulk.

The following results are described on a per-repeat basis to highlight how $T\Delta S^\circ_{\text{Wat}\cdot\text{Po}}$ is dependent on the *number and identity* of entities within the calyx.

1. Repeat 01: The first 40 ns of the bound simulations were taken for analysis because there are four relatively stable waters within the calyx. All waters apart from Wat175 have similar $T\Delta S^\circ_{\text{Wat}\cdot\text{Po}}$ values of ~9.5 kJ/mol (**Table.5.7**). As Wat175 occupies the pole position, it is the most ordered, and its removal would ostensibly provide a large favourable contribution of 14.3 kJ/mol to $T\Delta S^\circ_{\text{Glo}}$. The calculated values are much greater than the theoretical limit proposed by Dunitz because waters in the bound state are not well sampled. As water molecules are indistinguishable from one another, the coordinates from all four bound waters were concatenated to create a 160 ns trajectory. As the underlying distributions are better sampled, $T\Delta S^\circ_{\text{Wat}\cdot\text{Po}}$ is reduced to 8.68 kJ/mol; a value within the theoretical limit. Further reductions in $T\Delta S^\circ_{\text{Wat}\cdot\text{Po}}$ are expected on increasing sampling to match that obtained for the free state (§4.3.6). Though the calculated value for $T\Delta S^\circ_{\text{Wat}}$ will become larger upon taking the orientational contributions of waters into account, improved sampling and water-water correlations [450,451] are expected to significantly reduce the overall magnitude of $T\Delta S^\circ_{\text{Wat}}$ so that values fall within the theoretical limit. The extent to which these reductions would apply to the pole position water is currently unknown.

2. Repeat-07: As the number and identity of bound entities were believed to affect TΔS°$_{\text{Wat•Po}}$, the first 40 ns of this repeat was taken to evaluate the single pole position water. The calculated value of 13.3 kJ/mol is lower than that obtained from the water occupying the pole position in repeat-01 (14.3 kJ/mol) because it is likely that the other less ordered waters in repeat-01 acted as a stabilising force. If the bound trajectory was better sampled it is expected that protein conformational changes will perturb the position of the pole position water to a greater extent than observed here.

3. Repeat-03: A 40 ns trajectory segment (between 40 ns to 80 ns) was extracted to assess the effect of bound ligand in (de)stabilising the water at the pole position. As TΔS°$_{\text{Wat}}$ for Wat174 is 16.0 kJ/mol, it can be surmised that the ligand's hydroxyl group stabilises the pole position water to a greater extent than the presence of other bound waters in repeat-01. This is probably because the larger ligand is relatively less mobile than bound waters which are small and have greater facility to shift their positions and orientations.

The analysis demonstrates that the majority of bound waters are expelled upon internalisation of 3c9. After ensuring adequate sampling and including orientational contributions, it is expected that the displacement of these waters will make a significant contribution to TΔS°$_{\text{Glo}}$. Water bound at the pole position is anticipated to make the largest contribution and there is a possibility that the magnitude of this value might be larger than the theoretical limit. A weakness of *in silico* studies is the computational expense associated with accruing a sufficient amount of repeats. Thus, finding the exact number of waters displaced upon binding different ligands is a formidable task. Wet experimental techniques such as ultrasonic densimetry are better suited to obtain this information because these methods report the average of a massive number of molecules in solution. However, MD analysis does provide an atomistic picture that aids the understanding of systems and the development of better inhibitors.

5.4.0. Conclusion

5.4.1. T2FE methodology

The data is only semi-quantitative and requires further development, but the insights offered are valuable as this initial foray explores future directions for research. Ideally it would have been preferable to generate many simulation repeats in order to develop MSMs that would provide data on the free energy landscape and kinetics associated with the internalisation process. However, access to the appropriate hardware (GPUs and sufficient hard disk space) was limited, sampling was likewise limited. Nonetheless, the internalisation protocol is of interest and the insights offered are valuable. As it stands, it can be used for the following tasks amongst others:

1. If complexed protein-ligand X-ray or NMR spectroscopy structures are unavailable for MD, bound ligand poses can be discovered with a minimum of bias. The ability to execute this has been tested by binding a different set of ligands to rat odorant binding protein 3 (ROBP3) [452]. The method yielded 32 independent internalisation events (data not shown).

2. Mechanistic details about the binding process can be uncovered and potentials of mean force (PMF) obtained by extracting key structures and utilising them in pathway techniques such as umbrella sampling. Furthermore, viable reaction coordinates can be identified

3. The desolvation processes accompanying ligand binding can be characterised and quantified.

4. Key H-bonding interactions along the pathway can be identified.

Possible ideas for future research include the mutation of arginine residues such as ARG60 and ARG39 to ablate binding. Interestingly other research on other lipocalins has suggested that TYR84 on the L5 loop is important for binding [392]. The results obtained here suggest that mutation to a phenylalanine or alanine might actually destabilise the closed conformations of gateway residues and thus improve binding. Indeed, mutation of the analogous tyrosine in rat odorant binding protein 3 failed to greatly perturb binding [452].

The ability to run constant pH simulations on GPUs allows testing of the possible pH dependency of MUP, and the implementation of grid inhomogeneous solvation theory (GIST) into AMBER enables easier quantification of site-specific thermodynamic values associated with water occupancy.

5.4.2. Ligand & cavity desolvation contributions to the global entropy

In light of the discussion so far, it would be useful to reconsider the global entropies of binding of ligands to MUP with the following hypothetical argument. In §5.3.4.2 the evidence indicated that the majority of the ligand desolvation entropy calculated from vapour-solvent partition equilibrium experiments would be available to the system. The energetic contribution from this source is quite large and values for n-alkanols reported in Malham et al. (2005) are plotted in **Fig.5.12** [36].

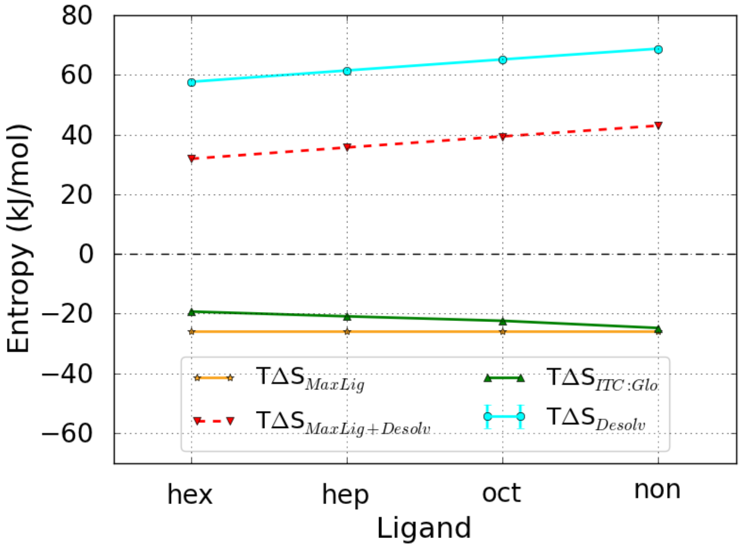
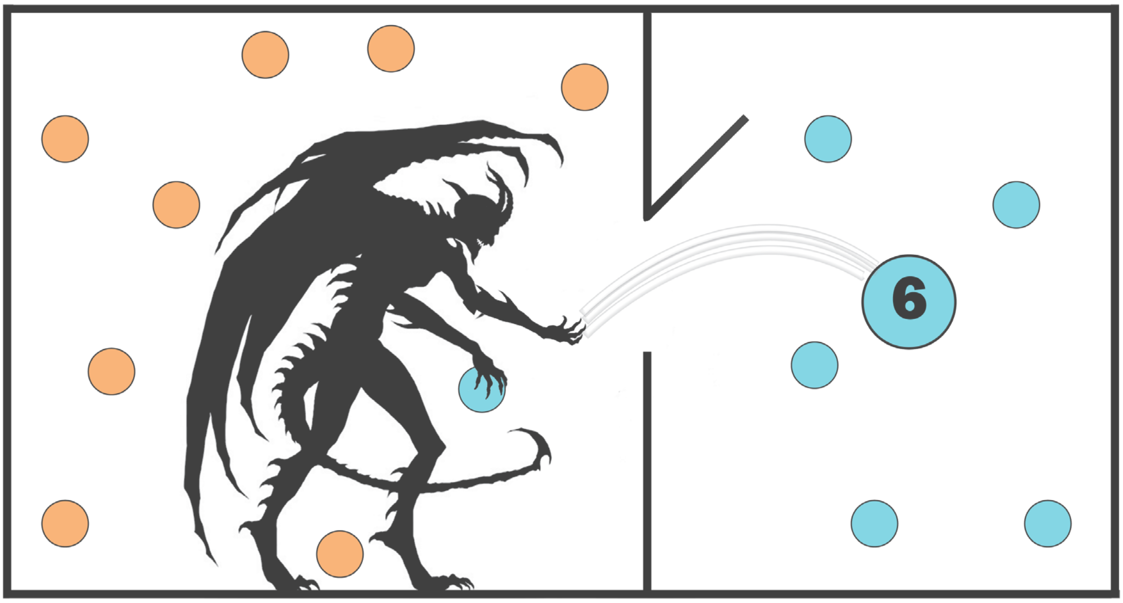


Fig.5.12. Hypothetical desolvation argument. See main text for details.

The initial hypothesis put forward in that paper and others that dealt with the binding of IPMP and IBMP to MUP proposed that the atypical decrease in $T\Delta S°_{Glo}$ across the series was primarily derived from the loss of ligand DOF [36,89,163,453,454]. Whilst an unfavourable entropy contribution from the protein was observed in the case of IBMP [89], this was deemed to be invariant in the case of binding different n-alkanols [36]. Furthermore, data based on computational and crystallographic analysis suggested that a similar number of bound waters were displaced on ligand binding [36]. These are reasonable assumptions when taken in the context of calculating the $T\Delta\Delta S°_{Glo}$ between different binding interactions to the same protein [36]. However, on considering the global data, it is difficult to see how the observed ITC values can be recreated by counterbalancing the loss of ligand DOF with such a large desolvation entropy. Both Roy et al. (2010) and Irudayam et al. (2009) estimate that on binding, the loss of IBMP DOF incurs an entropic penalty of only around -22 kJ/mol [91,164]. Even if the actual entropic cost is set at the higher value of 25.8 kJ/mol as proposed by Turnbull et al. (2004), the ligand desolvation entropy cannot be counterbalanced. Indeed, the gap between experimental $T\Delta S°_{Glo}$ values and calculated values ($T\Delta S°_{MaxLig}$ + $T\Delta S_{Desolv}$) ranges from +51.2 to +67.8 kJ/mol (**Fig.5.12**) [313]. Furthermore, this gap can only be widened if cavity waters are displaced on ligand binding. This is a very real possibility as analysis of the T2FE simulations in §5.3.4.5-6 indicated that bound waters do become displaced on binding 3c9, and the associated entropic benefits are likely to be significant. Logically, the only remaining force able to compensate this surplus is a large unfavourable protein contribution. Thus, this is the topic of the next chapter.



Maxwell's Daemon

Chapter 6.0: The protein response to ligand binding

6.1.0. “Dr. Livingstone, I presume?”

In the search for the molecular rationale underpinning the global ITC trends that describe the binding of ligand panels to MUP, both ligand and desolvation contributions have been duly considered (**Fig.3.1** & Chapters 3.0-5.0). The results thus far indicate that when summed, these contributions do not match $T\Delta S^{\circ}_{Glo}$ measurements. In fact, the large ligand desolvation contribution results in the calculated entropy of binding overshooting global ITC measurements by ~ +51.2 to +67.8 kJ/mol (**Fig.5.12**). As the only component left to be assessed is the protein, it is logical that this is the missing piece in the puzzle. The restriction of the many DOF present in the protein could conceivably yield a large unfavourable entropy contribution that could counterbalance the favourable desolvation contribution. Thus, this is the focus of this chapter.

Intuitively, protein and ligand DOF are expected to become restrained upon binding and indeed, studies on the binding of Biotin to streptavadin suggest that cooperativity of interactions can result in structural tightening that stabilises the complexed form. Hence, in these types of bimolecular interactions, the enthalpy is favoured at the expense of the entropy [33,455]. However, this need not always be the case, and an example of structural relaxation was reported on the binding of p53 to the chaperone protein HSP90 [456]. Typically, systems that possess favourable entropies of binding like ubiquitin are associated with the hydrophobic effect and the release of ordered solvent molecules from the binding interface (§1.1.6) [457–460]. Yet there are several entropically dominated systems whose binding signature could be dominated by the protein contribution instead of solvent reorganisation [461]. Two examples include the binding of ions such as Zn^{2+} to conantokin-G/ conantokin-T [462] and Ca^{2+} to phospholipase D [461,463]. Another study by Tzeng and Kaodimos (2012) investigated the binding of a set of catabolite activator protein mutants to the same DNA sequence [464]. As the location of the mutations was distal to the binding site, there was minimal perturbation of the binding interface. Hence, variations in the desolvation, translational and rotational entropies of binding were deemed to be negligible and the precise protein conformational entropy

($T\Delta S_{Prot}$) could be more accurately ascertained. Within a set of mutants, $T\Delta S°_{Glo}$ had a range of ~62 kJ/mol, whilst $T\Delta S_{Prot}$ spanned a much larger range of ~146 kJ/mol. As the solvation thermodynamics were found to be invariant between mutants, it was proposed that the distal mutations could modulate the large scale redistribution of the protein's conformational entropy. This meant that fast internal motions within the protein played a prominent role in aiding or hindering binding [464].

The thermodynamic role of the protein in MUP is far from clear because the two main papers that studied this important component provided conflicting results. The first by Zidek et al. (1999) evaluated the binding of SBT, and reported an *increase* in backbone dynamics in 68 residues, whilst the remaining 63 measured values did not display significant changes [100]. In contrast, the second study by Bingham et al. (2004) found that the backbone exhibited an overall *decrease* in mobility on binding IBMP [89]. The same study reported that though methyl containing sidechain within the pocket decreased in mobility in the presence of ligand, residues distal to the binding site displayed a compensating increase in mobility. This phenomenon was termed entropy-entropy compensation and it highlights how dynamics can be redistributed throughout the protein's amino acid network. However, the lack of corroboration between the different experimental studies on MUP mean that the precise contribution $T\Delta S_{Prot}$ makes to $T\Delta S°_{Glo}$ is currently debatable.

6.1.1. Methods that investigate the protein entropy

The problem of obtaining experimental data that can be used to corroborate MD results is pronounced in the case of the thermodynamic decompositions employed in this work because it is very difficult to find any experimental corollary with good signal to noise ratios. NMR spectroscopy and MD have enjoyed a close relationship as data from the former is often used to parameterise and improve the latter. The calculation of Lipari-Szabo generalised order parameters (S^2) are an example of an experimental method that can obtain entropies on a per-bond basis [465,466]. However, the data obtainable from NMR spectroscopy is bound by technical limitations that are not applicable to MD (§1.3.4). For instance, NMR spectroscopy is usually limited to analysis of amide and methyl bond vectors, whilst MD simulations have no such limitations. Depending on the system, it can also be difficult to obtain comprehensive sets of clearly resolved peaks and thus, data on bond vector entropies are often incomplete. Moreover, MD is not limited by the rotational tumbling of the protein and can explore timescales of motion in excess of what is achievable with conventional ^1H-^{15}N dipole-dipole T_1 & T_2 relaxation experiments which generally take place in the picosecond timescale. Maragakis et al. (2008) reported that the common issue of systematic discrepancies between S^2 values obtained from the two techniques could be better addressed by including external DOF from the MD analysis as it better captures the dependence of relaxation rates on molecular tumbling [467]. Other techniques such as Car-Purcell-Meiboom-Gill (CPMG) relaxation experiments [468,469] can obtain data on longer millisecond timescales but these experiments tend to be intrinsically insensitive and again have limitations on what bond vectors can be analysed. Thus, these methods also typically report on a small subset of nuclei.

NMR spectroscopy techniques offer the best possible avenue through which site-specific changes in the protein can be experimentally measured. However, as discussed above, these techniques are subject to many limitations which mean that they *do not* usually give an accurate estimate of the true magnitude of $T\Delta S_{Prot}$ e.g. Only a subset of bond vectors are captured and typical methodologies assume that no correlations between bonds exist. *The total conformational entropy of the protein can only be accurately quantified by assessing every DOF in the macromolecule. This is because the summation of per-bond entropy contributions has the propensity to accumulate or cancel in unpredictable ways as new per-bond entropy data is added to the data pool used to calculate $T\Delta S_{Prot}$.* In contrast, MD has the ability to overcome these technical limitations but is hamstrung by difficulties in accruing ergodically sampled trajectories. Additionally, theoretical calculations have been historically limited by compute power. For these reasons, researchers typically only assess subsets of the available protein DOF using techniques such as the QHA and histogramming methods [470]. For example, limiting analysis to bonds within the protein backbone saves considerable computational resources. The likely presence of significant correlations amongst the many bond vectors have resulted in the development of a variety of computational techniques that assess 2nd or higher order correlations (§3.1.2). One of the most noteworthy was presented by Fenley et al. (2012) [471]. They calculated the conformational entropy of all 889 bond torsions within BPT1, and used the maximum information spanning tree (MIST) algorithm to account for higher order correlations [472,473] (§6.1.2). As discussed in §3.1.2, when calculating higher order correlations, a great many data points are required to counter the rarification of data points in higher dimensional volumes. Thus, the authors work was greatly facilitated by the small size of the protein (58 residues) and the availability of 1 ms worth of simulation data. Some of the pertinent findings of this study are further discussed in the next section.

6.1.2. Entropy-enthalpy transduction & key control variables

The strength of almost all wet techniques is that experimental samples contain an enormous numbers of molecules and the consequent averaging provides well converged results. Yet, this is also a weakness in that averaging makes it difficult to separate the different structural and dynamic features exhibited by the main clusters of substates that constitute the commonly reported ground state. For example, the study based on the ultra-long 1 ms simulation of BPT1 discovered that the protein possessed multiple thermodynamic substates; each with distinguishable thermodynamic signatures [471].

Whilst the free energy for each cluster was nearly identical, their component entropies and enthalpies could differ widely. The authors hypothesised that as the free energy differences were so small, even a small perturbation (such as mutation or an incremental ligand structural difference) could result in large shifts in the underlying entropies and enthalpies, and thus change the occupancy of the identified clusters. Moreover, this could transpire with minimal perturbation to the free energy.

The authors examined the role of control variables i.e. key interatomic distances or torsional ranges that could drive the protein into a distinct cluster when computationally modulated by the researcher. The overall thermodynamics of a binding interaction can be conceptually subdivided into an "intrinsic" portion which constitutes *direct* ligand interactions with receptor and a "transduced" portion which encompasses the *long-range* modulation of protein conformation resulting from ligand interactions with key control variables. As local thermodynamic driving forces may be distinct from global forces, the agent responsible for perturbing the system could be viewed as "transducing" a local force into quite a different global thermodynamic signature. Identifying key control variables in the protein could therefore dramatically accelerate the process of rationally designing small molecule inhibitors. Yet, this is a difficult problem to deal with as it necessitates extremely long MD simulations or creative methods to ensure all conceivable conformations are adequately represented.

6.1.3. Objectives

The results presented below seek to assess three principle objectives.

1. The results obtained in Chapter 4.0 suggested that 3Z-olefins undergo increased dynamics compared to n-alkanols (§4.3.7). Therefore, it is of interest to establish whether a corresponding enthalpic difference can be measured from the simulations. The number of inter-atomic contacts is used as a cost-effective proxy for the enthalpy.

2. A modified version of the program order [360] is used to calculate the protein's conformational entropy for n-alkanol and 3Z-olefin complexes. The method underlying the program is based on that proposed by Yang and Kay (1996), and this computational analysis allows a rough estimate of summed $T\Delta S_{Prot}$ contributions derived from 8 different complexes [96,361].

3. The protein entropy analysis necessarily involves the calculation of per-bond entropy differences. Thus, trends in this data will be utilised to identify possible control variables.

6.2.0. Methods

6.2.1. Contacts

The number of contacts was assessed with the “contacts” command in cpptraj. The command works by assessing all contacts within a specified distance between the atoms specified in a search string. Thus, if a single residue is selected, the command only reports the number of intra-residue contacts that fall within this distance. With this in mind, different instances of the command were run on each 1.2 µs apo and holo trajectory where appropriate:

1. All contacts for every protein residues (1 to 157 in one command instance)

2. Ligand only (self contacts)

3. Ligand versus a single residue (x157 discrete command instances)

4. Residue only (self contacts x157 discrete command instances)

Adding or subtracting various combinations of these different datasets allowed the calculation of:

A. PL interfacial contacts: Protein and ligand intermolecular contacts with no protein intramolecular contacts. Requires the holo trajectories only. The average number of contacts per residue was obtained by dividing the sum of all contacts by the number of frames utilised.

B. All PL contacts: Includes inter- and intramolecular protein-ligand contributions. The difference in contacts was obtained by subtracting average number of contacts in the receptor from the holo complex.

C. All PP contacts: Intramolecular protein-protein contributions only. The difference in contacts was obtained by subtracting average number of contacts in the receptor from the bound state.

A cut-off of 7 Å was used for all calculations as the distance should be large enough to remove most residue self interactions to obtain an accurate estimate of the number of interfacial contacts. A tighter cut-off of 3-4 Å would technically be more accurate, but the larger distance preserves the necessary information.

6.2.2. Computational calculation of protein entropy

The conformational entropies of bond vectors were calculated with program order [360]. The method used in the program is described in §4.2.5.3. The protein backbone was RMS fitted to the reference structure used in §4.2.2 prior to analysis. The key difference between the calculation of the ligand orientational entropy and the protein conformational entropy is in the RMS fitting. When calculating the ligand orientational entropy, ligand dynamics is not limited by explicitly fitting any of its coordinates. Thus, it is allowed to move freely relative to the frame of the fitted protein i.e. it possesses both T&R motion. A subsequent processing step moves each ligand bond vector to the origin so that the bond rotates around this central point. This removes translational motion while preserving rotational motion. In contrast, the motion of protein bond vectors is constrained to the frame of the protein which has had both gross translational and rotational motions removed.

6.2.3. Protein H-bond occupancies

Hydrogen bond (H-bond) analyses on the MD generated ensemble of structures were accomplished using the ptraj module from AMBER, and custom python scripts. Analysis was performed on all 1.2 x 10^6 frames for all ligands from both panels and every solute-solute H-bond interaction was recorded. Note that percentage H-bond occupancies are calculated for the apo and holo simulations separately. Subsequent to this, the apo occupancy for a given H-bond was subtracted from holo occupancy. Thus, positive values indicate that there is a large increase in holo H-bond occupancies compared to that of the apo state. In a large number of instances, the difference in H-bond occupancy between apo and holo simulations is very small i.e. (-5 to +5%). These were discounted from the analysis and the importance of H-bonds were categorised into the following occupancy ranges.

1. Stabilising strong: Greater than 30.01%

2. Stabilising medium: Between +15.01% and +30.0%

3. Stabilising weak: Between +5.0% and +15.0%

4. Destabilising strong: Less than -30.01%

5. Destabilising medium: Between -15.01% and -30.0%

6. Destabilising weak: Between -5.0% and -15.0%

Note that it is the holo complex that is stabilised or destabilised with respect to the apo receptor.

6.2.4. Principle component analysis

Videos of largest linear protein displacements were obtained using principle component analysis as described in §3.2.6. All 1.2 x 10^6 frames were used. Only protein backbone atoms were evaluated.

6.3.0. Results & Discussion

6.3.1. Protein-ligand contacts as an enthalpic proxy

Malham et al. (2005) proposed that the favourable enthalpy of binding was the result of increased protein-ligand interfacial contacts [36]. In order to assess this, the number of contacts made within a 7 Å cut-off was assessed as a proxy for the enthalpy. The number of contacts is high because of the relatively large cut-off. This was needed to remove the intramolecular contacts of larger protein residues and recover the interfacial contribution (§6.2.1). Though such an analysis does not provide energetic values, the simplicity of the analysis avoids the difficulties associated with more computationally expensive techniques such as MM(P/G)BSA. The average number of contacts is subdivided into the three categories depicted in **Fig.6.1**.

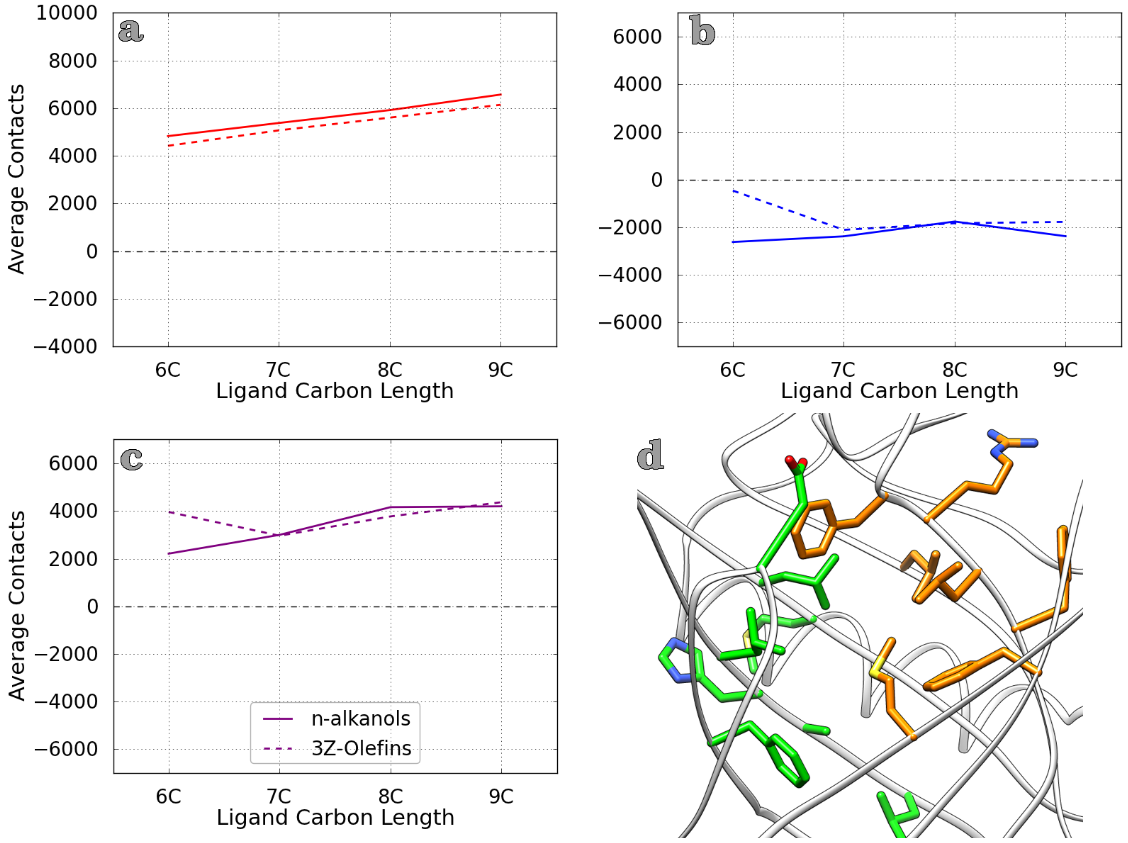


Fig.6.1. Average number of contacts for n-alkanol and 3Z-olefin ligand panels. **(a)** Protein-ligand intermolecular contacts only. No intramolecular contacts. Represents pure PL interfacial contacts **(b)** Protein-protein intramolecular contacts without any ligand contribution. **(c)** All protein-ligand and protein-protein contacts. Obtained by subtracting the receptor from free **(d)** Image of MUP with residues colour coded to show which residues 3Z-olefins (green) and n-alkanols (orange) preferentially contact. Per-residue data obtained from **Fig.6.2**.

1. The average number of protein-ligand interfacial contacts for various complexes (**Fig.6.1.a**). Intramolecular contacts have been removed, so this encompasses protein-ligand intermolecular contacts only. The number of contacts increases linearly as ligand length increases. This suggests that enthalpically favourable interactions occurring at the binding interface contribute linearly to ΔH_{Glo} trend lines obtained from ITC. On average, 3Z-olefins systematically make fewer contacts with binding site residues compared to their respective n-alkanol analogues. As discussed in §4.3.5, this is because ligands preferentially bind to different H-bond partners and as a result of these interactions, 3Z-olefins "fit" better into the cavity compared to n-alkanols in both ligand pose categories. Thus, differences in the location of these ligands in the cavity are directly responsible for their reduced interfacial contacts.

2. The difference in averaged protein-protein contacts between apo and holo simulations (**Fig.6.1.b**). Values have been obtained by subtracting apo from holo contacts and do not include any protein-ligand contacts. The purpose of this graph is to isolate the protein-protein contribution from the total number of contacts which is inclusive of protein-ligand contacts (**Fig.6.1.c**). The shape and slope of the trend lines highlight convergence issues present in the 1.2 μs length simulations. Firstly, the end of the n-alkanol plot forms a "hockey stick" shape because oct and non have a similar number of contacts. The same disruption to linearity was also observed in summed n-alkanol positional and orientational entropies, and this was demonstrated to be indicative of imperfect sampling. As demonstrated in §4.3.9.2, additional sampling for the oct data point assisted the recovery of a linear trend. Secondly, the data point for 3c6, the six carbon 3Z-olefin is out of line with the rest of the unsaturated ligands and is clearly too high. Again, this is a sampling issue which is likely to be the result of the protein preferentially visiting conformations that are unusually tight/compact. The precise reason for this is currently unknown, but as detailed in the following sections, trends obtained from congeneric ligand panels greatly assist the interpretation of imperfect data.

Ignoring the value for 3c6, the magnitude of values plotted in **Fig.6.1.b** ranges from -2374 to -1758. The negative sign indicates that when protein-ligand interactions are neglected, the apo receptor has a greater number of protein-protein contacts than the holo complex. The pocket is ~14 Å across at its widest point, and an increase in apo protein-protein contacts is likely because binding site residues can come closer to one another in the absence of ligand. Conversely, in the holo complex, the same residues are further apart due to the presence of the ligand. Despite the sampling issues, extrapolating the trend lines of both panels indicates that there is an increase in holo protein-protein contacts with increasing ligand length because the apo receptor makes an identical contribution to each data point. However, it is unclear whether there is a systematic differential between comparable n-alkanol and 3Z-olefin ligands. With better sampling and a tighter cut-off, this could be determined in the future. *Nonetheless, this result supports the idea that protein structural tightening does occur, and this* *suggests that the favourable enthalpy of binding is not solely derived from protein-ligand interfacial interactions.*

3. The difference in the total averaged contacts between apo and holo simulations (**Fig.6.1.c**). Includes inter- and intramolecular protein-protein and protein-ligand contacts. Values have been obtained by subtracting apo from holo contacts. Protein-protein contacts account for the bulk of all contacts, but sampling issues introduce some noise into the trend line. Once all protein-ligand interactions are accounted for, the values have a positive value as the holo entity possesses more inter- and intramolecular contacts than the apo receptor. Though the slope of the trend lines in **Fig.6.1.b** indicated that protein tightening occurs as ligand size increases, protein-ligand interfacial interactions ostensibly make a bigger contribution because their inclusion make the slopes much steeper. However, this should be taken with caution due to the lack of energetic values and the use of a relatively large cut-off.

The number of contacts were also analysed on a per-residue basis, and this revealed that ligands in the two panels exhibit markedly different preferences for binding site residues (**Fig.6.1.d**). Whilst n-alkanols make more contacts to residues in the omega loop, 3Z-olefins make contacts to residues at the opposite side. This is a consequence of different patterns in direct protein-ligand H-bonding (§4.3.5). A detailed per-residue plot depicting the difference in the number of averaged number of interfacial contacts between the two panels is presented in **Fig.6.2.** Values have been obtained by subtracting 3Z-olefin per-residue contacts from n-alkanols. At either extreme, positive values indicate that the residue makes more contacts with saturated ligands, whilst negative numbers represent more contacts with unsaturated compounds. Note that using a 7 Å cut-off results in some residues outside the binding pocket picking up a small number of spurious protein-ligand contacts e.g. LEU151 and PHE134. Full per-residue graphs for both panels are located in the appendix (**Fig.A3.1.1-2)**.

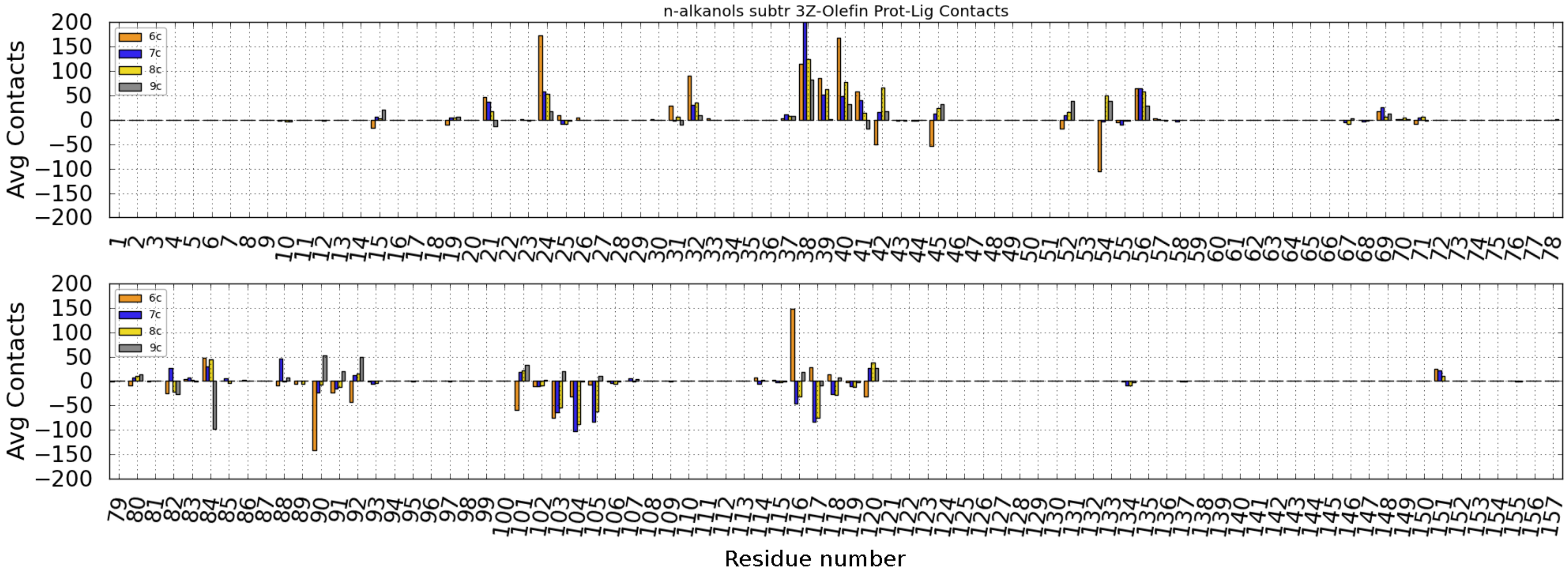


Fig 6.2. Average per-residue protein-ligand interfacial contacts. Obtained by subtracting 3Z-olefin from n-alkanol contacts (Fig.A3.1.1-2). Differences for 6C ligands show out of trend behaviour. As explained in the main text, this is probably because the 3c6 holo complex has not fully converged.

6.3.2. Per-bond entropy analysis & control variables

Only the protein's entropic contribution is left to be assessed in the system. Hence, a rough 1st order estimate for each 1.2 μs holo and apo simulation was obtained by calculating per-bond entropies for the bonds listed below. Per-bond entropy differences ($T\Delta S_{Prot\text{-}pb}$) were calculated by subtracting the value obtained for the receptor from the complex.

1. Amide bonds: Amides in all residues apart from PRO93, PRO124, and GLU1 in the N-terminus were assessed.

2. Aromatic bonds: Conformational entropies were calculated for C_α-C_β and C_β-C_γ bonds for each aromatic residue. As these bonds are expected to be highly correlated, entropy values were averaged because the facility to calculate 2nd order entropies was not available at the time of writing.

3. Methyl bonds: Entropies for methyl bonds were calculated by assessing the motion of the terminal methyl containing bond. In the case of valine residues which contain symmetrical methyl groups, the entropy values were averaged as a rough method to account for correlations.

As discussed in §6.1.1, a large amount of correlations between bonds are expected when calculating $T\Delta S_{Prot}$. However, as all the simulations which have been processed the same way, this 1st order treatment should allow adequate comparison of per-residue trends between the different holo complexes. The work thus far indicates that even after using 1.2 million snapshots, the simulations have not fully converged. However, inaccuracies can be readily identified by deviations in expected trends as the results are derived from multiple independent simulations that examine the binding of ligands with small incremental structural differences. The analysis presented in this section provides strong evidence that there is a differential protein response which is dependent upon the identity of the bound ligand. In addition to detailed per-residue graphs (**Fig.6.4),** a secondary structure graph is also presented (**Fig.6.3**). The graph is created by summing up $T\Delta S_{Prot\text{-}pb}$ values for all bonds within canonical secondary structure elements (**Fig.1.20**). The resulting reduction in both data dimensionality and granularity facilitates the easy identification of which areas of the protein are most affected by ligand binding. The most important trends and conclusions will be briefly discussed.

In both panels the protein generally becomes more restrained as the bound ligand becomes larger (**Fig.6.3**). Furthermore, it is clear that the protein is much more restrained in the presence of n-alkanols compared to 3Z-olefins. The protein's β-clam fold is made of the orthogonal arrangement of strands *b-d* (residues 41-73) with respect

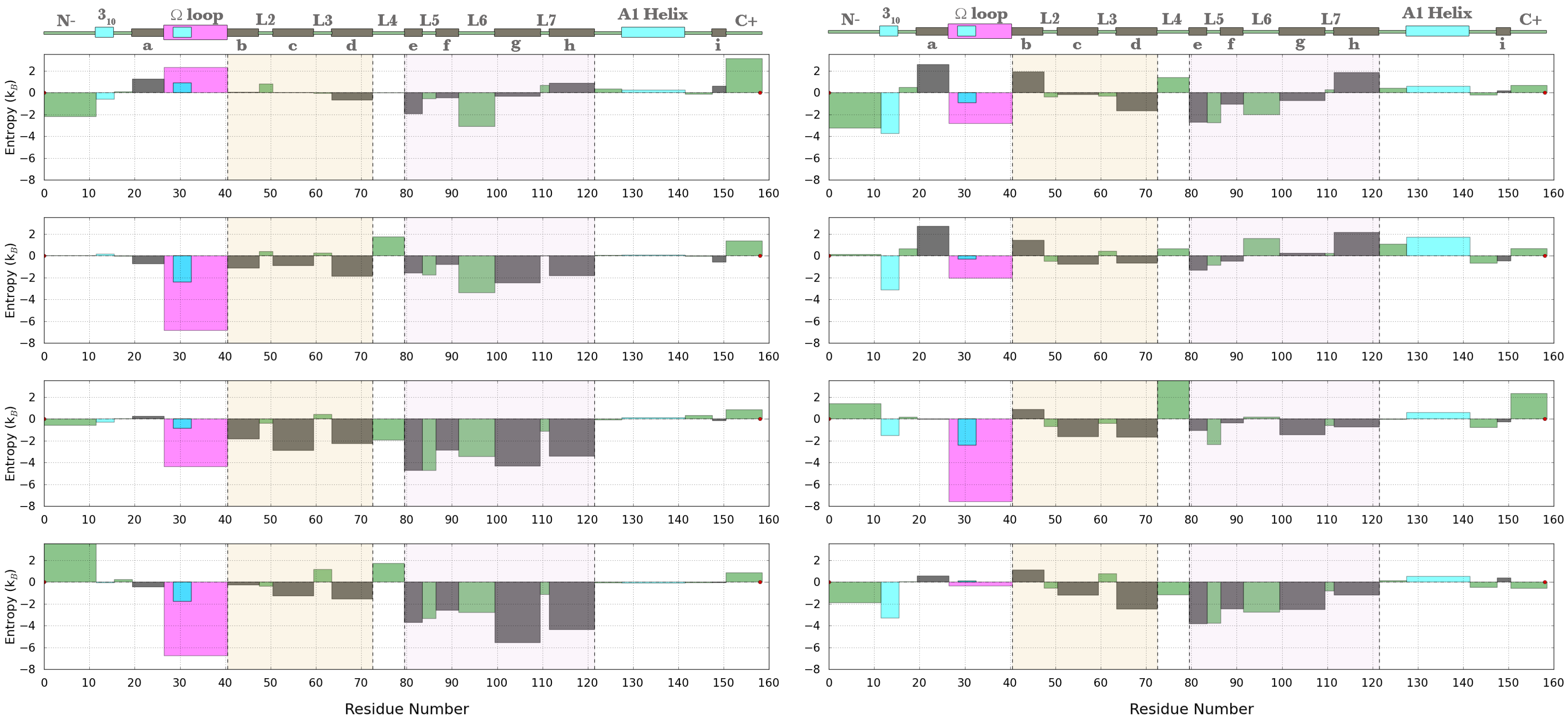


Fig.6.3. Secondary structure bar graph plots the summed 1st order $T\Delta S_{\text{Prot-pb}}$ values for all bond vectors within canonical secondary structure elements. The secondary structure sequence at the top of the panels provides the key for colours used in the bar graph.

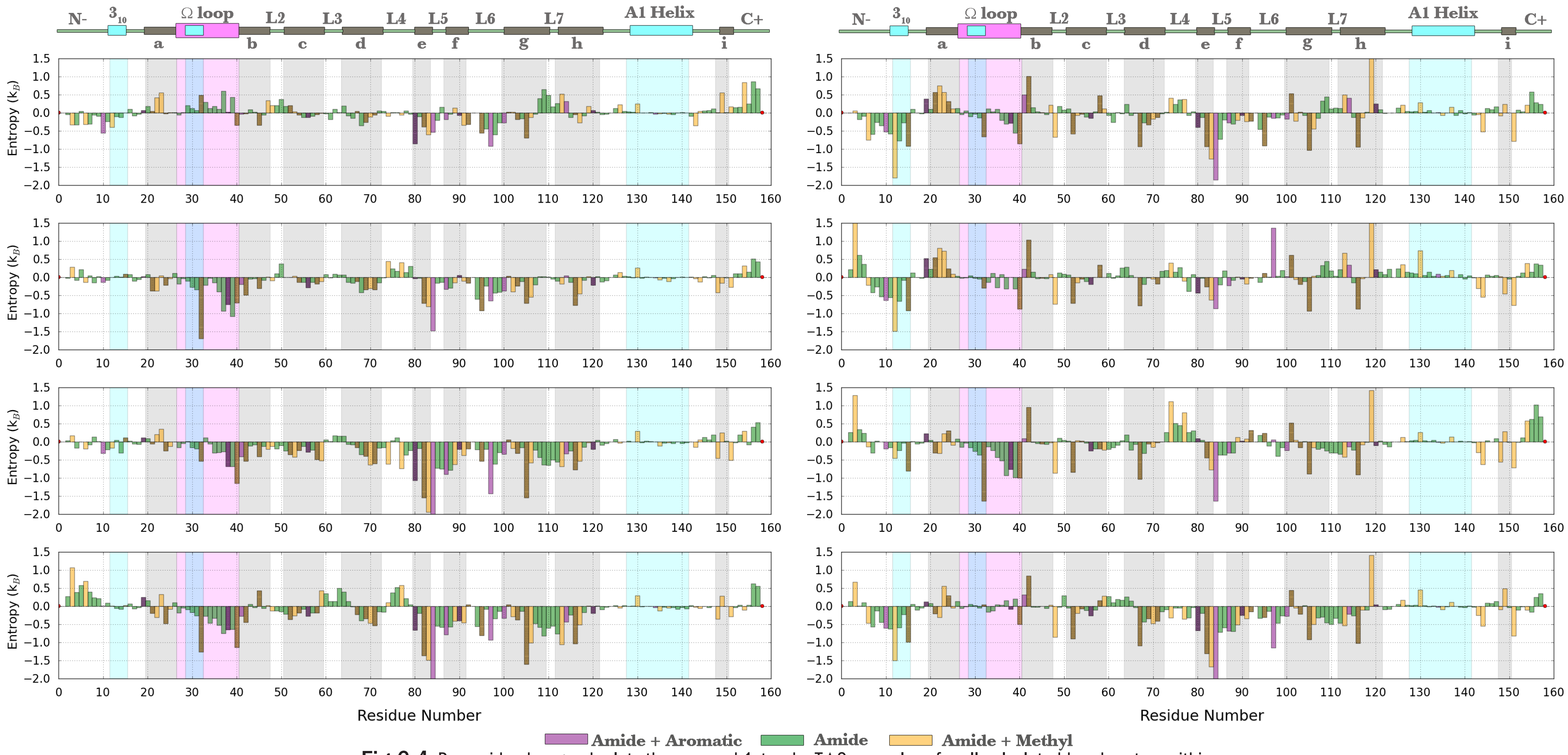


Fig.6.4. Per-residue bar graph plots the summed 1st order $T\Delta S_{Prot\text{-}pb}$ values for all calculated bond vectors within a given residue. Residues containing methyl or aromatic bonds have the $T\Delta S_{Prot\text{-}pb}$ values of these groups added to the amide contribution. Colour legend for residue type is at the bottom of the figure. Also note that residues with sidechains located within the binding site are cross-hatched for easy identification.

to *e-h* (residues 80-122). It is the latter collection of secondary structure elements that are primarily affected by the binding of saturated compounds whilst the former displays a more moderate response. Residues 41-73 are located near the entrance to the protein, and it is possible that this mix of aliphatic and aromatic residues is more densely packed. Thus, this could potentially inhibit large shifts in mobility. More importantly, the contacts analysis indicated that relatively fewer protein-ligand contacts are made to residues 41-73 compared to the larger range of residues within 80-122 (**Fig.A3.1.1-2**).

At this point, the concept of control variables described in §6.1.2 is revisited. Analysis of $T\Delta S_{\text{Prot-pb}}$ trends could shed some light on possible mechanisms by which local ligand interactions potentially drive global protein conformations. The work in chapter 4.0 suggested that a control variable contender was PHE90 (§4.3.4). This residue can flip between "open" and "closed" conformations so as to allow or hinder ligand access to *cal3* respectively. As previously discussed, longer ligands have an increased probability of being displaced into *cal3* and shunting PHE90 aside by virtue of their increased bulk. Thus, this was thought to be a likely control variable as it possesses one of the largest changes in terms of both chemical shift perturbation (data not shown), and per-residue contact analysis (**Fig.A3.1.1-2**). Though inter-panel $T\Delta S_{\text{Prot-pb}}$ differences are present for this residue, the magnitude of the contribution is fairly small. Despite this, it is entirely plausible that the subset of bond vectors used for analysis does not fully capture the entropic contribution from PHE90. Nonetheless, when PHE90 is displaced into the open position, the ligand can adopt vertical poses located further within the heart of the calyx. This has a knock-on effect of increasing the sidechain packing of adjacent residues such as LEU105 and LEU116 which become increasingly restricted as n-alkanol ligand size increases and vertical poses become more stabilised. *Per contra*, in 3Z-olefin complexes, the amount of restriction observed in these residues is relatively constant (~ -1.0 k_B) because unsaturated ligands have greater facility to switch between vertical and horizontal poses (§4.3.7.2). Thus, the underlying distributions of LEU105 and LEU116 bond vectors in various 3Z-olefin complexes possess greater similarity, and intra-panel $T\Delta S_{\text{Prot-pb}}$ values for these residues are relatively invariant.

LEU105 and LEU116 are located in the *g* and *h* strands respectively. Bingham et al. (2004), found that on binding IBMP, residues distal to the binding site in the L6 loop stiffened. A fact substantiated by microsecond simulations [91]. As the vertical ligand pose increases direct protein-ligand contacts, displacement of PHE90, LEU105 and LEU116 could have long-range effects on the dynamics of residues distal to the binding site. Additionally, TYR97 in L6 is well positioned to affect the dynamics of the N-terminal region, and possibly modulate other loops such as the L2 loop by a conformational relay mechanism similar to that described by Bingham et al. (2004) [89]. The entropy differences of residues 92-121 within the secondary structure elements that span L6 --> L7 display significant inter- and intra-panel differences. In n-alkanol complexes there is a general decrease in residue mobility that is correlated with ligand size and this indicates that changes in this area must be affected by greater occupation of vertical poses. Though the dynamics of this region within 3Z-olefin complexes is a lot more variable, the degree of restriction is ameliorated compared to analogous n-alkanol complexes. Indeed, residues such as LEU101 and LEU119 display a *consistent increase* in $T\Delta S_{\text{Prot-pb}}$. Note that LEU119 is located in-between flexible glycines (GLY118 and GLY121) in the *h* strand, and it is possible that that ligand H-bonds to TYR120 assist in the destabilisation of LEU119. Thus, it is unlikely that $T\Delta S_{\text{Prot-pb}}$ gains and losses can purely be described in terms of differences in non-directional interfacial interactions. Strong, directional protein-ligand H-bonds can also modulate the dynamics of residues distal to those in the binding site via the action of coupled protein-protein interactions.

So what is the root cause for greater $T\Delta S_{\text{Prot}}$ losses in n-alkanol complexes compared to 3Z-olefins? As discussed in §4.3.5, n-alkanols preferentially H-bond to residues within the Ω loop whilst 3Z-olefins tend to H-bond to residues such as ALA103 and TYR120 which are located lower down within the calyx (**Fig.4.20**). As all ligands increases in size, an increase in H-bonds occupancies to Ω loop residues such as PHE38 and LEU40 is observed. In n-alkanol complexes, the increase in these interactions becomes more pronounced as ligand size increases, and these directional forces act to stabilise and limit dynamics at the front of the Ω loop. Additional stabilising interactions increase H-bond occupancies between TYR84 on the L5 loop, and ASN37 on the Ω loop at the expense of H-bonds to the upper part of the c strand. Furthermore, additional H-bond mediated tightening is observed near the tops of the *e*, *f* and *g* strands between ASN107, ASN88 and various backbone atoms. As a result of these key interactions, n-alkanol complexes generally see a greater increase in inter-strand H-bonding between residues 80-122 compared to 3Z-olefins (**Fig.6.5**). Thus, this coupled network of stabilising interactions results in greater $T\Delta S_{\text{Prot}}$ losses in n-alkanol complexes versus 3Z-olefins. An issue with this proposition is that the 3c8 holo complex also displays a large amount of restriction in Ω loop residues but the ligand does not H-bond to this secondary structure element for much of the simulation. The precise reason for this discrepancy is currently unknown, but the highly unfavourable $T\Delta S_{\text{Prot-pb}}$ value for ILE32 indicates that a separate confounding protein-protein interaction may be present. The Ω loop is relatively large and difficult to converge. It is capable of complex multi-segmented motions that could be influenced by the auxiliary dynamics of a number of other secondary structure elements. Thus, further sampling is necessary to make an unqualified conclusion on this out-of-trend data.

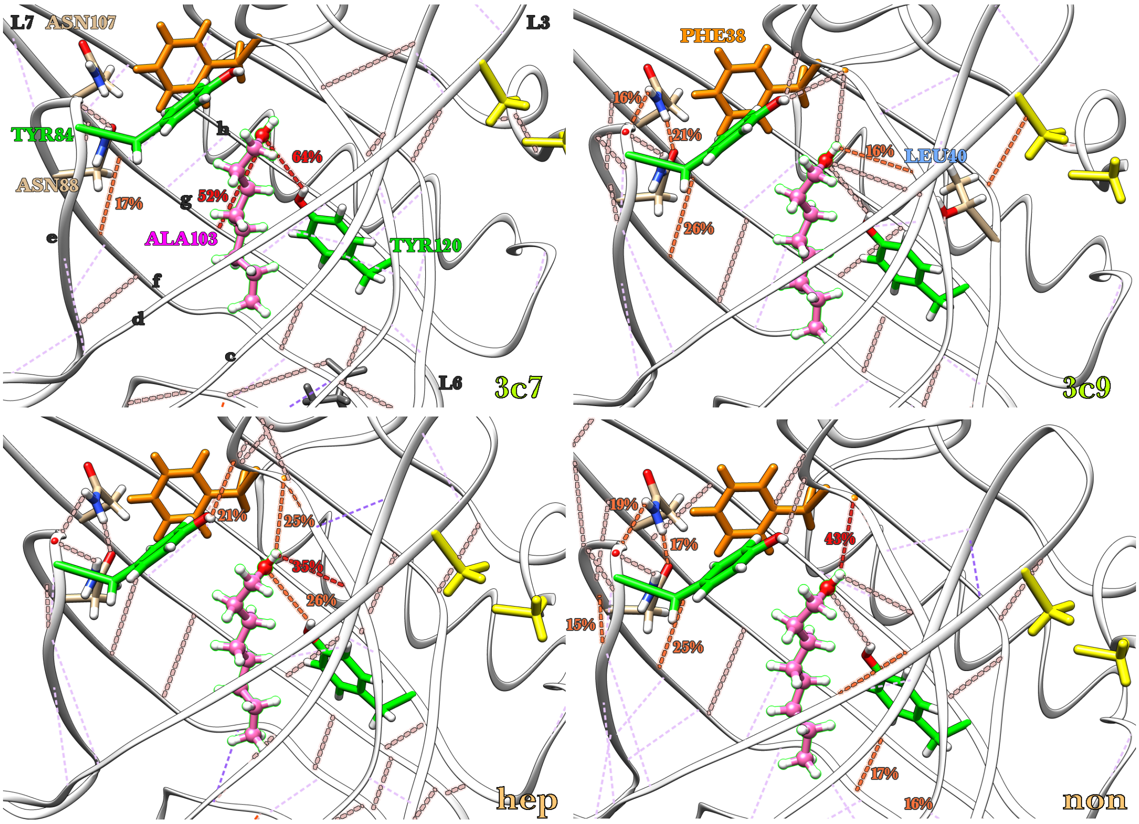


Fig.6.5. Differences in inter-strand H-bonds in various MUP complexes. High occupancy, complex stabilising H-bonds are coloured red, medium - orange, and low - salmon. Complex destabilising H-bonds are unemphasised and are coloured three shades of blue. The occupancy of stabilising H-bonds generally increases with ligand size. Moreover, saturated ligands promote greater positive H-bond occupancies compared to unsaturated ligands. Such a pattern is seen in the H-bonds between *f* and *g* strands. Also note the H-bond locking mechanism between ASN88 and ASN107 act to pull the L5 and L7 loops together whilst H-bonds between TYR84 and the Ω loop act as another immobilisation mechanism.

In contrast, unsaturated ligands do not tend to make H-bonds to the front of the Ω loop and the protein undergoes comparatively less tightening within the areas just discussed. However, the binding of 3Z-olefins elicits intriguing protein responses in two particular areas. Firstly, an increase in the dynamics of the A1 helix is observed. As discussed, LEU119 possesses one of the largest increases in mobility on binding. It is located outside the cavity on the *h* strand and pushes against the A1 helix. This is possibly due to synchronous ligand mediated H-bonds to ALA103 and TYR120 which are respectively located within the *g* and *h* strands. The dynamics of external sidechains are worthy of further investigation but because the current analysis does not include data on all sidechain bond vectors it is difficult to make any firm conclusions. For example, what is the role of arginine residues that link the A1 helix to the Ω loop (chapter 5.0)?

Secondly, a consistent increase in order is observed in the short 3_{10} helix located at the N-terminus (residues 12-15). The precise cause of this is currently unknown, but it is pertinent that a similar amount of tightening is not observed in n-alkanol holo complexes. Zidek et al. (1999) reported that the protein backbone entropy *increases* upon binding the ligand SBT as 68 residues showed significant increases in mobility [100]. Moreover, a single residue, *PHE10 located within the conserved 3_{10}-helix in the N-terminus underwent a significant reduction in mobility*. These observations regarding the N-terminus were supported by MD simulations presented by Macek et al. and it was also noted that the ligand did not remain in the pose seen in crystal structures [101,102]. These simulations also made a link between the ligand's interaction with residues (ALA103 and LEU105) within the calyx and the disruption of H-bonds between residues at the base of *g* and *h* strands near the L6 loop [101,102]. As shall become apparent, this is of interest because bound SBT complexes have been shown to promote a favourable $T\Delta S_{Prot}$ contribution, and 3Z-olefins, (which also H-bond to ALA103), suffer reduced $T\Delta S_{Prot}$ losses compared to n-alkanols. Thus, it is possible that protein-ligand interactions with residues near ALA103 act to increase protein dynamics via a conformational relay mechanism.

At this juncture, two videos are presented to highlight what has been discussed so far. Both videos show the first largest principal component mode for the receptor (**video.6.1**) and all holo complexes (**video.6.2**). You will have to loop the videos in your video player. The program VLC media player is suggested to avoid codec issues. Note the following points:

1. Increasing immobilisation and synchronous coupling of L5 and L7 loops as n-alkanol ligand size is increased. A similar phenomenon, albeit greatly reduced is observed in 3Z-olefins.

2. Difference in N-terminal motion between 3Z-olefin and n-alkanol holo complexes.

3. The Ω loop is capable of various multi-segmented motions. In n-alkanol holo complexes, LEU40 (which is at the front base of Ω loop) gets progressively more restricted as ligand size increases. In all 3Z-olefin complexes apart from 3c8, the amount of restriction in that area is reduced.

On synthesising these disparate pieces of data, it is possible to hypothesise that the protein is capable of modulating its response to ligand binding via dynamic rearrangement of the N-terminal region which lies beneath the base (ILE15, ILE45, LEU52, ILE92, and LEU101) of the binding cavity (**Fig.6.6.a**). Two out of three of short SCRs are localised in this zone (§1.3.1). PHE10 is within the highly conserved SCR1 sequence adjacent to the N-terminal 3_{10}-helix which (in common with other members of the Calycin family) possesses the distinct structural characteristic of forming multiple H-bonds (over a conserved inward-pointing tryptophan) to an arginine or lysine [474].

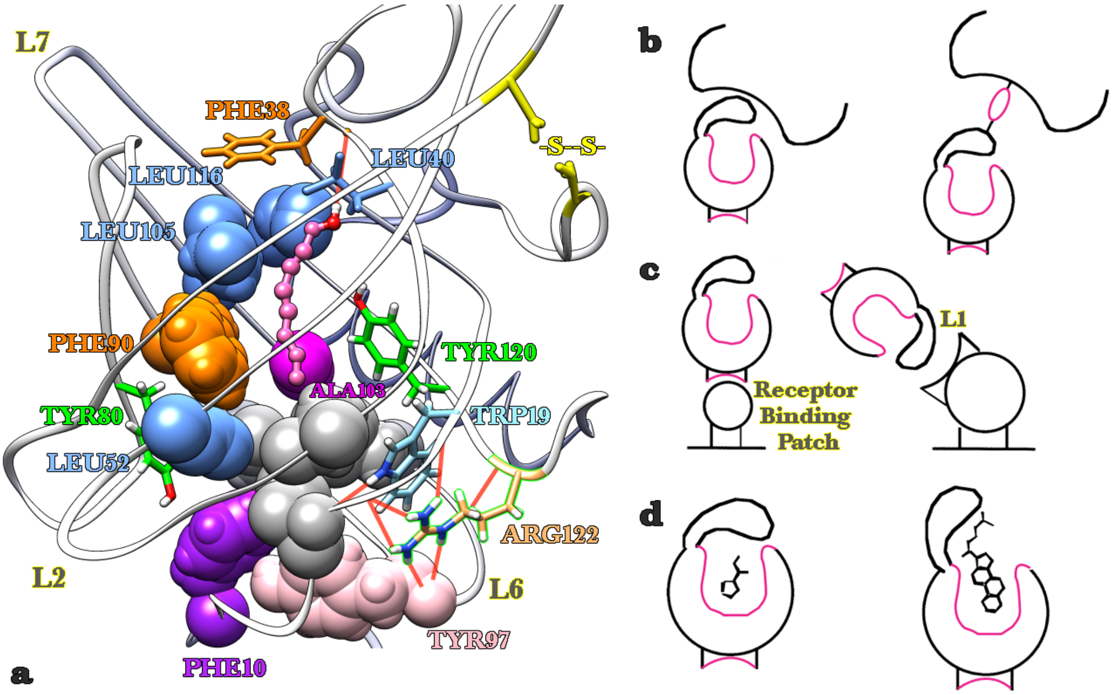


Fig.6.6. **(a)** Image depicting the positions of residues described in the text. Base of the cavity coloured grey. **(b)** Two modes of binding to soluble macromolecules: non-covalent and covalent association. **(c)** Binding to cell membrane receptors via Ω loop or receptor binding patch at N-terminus. **(d)** Binding to small and large ligands. Larger ligand is only partly enclosed and protrudes out of cavity, whilst the smaller ligand is fully enclosed. Text and images for panels b-d are adapted from Flower (1996) [474].

Additionally, SCR2 encompasses the (now familiar) L6 loop and the bottom portions of strands *f* and *g*. Strikingly, strand *h* along with the aforementioned arginine constitutes SCR3 and the outward facing residues in this strand are well placed to interact with Helix A1 (**Fig.1.20.b**). Finally, the literature hypothesises that these conserved SCRs form the basis of a receptor binding site (**Fig.6.6.c**) [474–476]. If this were the case, ligand binding to lipocalins could induce dynamic changes to the base of the protein so that cell receptors are modulated. Indeed, MUP-1 has several postulated functions and circulatory MUP-1 has been shown to regulate glucose and lipid metabolism, and there is the suggestion that it can bind its own cognate receptor in hepatocytes [477]. Moreover, a ligand that promotes male-male aggression has been shown to bind to MUP-1, which goes on to stimulate sensory neurones within the nasal cavity [478].

6.4.0. Conclusion

6.4.1. Summed protein contributions to the global entropy of binding

To assess whether $T\Delta S_{Prot}$ provides a significant contribution to the $T\Delta S°_{Glo}$ trend lines obtained from ITC, entropy differences for all bond vectors were summed. This yields a 1st order estimate as it assumes that all bond vectors are independent from one another. If correlations between bond vectors were taken into account, the magnitude of the final sum would be reduced (§3.1.2). On the other hand, the true 1st order protein contribution is likely to be underestimated because only a relatively small subset of the total number of bond vectors has been captured. As discussed in §6.1.1, NMR spectroscopy is the only wet experimental technique that can report on protein site-specific entropy changes with any accuracy. Despite this, the difficulty of capturing data for all bond vectors means that it is currently unlikely to give an accurate estimate of the true protein entropy. Whilst analysis of amides in the protein backbone and methyl groups in sidechains might give a rough indication of the protein contribution, it is important to remember that entropy differences can accumulate or cancel as more DOF are included in the summation. Thus, the missing data is a serious issue. The only technique that has the potential to capture all bond vector data and account for correlations is MD, but this is a computationally difficult task because of difficulties associated with accruing sufficient sampling and the raw compute capabilities required to process the data (§6.1.2). For these reasons the work presented in this chapter suffers from the same limitations associated with using experimental order parameters. Nonetheless, $T\Delta S_{Prot}$ trends calculated from multiple independent simulations allow the assessment of whether there is a differential protein response that is dependent on the identity of the bound ligand. The analysis allows the following facts to be established:

> **1.** Do the computationally calculated $T\Delta S_{Prot}$ trend lines provide linear contributions to the linear $T\Delta S°_{Glo}$ trends obtained from ITC? This would be indicative of a graded protein response which is dependent on ligand size.
>
> **2.** Is there is a systematic offset between calculated $T\Delta S_{Prot}$ values between n-alkanol and 3Z-olefin complexes that arises from a simple ligand structural modification? Such a difference would indicate that differences in protein dynamics contribute to the improved $T\Delta S°_{Glo}$ observed on binding unsaturated compounds compared to saturated compounds (**Fig.3.1**).

As discussed in §5.4.1.2, the contribution from ligand desolvation entropy on binding is very large and favourable. This cannot be counterbalanced with the relatively small contribution yielded by the loss of ligand DOF. Deductively, only the protein has the requisite number of DOF that can potentially produce the unfavourable contribution

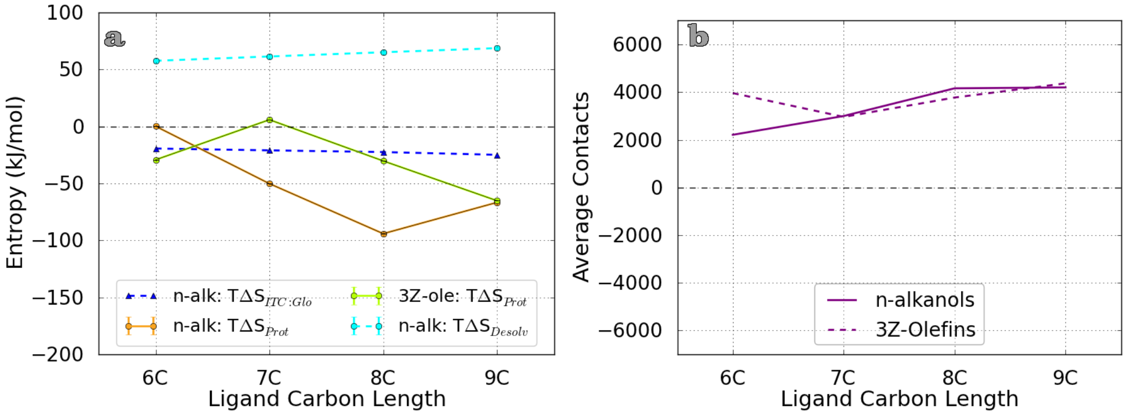
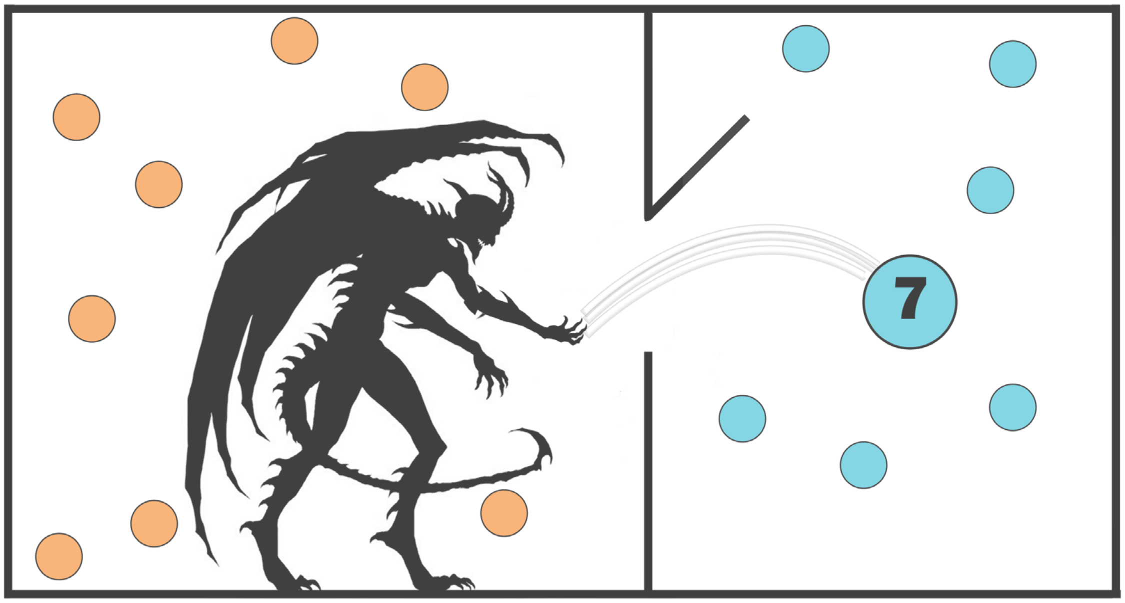


Fig.6.7. (a) Summed 1st order TΔS$_{\text{Prot}}$ values for both ligand panels. Ligand desolvation and experimental ITC values are plotted for comparison. **(b)** All protein-ligand and protein-protein contacts. Reproduced from **Fig.6.1**.b. Note that the data point for 3c6 displays an anomalous amount of protein-protein contacts (§6.3.1).

required to compensate desolvation and bring the calculated values within the range measured by global ITC. The results obtained from the simulation data indicate that this is indeed the case (**Fig.6.7**). However, the linearity of the trend lines is disrupted in both panels. Fortunately, prior analysis provides the rationale as to why this occurs. Firstly, in the case of n-alkanol holo complexes, the analysis in §4.3.9.2 indicated that the data points obtained for bound oct, and (to a lesser extent) non suffered from sampling problems. Linear trends for ligand orientational and positional entropies would be recaptured upon extending sampling (**Fig.4.28**). As there are many DOF in the protein and the correlations between them are not accounted for, differences in the n-alkanol trend line are exaggerated. In contrast, 3Z-olefins exhibit sampling issues for the smallest bound ligand, 3c6. As demonstrated in **Fig.6.7.b,** the 3c6 holo complex exhibits an anomalous amount of protein-protein intramolecular contacts which is likely to contribute to protein structural tightening. This would undoubtedly result in the TΔS$_{\text{Prot}}$ contribution of 3c6 being far more unfavourable than it ought to be. The same contacts data also supports the non-convergence of the bigger n-alkanols, oct and non. Nonetheless, after taking non-convergence into account, TΔS$_{\text{Prot}}$ makes a significant unfavourable contribution to TΔS°$_{\text{Glo}}$ and is in the right range to compensate the favourable desolvation entropy. It is expected that on including all available protein DOF and accounting for higher order correlations, values constituting the trend line will be further affected. This should correct the offset and slope of the trend lines to complement that obtained for the thermodynamic contributions produced by the other components within the system. It is also possible to conclude that the protein does display an amplified, albeit graded response in which greater reductions in TΔS$_{\text{Prot}}$ are correlated with the size of the bound ligand. We are currently in the process of acquiring T_1 & T_2 relaxation data and the preliminary results support the argument for a differential protein response (data not shown).

It is particularly interesting that a seemingly small unit modification which introduces a cis-3-4 double bond into saturated ligands results in $T\Delta S_{Prot}$ losses being greatly ameliorated. The contacts data suggests that this would be at the expense of the enthalpy. Though small differences in ligand structure can promote large shifts in the underlying enthalpies and entropies, the free energy often seems immutable because the system rebalances itself (**Fig.3.1**). When viewed from the perspective of drug design, this phenomenon is rightly considered an obstacle to the discovery of efficacious small molecule inhibitors. However, such a perspective often misses the beauty and complexity of biological systems. The free energy may remain static on binding ligands with small structural differences, but shifts in the underlying entropic and enthalpic contributions report on shifts in dynamic and structural features exhibited by the protein. These differences can contribute to variations in physiological function because different small molecules can mediate the modulation of the protein's conformations and intermolecular interactions with its cognate receptor. This in turn, can result in the differential regulation of downstream pathways within the system. The highly coupled nature of protein-ligand interactions in this system has important implications for the rational design of inhibitors in other systems that exhibit EEC because the effects of the tiniest changes are not always predictable or obvious. The investigations carried out in this thesis indicate that successful drug design would be maximised by considering systems holistically as coupled analyses of the dynamic and structural components of binding offer a compelling route through which EEC can potentially be bypassed or leveraged.



Maxwell’s Daemon

Chapter 7.0: Conclusions

7.1.0. Summary

This thesis has focussed on decomposing the global entropies of binding measured by ITC (**Fig.3.1**). There are many experimental techniques that are proficient at measuring global thermodynamic values such as ITC, but these methods typically encounter difficulties in providing rationale for the atomistic origins of observed binding signatures. This is primarily because the different components in the system can give rise to convoluted thermodynamic contributions that augment and negate one another. Thus, the assumptions underpinning most experimental decompositions are too broad, and this can obscure the subtleties of molecular interactions at the atomistic level. MD simulations provides good models of atomistic data that is extremely amenable to decomposition because its analysis relies upon building up desired thermodynamic values from the smallest base interactions. Traditionally, MD analysis has focussed on the characterisation of the structural components of binding because presumably, such interpretation is more intuitive, easier to accomplish and has already been richly defined. In contrast, the work presented in this thesis has principally focussed on the dynamic component of binding, an esoteric, niche field. The approach has yielded dividends in that, the calculation of per-unit entropies offers a way through which the behaviour of small molecules, macromolecules and solvent can be readily characterised. Synthesising this information with the structural data obtained from spatial information about the architecture of the binding cavity, intermolecular contacts and H-bond analysis has allowed a detailed picture of bimolecular interaction in MUP to be painted.

But, what is the benefit of studying the decomposition of the global thermodynamic signature of MUP? As discussed in the introduction, MUP is a well studied system which is amenable to a variety of biophysical techniques, and the results from *in silico* investigations can be readily validated against the body of work published in the literature. However, it should not be merely regarded as a toy problem and the methods developed here have broad applicability to a variety of other systems for the following reasons:

1. Lipocalins constitute a large family of proteins, and the results discovered possess a measure of transferability. For example, the development of engineered anticalins is an active field of research.

2. As detailed in §4.1.1, there are a large number of promiscuous proteins and ligands and the techniques developed here can guide the rational design of drugs. Holistically understanding the molecular relationships between structural and dynamic elements could allow the fine-tuning of activity and specificity in a similar vein to that exhibited by natural compounds.

A key factor that makes MUP an interesting system to study is its atypical binding signature and its ability to promiscuously bind many different ligands. ITC analysis indicated that on binding a panel of n-alkanols that differed incrementally in size by a single methylene group, the enthalpy became progressively more favourable, whilst the entropy became more unfavourable. As both the protein binding cavity and ligands were predominantly hydrophobic, it was initially expected that the binding signature should be entropically dominated. Malham et al. (2005) proposed that as the cavity was suboptimally hydrated, binding did not benefit from the expulsion of ordered water molecules on ligand binding. This contribution usually produces an extremely entropically favourable contribution and is considered a hallmark of apolar association [36]. In its absence, the enthalpy dominated the binding thermodynamics due to an inequality in dispersive interactions between endpoint states. At the same time, the smaller, progressive loss of entropy across the series was proposed to be due to the loss of ligand DOF (**Fig.3.1**). These experimentally derived hypotheses were the starting point for the *in silico* investigations carried out in this thesis.

What makes MUP even more interesting is its ability to bind homologous panels of ligands which possessed double bonds. The 3Z-olefin panel presented itself as an ideal counterpoint to the n-alkanol panel because the binding of these ligands was characterised by an improved entropic signature that came about at the expense of the enthalpy of binding. Therefore, EEC resulted in a net free energy change that was negligible. It was further hypothesised that greater entropic favourability of binding exhibited by 3Z-olefins was the result of ligand pre-organisation or because the entropic penalty was prepaid during the process of chemical synthesis. The rationale for the underlying shifts in the underlying entropies and enthalpies of binding between these two panels formed the focus of much of the work executed.

Initially, chapter 2.0 attempted to obtain all three thermodynamic parameters of interest from TI calculations i.e. ΔH, ΔG and TΔS. Whilst the free energy was accurately captured, entropies and enthalpies suffered in accuracy by approximately an order of magnitude. Thus, chapter 3.0 sought to capture the entropy more accurately by calculating the loss in torsional DOF. This approach is commonly used in research to characterise the ligand conformational entropy, a subcomponent of $T\Delta S_{Glo}$. If the linear decrease in $T\Delta S_{Glo}$ across ligand panels was principally derived from the loss of ligand DOF, this calculation should provide conclusive evidence for this proposal. Furthermore, there should be a systematic differential between n-alkanol and 3Z-olefin panels. However, the analysis revealed that the contribution from the conformational entropy was negligible. This was primarily because ligands bound to MUP retained significant residual motion. Despite this, analysis of per-dihedral entropies allowed hypotheses to be made regarding ligand interactions and dynamics within the cavity.

Chapter 4.0 primarily focussed on assessing the loss of the ligand's translational and rotational DOF on binding. These contributions are not commonly assessed in the literature because the calculation of the translational contribution is controversial, and most (if not all) methods that deal the rotational contribution are limited to relatively rigid molecules. Two new methods were developed that addressed some of the limitations of previous techniques. Inaccuracies in previous methods could arise as a result of suboptimal sampling and the use of functional forms that fail to characterise the multimodal distributions that are likely to be present in many biological systems. Both the methods presented benefited from long 1.2 μs simulations and did not rely on any functional form to describe the distributions. Furthermore, analysis of the rotational contributions to $T\Delta S_{Glo}$ was executed on a per-bond basis and this allowed the principal rotations of flexible molecules to be characterised. The partitioning of the rotational contribution revealed significant differences between the dynamics of the different ligands analysed. When holistically juxtaposed with COM displacements and H-Bond patterns, the analyses revealed a clearer picture of the nature of ligand dynamics and structural interactions within the context of the binding cavity. Whilst longer n-alkanol ligands tend to remain localised within defined areas within the calyx, 3Z-olefins have the ability to rapidly switch positions and orientations as a result of the rigid cis-3-4 double bond. Thus, the improved entropies of binding of unsaturated ligands compared to saturated compounds were not the result of ligand pre-organisation or due to the entropic penalty being prepaid during the process of chemical synthesis. Such mechanistic information is extremely useful and can be utilised in the rational design of small molecule inhibitors. Finally, the results indicated that summed translational and rotational entropy contributions to $T\Delta S_{Glo}$ were far more significant than the conformational contribution. This raises a doubt as to the wide-spread utility of the conformational entropy as a proxy for $T\Delta S_{Glo}$ and suggests that these other ligand contributions should also be assessed. Nonetheless, on accounting for all the ligand's DOF and roughly estimating the effect of correlations, it became apparent that this contribution (~ -15 to -25 kJ/mol) could not solely account for measured $T\Delta S_{Glo}$ values. This was primarily because the ligand desolvation entropy was very large and favourable (+57.7 to +68.8 kJ/mol for n-alkanols).

Chapter 5.0 involved the development of a MD protocol to examine the process of ligand internalisation. This approach provided mechanistic details on how the ligand gained access to the binding cavity and in the process also allowed the response of bound waters to be studied. The analysis indicated that most bound waters are expelled upon ligand entry and their release was associated with a significantly favourable entropic contribution to binding. Thus, after accounting for this and the ligand desolvation entropy, the loss of ligand DOF was an even more unlikely candidate as the primary causative agent for observed $T\Delta S_{Glo}$ values because the computationally calculated entropies of binding were much too favourable.

As the only remaining component in the system left to be assessed was the protein itself, chapter 6.0 assessed this contribution by obtaining per-residue entropies and examining protein-ligand contacts as a proxy for the binding enthalpy. The results revealed that there were systematic differences in the number of interfacial contacts made between the two panels of ligands. 3Z-olefin compounds made fewer contacts than their saturated analogues as a result of differential positions adopted in the cavity. Interestingly, on allowing for sampling issues, calculated per-residue protein entropies and the summed totals indicated that there was a differential protein response that was dependent on the identity of the bound ligand. Once again, 3Z-olefin holo complexes displayed a much reduced response. What is remarkable is how the binding of small molecules could:

1. Amplify the protein response

2. Elicit a vastly different response for comparably sized compounds as a result of small differences in ligand structure i.e. the presence or absence of a double bond.

The per-residue entropy analysis allowed the specific regions of the protein susceptible to ligand mediated modulation to be identified. This data could be correlated with experimental order parameters obtained from relaxation experiments, and the analyses supported an early theory in the literature regarding the nature of MUP's biological function. Moreover, the restriction of the many DOF within the protein provided an unfavourable entropy contribution of a magnitude capable of counterbalancing the favourable contribution yielded by desolvation. Thus, the global entropies and enthalpies of binding are the result of a combination of accumulating and cancelling contributions derived from the triumvirate of components ruling the system: Protein, ligand and solvent.

7.1.1. Future work

At the time of writing, most of the T_1 & T_2 relaxation data which will be used to experimentally corroborate the MD data has been obtained. There has a been a delay due to the upgrades taking place at the NMR suite in Leeds, but the initial results indicate that there are significant differences in the relaxation rates of the different compounds tested. A significant amount of work has also been carried out on 2nd order orientational entropy contributions, and this is currently in the final stages of testing.

Future directions include the mechanistic characterisation of other solvent exposed residues in MUP such as histidines. Such analyses would require the development of more advanced methods to characterise the H-bond network within proteins. This would be an effective tool that has broad applicability to a variety of different systems and situations.

7.1.2. Sampling

The issue of sampling has been addressed throughout this work and it is apparent that this is one of the primary weaknesses of MD simulations. When sampling is poor, there is a tendency to obtain nonsensical results, and it is often quite difficult to gauge when enough data points have been acquired. The approach taken in this thesis of testing panels of ligands that possess small incremental differences has been invaluable in identifying anomalous data points. Furthermore, well converged data points can be used to extrapolate deficient ones. These relationships allow ready assessment of 1st order entropy calculations and the computational expense associated with pinpoint accuracy is avoided. In many cases, it would be better to gain an understanding of how the system works in preference to pinpoint accuracy as this information can be more readily incorporated into the process of rational design. The simulations are relatively long (1.2 μs) compared to many of the studies published in the literature, but the results indicate that still far greater levels of sampling are required to obtain totally converged results. Such an extension would be particularly useful to further analyse the protein contribution.

It is interesting to read early papers near the beginning of this century that analyse simulations that are no more than tens of nanoseconds long. From today's perspective, these short simulations seem anachronistic. However, everything within life is a matter of perspective, and it is very likely that five to ten years from today, 1.2 μs will seem like a paltry computational effort. The nature of the subjectivity of experience is encapsulated in the story related about Zhuangzi, Huizi and the happiness of fish in the forward of this work. And so it is with science. As different scales are unlocked; whether they be related to distance or time, our view and understanding of the universe is modified and transformed. Then we realise that what we thought was true is naught but a screen masking a deeper truth.

On that note, all that is left for me to do is to sincerely thank all the people who have provided their time and knowledge to assist me in this endeavour, and to say “so long and thanks for all the fish!” [479]

Appendix A1

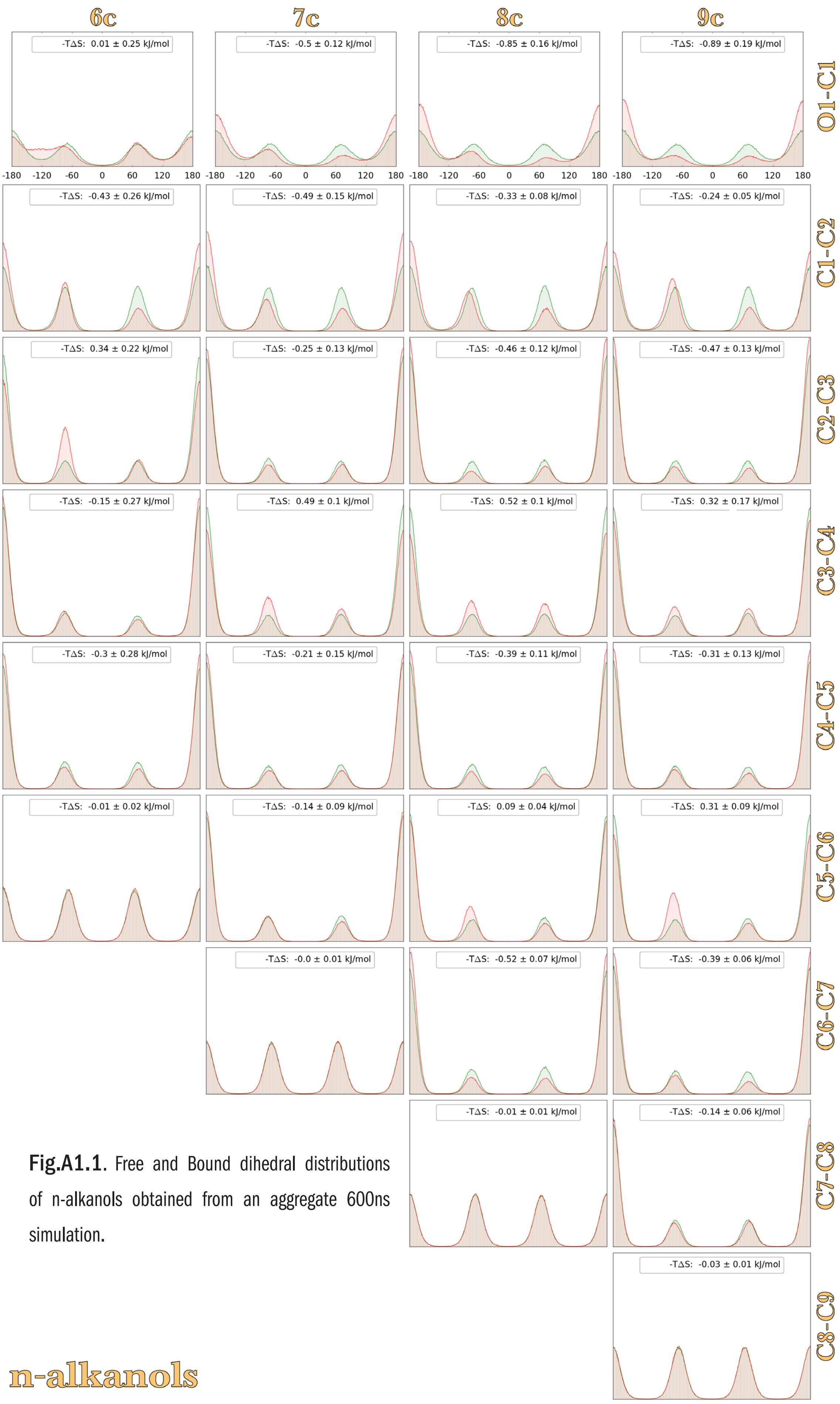


Fig.A1.1. Free and Bound dihedral distributions of n-alkanols obtained from an aggregate 600ns simulation.

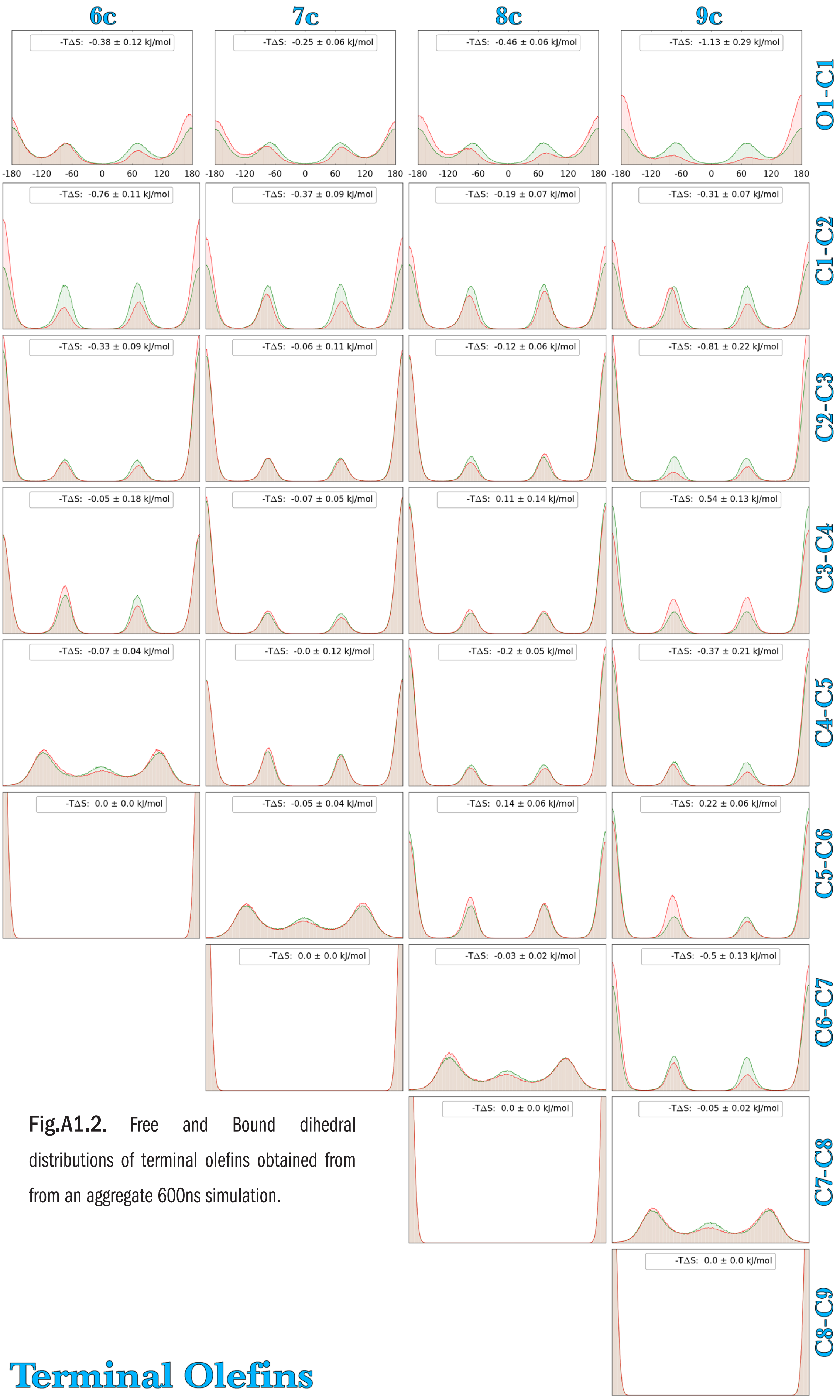


Fig.A1.2. Free and Bound dihedral distributions of terminal olefins obtained from from an aggregate 600ns simulation.

Terminal Olefins

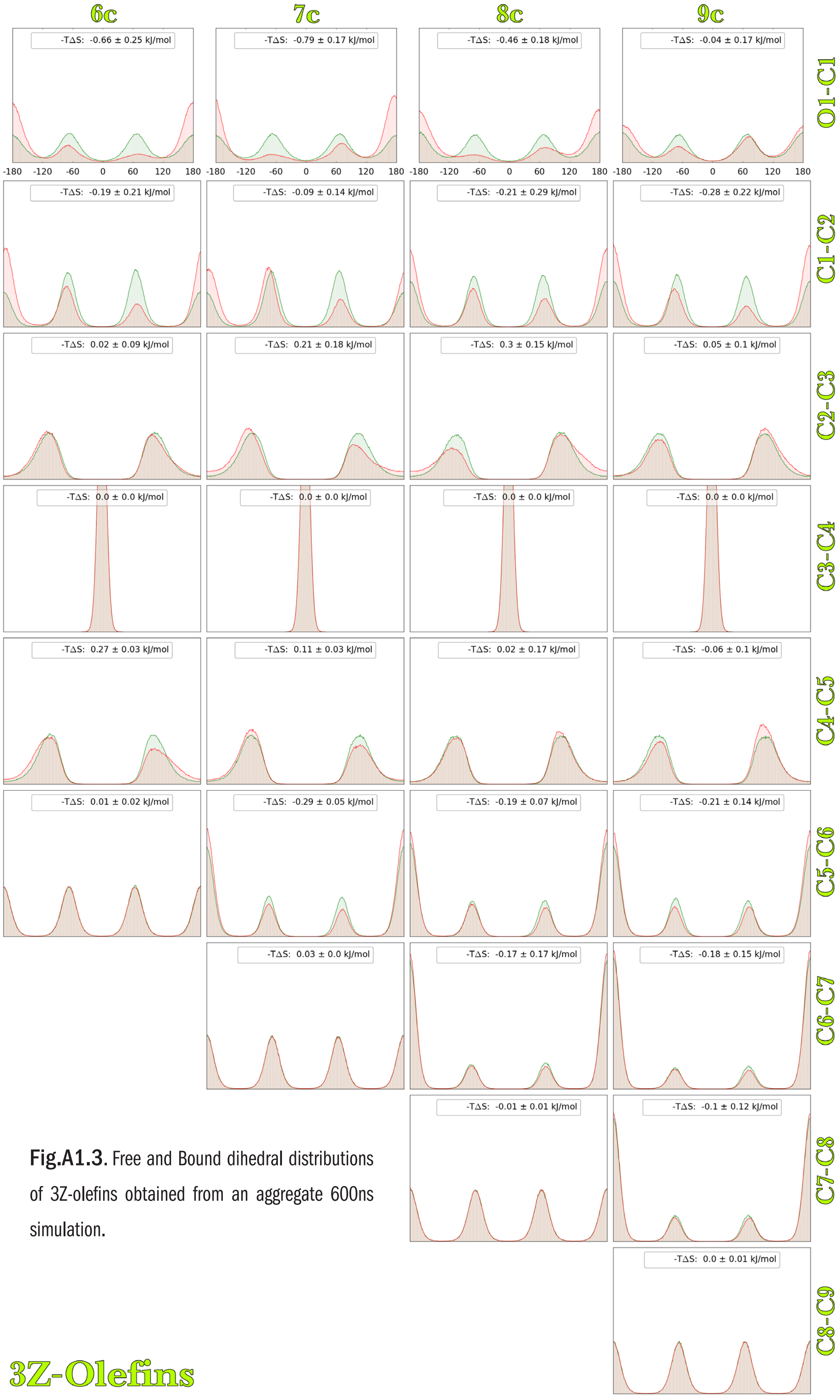


Fig.A1.3. Free and Bound dihedral distributions of 3Z-olefins obtained from an aggregate 600ns simulation.

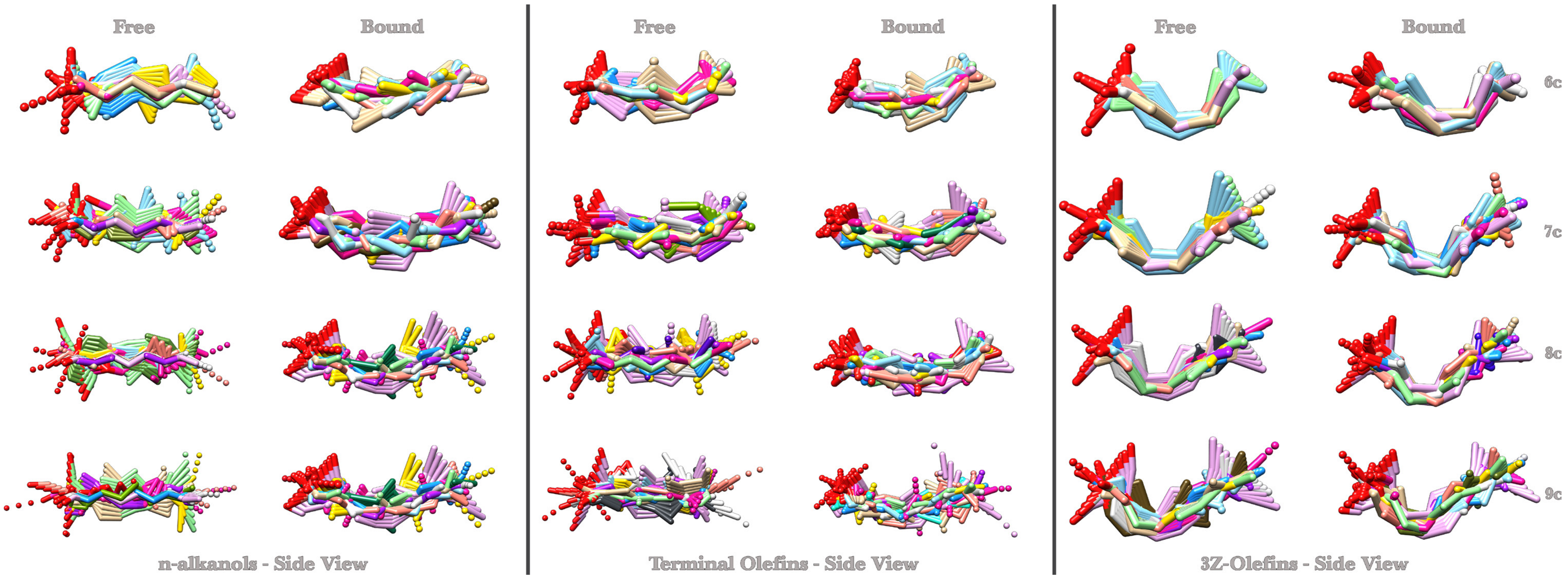


Fig.A1.4. Overlay of all ligand PCMs for both Free and Bound ligands from an aggregate 600ns simulation. View from the side.

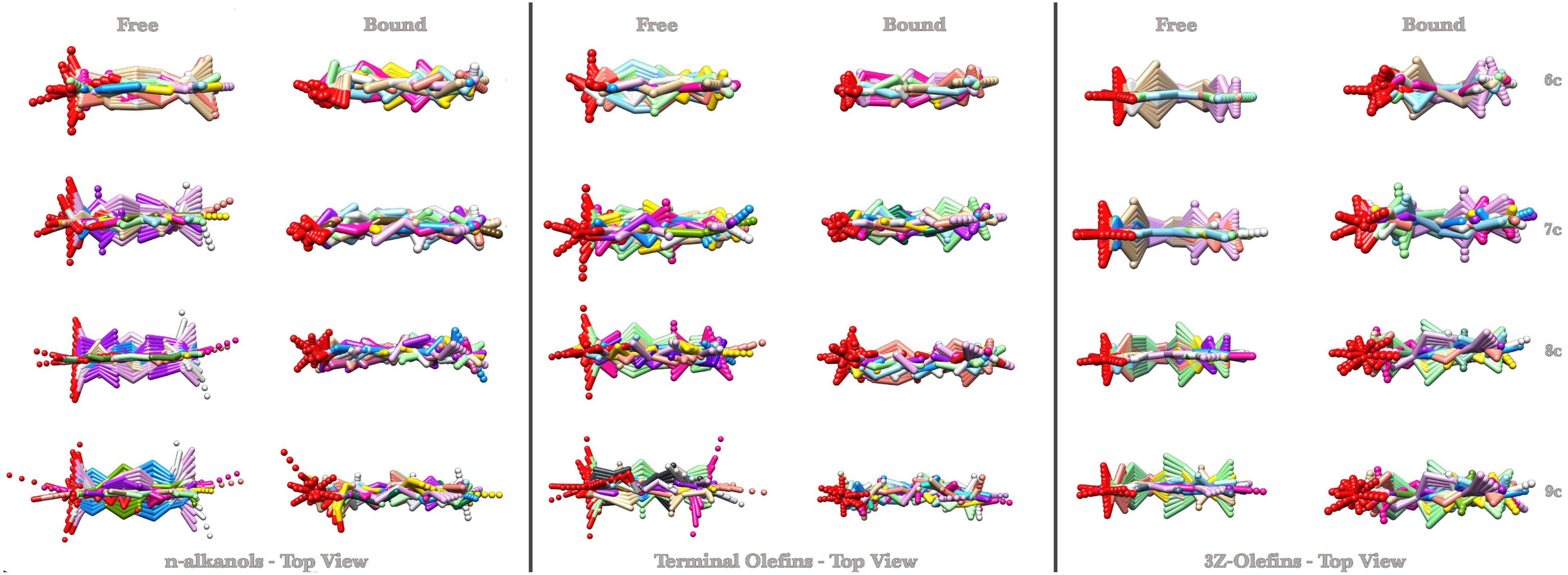


Fig.A1.5. Overlay of all ligand PCMs for both Free and Bound ligands from an aggregate 600ns simulation. View from the top.

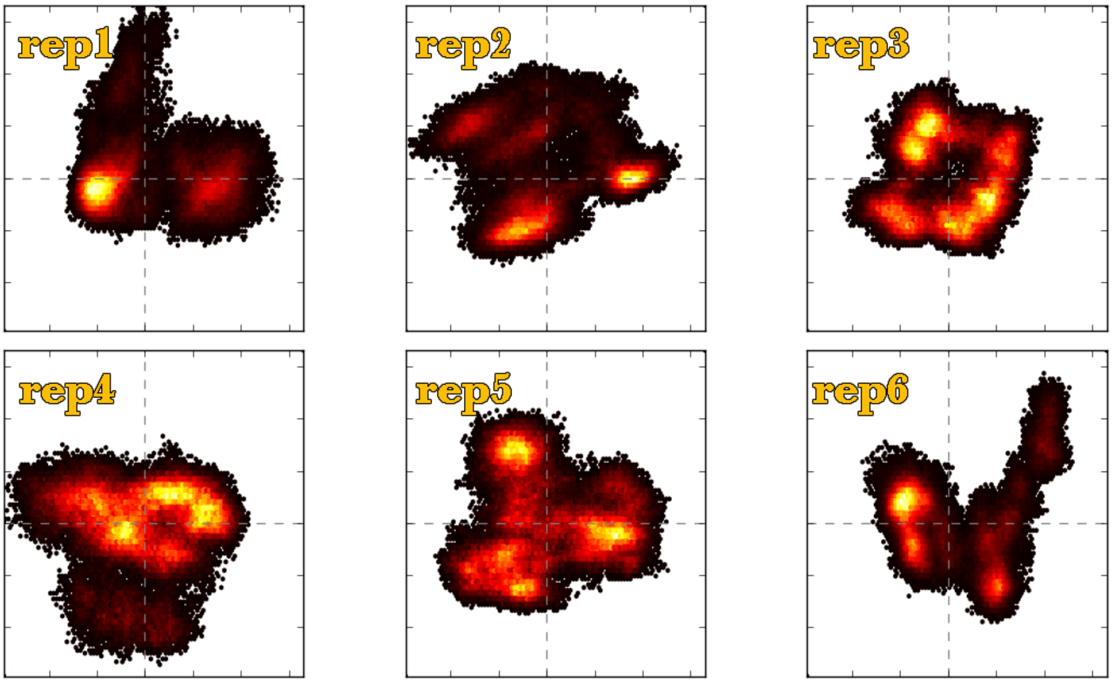


Fig.A1.7. PCM-1 versus PCM-2 showing protein conformational sampling whilst octanol is bound to MUP. The plot is generated from a single 1.2ms simulation and when compared to **Fig.A1.6** indicates that a simulation of twice the length of 6 shorter simulations does not sample conformational space as efficiently.

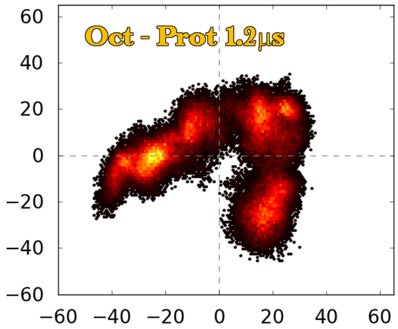


Fig.A1.6. PCM-1 versus PCM-2 showing protein conformational sampling whilst octanol is bound to MUP. Individual plots are shown for 6 100ns repeats.

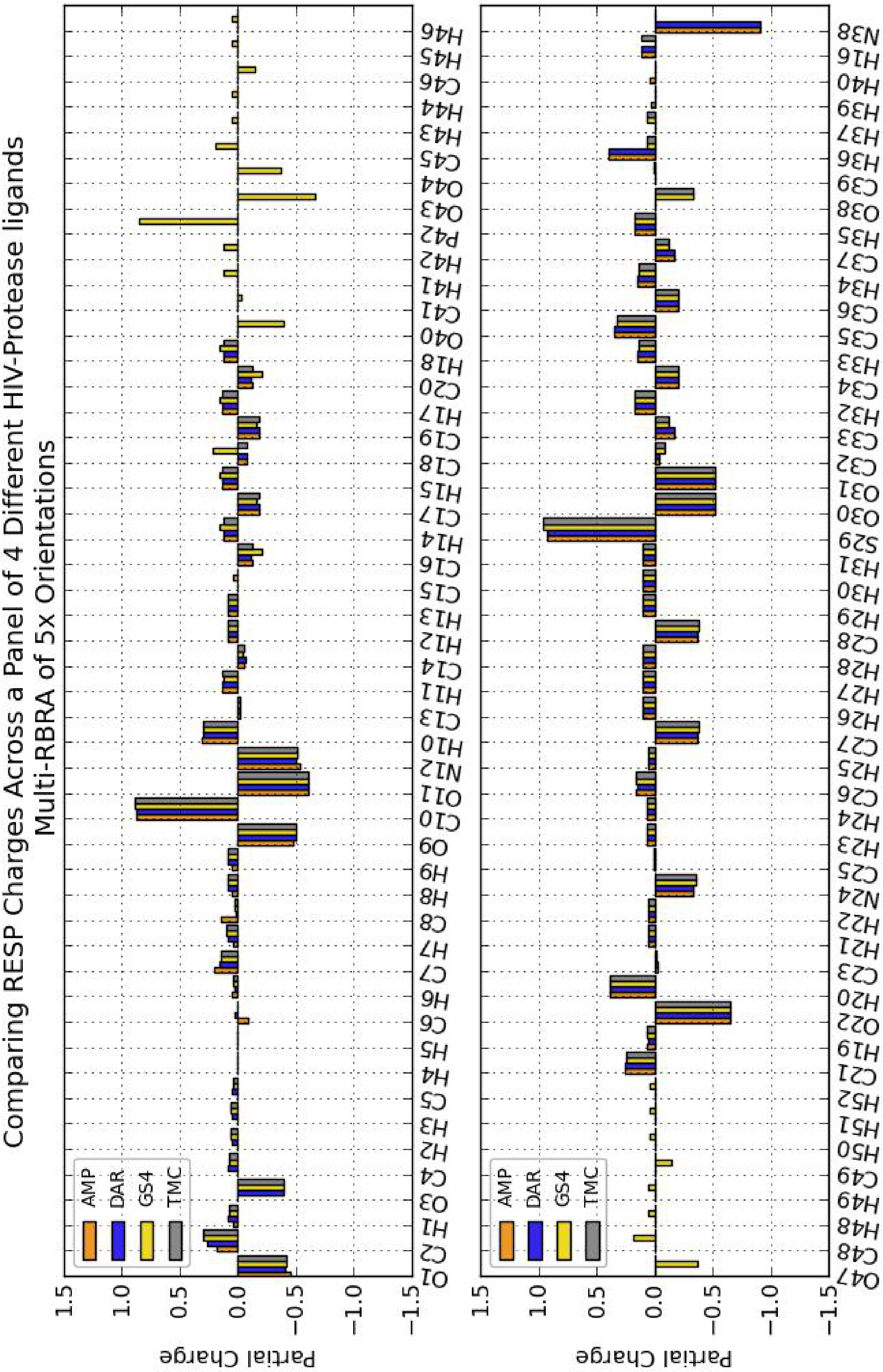


Fig.A1.8. Partial charges obtained for four HIV-Protease ligands obtained via a fragment based approach. Interchanging functional groups onto a common scaffold ensures a more consistent RESP charge derivation process.

Appendix A2

A2.1.1. The Sackur-Tetrode Equation

(Referred from §4.1.2.2)

The simplest method of assessing the loss in translational entropy on bimolecular binding is via application of the Sackur-Tetrode (ST) equation (**eqn.A2.1**). Independently derived by Otto Sackur (1880 - 1914) and Hugo Tetrode (1895 - 1931), it allows the calculation of the entropy of a monatomic ideal gas, which possesses three translational DOF for each atom. In classical mechanics, an isolated system in thermodynamic equilibrium can be chiefly described by the macroscopic variables NVE, which correspond to the number of particles in the system (N), the system volume (V) and the total energy of the system (E). This function can be written as the following where U is equal to the internal energy, m the mass, k_B the Boltzmann constant and *h* Planck's constant.

$$S(E,N,V) = Nk_B \log\left(\frac{V}{N}\left[\frac{4\pi mU}{3Nh^2}\right]^{\frac{3}{2}} + \frac{5}{2}\right) \qquad \textbf{eqn.A2.1}$$

In order to characterise the entropy of an ideal gas where the total energy and volume are held fixed, the most likely macrostate is calculated from the number of microstates available. Boltzmann realised that the best description of such a system is acquired by characterising the different states taken up by the positions (r) and conjugate momenta (p) of its constituent particles; variables that correspond to the potential and kinetic energy respectively. In the case of a single particle, position space (coordinate space) can be determined by discretising the volume into uniformly symmetric cells of arbitrary size along the three spatial axes: $\delta r_x \, \delta r_y \, \delta r_z$. The three momentum vectors can also be apportioned into cells within an imaginary space (momentum or k-space) which possess axes p_x, p_y and p_z, so that: $\delta p_x \, \delta p_y \, \delta p_z$. Discretising coordinate space is relatively straightforward but the solution to describing the "volume" of momentum space is more complicated. However, as the translational kinetic energy of the particle is constrained to equate to the internal energy we can write:

$$U = \frac{(p_x^2 + p_y^2 + p_z^2)}{2m} \qquad \textbf{eqn.A2.2}$$

This can also be written as $p^2_{x+}p^2_{y+}p^2_z = 2mU$ and momentum space can be mapped onto the surface of a sphere with a radius of $(2mU)^{1/2}$. As the energy of a large system cannot be specified exactly due to the Heisenberg uncertainty principle, the energy of the system fluctuates around its equilibrium value. Hence, the 2-dimesional area is turned into a 3-dimensional shell by multiplying by δp, and thus momentum space acquires dimensions of volume (**Fig.A2.1**). This also circumvents the issue of one of the momentum components being exactly specified; something that is impossible in a quantum mechanical treatment.

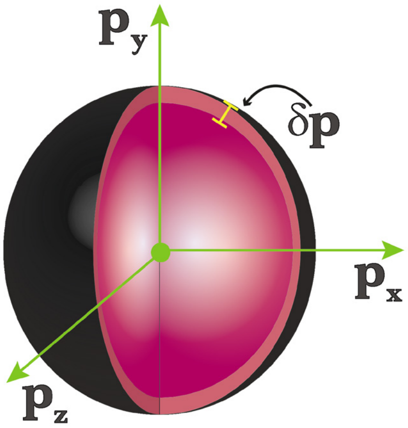


Fig.A2.1. Depiction of the momentum hypersphere with 3 principal axes labelled as P.

The total number of microstates for a one particle system with three dimensions, can be given by **eqn.A2.3**, where H = [δr δp]. H is raised to the third power because there are 3 degrees of translatory freedom and this cancels the dimensions of Ω and normalises this quantity. The choice of values for H is arbitrary and its size determines the granularity by which phase space is discretised.

$$\Omega_1 = \frac{V_r V_p}{H^3} \qquad \textbf{eqn.A2.3}$$

The total energy of a gas consisting of two particles is still constrained to sum to 2mU. So, $(p^2_{1x} + p^2_{1y} + p^2_{1z} + p^2_{2x} + p^2_{2y} + p^2_{2z}) = 2mU$ and the 6 axes in momentum space are described by a 6-dimensional hypersphere with a radius of $(2mU)^{1/2}$. In this case, swapping the positions of identical particles does not result in a distinct new state. Thus, **eqn.A2.4** is used to avoid double counting states.

$$\Omega_2 = \frac{1}{2}\frac{V_r^2}{H^6} \times (\text{momentum hypersphere volume}) \qquad \textbf{eqn.A2.4}$$

When extending the problem to the case of a gas containing N indistinguishable particles, the volume is divided into cells of $\delta r_{x1}\ \delta r_{y1}\ \delta r_{z1}\ \delta r_{x2}\ ...\ \delta r_{zN} = \delta r^{3N}$ and the momentum vectors are also divided so that - $\delta p_{x1}\ \delta p_{y1}\ \delta p_{z1}\ \delta p_{x2}...\ \delta p_{zN} = \delta p^{3N}$. This allows the structure of the various microstates available to the system to be characterised by dividing 6N-dimensional phase space into discrete units. Overcounting of the number of microstates is avoided by inserting the factor 1/N! (**eqn.A2.5**).

$$\Omega_N = \frac{1}{N!}\frac{V_r^N}{H^{3N}} \times (\text{momentum hypersphere volume}) \qquad \textbf{eqn.A2.5}$$

The area of a 3-dimensional momentum hypersphere is $4\pi r^2$, and the hyper-volume of a d-dimensional momentum hypersphere can be calculated using **eqn.A2.6**.

$$V_p = \frac{2\pi^{\frac{d}{2}}}{\left(\frac{d}{2}-1\right)!} r^{d-1} \qquad \textbf{eqn.A2.6}$$

Combining the last two equations yields **eqn.A2.7**.

$$\Omega_N = \frac{1}{N!}\frac{V^N}{H^{3N}}\frac{2\pi^{\frac{3N}{2}}}{\left(\frac{3N}{2}-1\right)!}\left(\sqrt{2mU}\right)^{3N-1} \qquad \textbf{eqn.A2.7}$$

As an ideal gas typically contains an enormous number of molecules (10^{23}), some factors can be discarded for ease of calculation by applying Stirling's approximation and taking the logarithm. This yields the Sackur-Tetrode equation (**eqn.A2.1**) [291–297].

Historically, classical thermodynamics could only produce relative entropies which were not third law compliant. The genius behind the derivations of Sackur and Tetrode was identification of the correct value for H needed to define the elementary cell volume used to discretise phase space as being h^n. n equalling the number of DOF and *h*, Planck's constant. This allowed a conceptual bridge to be made between particles such as photons on the small scale and the states taken up by massive particles, thus highlighting the importance of *h* and ensuring it became pervasive throughout physics. Their equation also adds the factor of 1/N! to remedy the "over counting" of N indistinguishable particles that occurs when applying a purely classical approach and thus avoids the Gibbs paradox. If this factor was not present, removing a partition separating two *identical* volumes of an ideal gas and allowing them to mix would result in a paradoxical increase in entropy. The corrective factor manifests as the last 5/2 term in **eqn.A2.1**. If the two gasses were distinguishable (e.g. argon and neon), this term would be 3/2. The quality of the ST equation is such that the theoretical values it provides for ideal gasses are currently considered superior to those obtained by experiment. A greater appreciation of the history around the derivation of this famous equation can be obtained by consulting Grimus (2013) [293,480,481].

A2.1.2. Measuring dynamics from crystal structures

(Referred from §4.1.2.6)

Since X-ray crystallography solved the first protein structure of myoglobin from sperm whale muscle in 1958, this technique has *dominated* our view of structural biology. As of April, 2014 there were 88,213 crystal structures versus 10,352 NMR spectroscopy structures deposited in the PDB [9,482]. NMR spectroscopy opens a window on the dynamics of macromolecules in the solution state by generating an ensemble of structures from a time scale ranging over nanoseconds to seconds. However, it is difficult to routinely obtain well resolved NMR spectra for proteins much larger than ~50 kDa without resorting to specialised methods and labelling techniques. Hence, solved NMR structures in the PDB tend to predominantly be under 10 kDa in size [483,484]. Though NMR spectroscopy can provide data on bound ligand dynamics, getting good signal to noise ratios can be problematic due to the relatively smaller number of ligand atoms containing suitable nuclei compared to those from the other constituents within the sample. Additionally, expensive labelling strategies may be required if the ligand does not naturally contain nuclei possessing nonzero spin (e.g. fluorine, hydrogen, etc). Despite these difficulties, MD simulations can reproduce chemical shifts measured by NMR spectroscopy with greater accuracy than that generated from crystal structures [485].

The magnitude of conformational changes undergone by proteins can be very large, as evidenced by the observation that deoxygenated haemoglobin crystals shatter upon exposure to oxygen [486]. However, the data X-ray crystallography provides on dynamics is limited and the static view it generates, imparts the impression that proteins and their bound ligands are a lot more ordered than the reality. This is because the crystallisation process favours recovery of averaged ground state structures from crystals that may contain ~10^{13} discrete molecules; each possessing limited conformational dynamics in the confines of the unit cells that constitute the crystalline lattice. Moreover, high resolution crystal structures often require a time period of seconds to hours to obtain good-signal to noise ratios. When this is combined with conformational averaging of a large ensemble of molecules, a further measure of dynamics is lost [483,487,488]. Despite this, many advances in the understanding of protein conformational changes have come about via the resolution of crystal structures of proteins obtained under different conditions, such as before and after ligand binding. Sometimes multiple ligand poses can be observed within holo structures such as that found in MUP, HIV PR, T4 lysozyme, neuraminidase, thymidylate synthase and cytochrome P450cam [253–256].

Frauenfelder et al (1979) used crystals of metmyoglobin obtained at four temperatures (220-300 K) to extract dynamic information from conformational substates of the protein [489]. The Debye-Waller factor (or B factor) measures the reduction in X-ray scattering caused by atomic fluctuations that atoms undergo due to thermal motion and this can be used to indirectly obtain their mean square displacements. A low value predicts structural rigidity whilst a high value, flexibility. The value contains a conformational component that describes the different substates occupied by atoms or groups of atoms, whilst the vibrational term describe static dynamics and lattice disorders [489,490]. The temperature dependence of the latter allows its estimation via linearly extrapolating B factors obtained at different temperatures back to 0 K [489,491].

However, linearity is not observed for all protein systems. For example ribonuclease A displays a “glass transition” or “kink” in the B factor trend line between 212 and 228 K. As substrate is not observed to bind to the protein below this transition point, it was suggested that flash cooling increased crystal packing and hinders the conformational movements required for binding [335,489,491]. The majority (~90%) of protein crystals have been obtained at temperatures of ~100 K because rapid cooling with cryogens such as liquid nitrogen or propane, limits the formation of damaging ice crystals and drastically reduces radiation damage caused by X-ray sources [487,491]. A study on high resolution crystals obtained at both room and cryogenic temperatures were conducted on thirty proteins by Fraser et al (2011). Their finding demonstrated that the cooling process reduced the degree of residual thermal motion, created a more compact averaged structure, increased lattice contacts (due to contraction of the unit cell), and decreased conformational populations. Furthermore, cryocooled crystals of the signalling switch enzyme HRas did not sample the catalytically competent conformations that were accessible within crystals obtained at room temperature [335].

With respect to structural dynamics, the most exciting recent developments include methods such as time-resolved Laue diffraction and time-resolved wide angle X-ray scattering (WAXS). While the former directly operates on crystal structures, the latter has been applied to proteins in solution, and is thus able to obtain a variety of temporally resolved structures that are unfettered by the crystalline lattice. Both methods add the dimension of time by capturing multiple intermediate conformations of a light-dependent structural reaction (such as ligand binding) that are induced by laser pulse excitation. If standard X-ray pulses are utilised, a resolution of ~100 ps can be obtained. Newer X-ray free electron lasers (XFELs) deliver a rapid series of femtosecond pulses that minimise sample damage and allow resolutions of 10-100 fs to be obtained. This allows crystal structures of excited states to be obtained that offer a much better picture than the ground states typically offered by crystallography [487,492,493].

A2.1.3. per-bond orientational vs. internal entropy

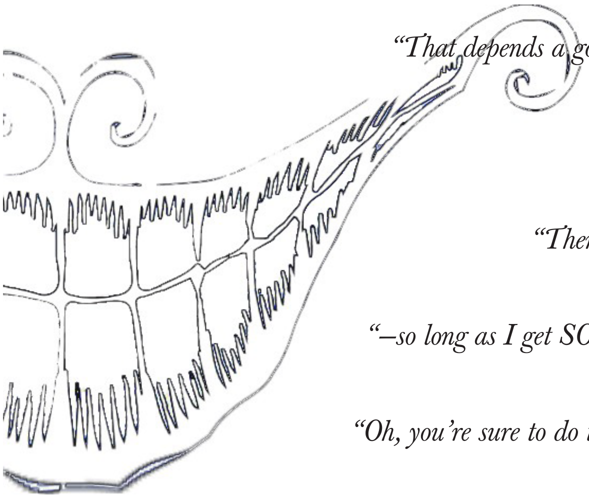
(Expanded Point 4 from §4.3.7.1)

On increasing ligand length a characteristic “alternating” pattern near the centre of the molecule is observed. The pattern is similar to that obtained from per-dihedral analysis and reports on crankshaft types of motion that occur on long linear chains confined within small volumes (§3.3.6.2 & **Fig.3.13**). Increases in per-dihedral $T\Delta S_{In}$ (within C3-C4 and C5-C6) were associated with increased conformational transitions between trans and gauche states. However, comparable increases in $T\Delta S_{Or}$ are associated with increased movements in C2-C3 and C4-C5 bonds and the peaks and troughs of the alternating pattern are inverted (**Fig.4.24**). The apparent anti-correlation is the result of differences in the method by which the two subsets of the entropy are measured. If the bond vectors that lie between atoms C2 to C5 in oct are taken as an example, it is observed that the orientational fluctuations of C2-C3 and C4-C5 bonds are relatively greater than that of C3-C4. The internal torsion of C3-C4 is defined by four atoms (C2, C3, C4 & C5) and the dihedral angle is the angle between the plane intersecting C2, C3, C4 atoms and the plane intersecting atoms C4, C5, C6. Thus the dihedral distribution is created by the relative change of C2-C3 and C4-C5 bonds with respect to one another and it logically follows that if these bonds have large $T\Delta S_{Or}$ values, there must also be a large amount of internal torsional interconversions of the C3-C4 bond between trans and gauche conformations. On the other hand, $T\Delta S_{Or}$ is obtained from the relative orientation of the C3-C4 bond after its translation through space has been removed and this measurement implicitly encompasses contributions made by bond torsions. $T\Delta S_{Or}$ is calculated by evaluating the distribution created by isolated bond vectors using spherical coordinates. On the face of it, each bond vector is equivalent to a diatomic molecule possessing two rotational degrees of freedom. However, we cannot escape the fact that the coordinates used as input for the calculation of $T\Delta S_{Or}$ implicitly contains an inseparable contribution from internal motions due to the polymeric nature of the ligand i.e. any given orientation of a bond vector has implicitly been affected by internal factors relating to the conformation of its neighbours. In theory, it should be possible to obtain the total “pure” first order rotational entropy ($T\Delta S_{Ro}$) by subtracting the *summed* values for $T\Delta S_{In}$ from that of $T\Delta S_{Or}$ (§4.3.7).

A2.1.4. The implications of non-conforming non-convergence

(Referred from §4.3.9.1)

"Cheshire Puss," she began, rather timidly, as she did not at all know whether it would like the name: however, it only grinned a little wider. *"Come, it's pleased so far,"* thought Alice, and she went on. *"Would you tell me, please, which way I ought to go from here?"*



"That depends a good deal on where you want to get to," said the Cat.

"I don't much care where–" said Alice.

"Then it doesn't matter which way you go," said the Cat.

"–so long as I get SOMEWHERE," Alice added as an explanation.

"Oh, you're sure to do that," said the Cat, "if you only walk long enough." [81]

The above passage is one of the most extensively quoted words from Lewis Carroll's seminal piece of fiction - Alice in wonderland [494]. In addition to the numerous themes regarding being and non-being - arguments related to proto-atomistic theories of matter (see Democritus versus Parmenides), the philosophical implications of ostensibly nonsense scenarios have broad applicability to a number of situations [494-496]. For example, John Kemeny framed the Cheshire cat's final response as an opposition between science and ethics, whilst Perusco et al. (2006) provide an interpretation centered within the context of new technologies and their impact on society [497-498]. Much can be read into the nonsensical adventures of Alice, but therein lie the power of the work. Like the cut-up technique popularised by William Burroughs, the juxtaposition of the absurd with the logical synthesises new meanings and insights that can be applied to diverse contexts [499]. Like Alice, the researcher has embarked on a journey; the end of which ideally marks the discovery of knowledge that illuminates the phenomena studied. To finish, they should be open to the path opened by the results and not be distracted by paths that lead away from the true destination. Unlike Alice who naively does not have a clear idea of her goal or even how to recognise it once it has been attained, researchers must have a comprehensive framework that allows them to recognise that they have reached their destination. In the case to hand, MD experiments are plagued by doubts regarding the accuracy of the force field and sampling issues. To recognise when simulations match reality and avoid mimsy conclusions, the following checks should be implemented:

1. Monitoring the change in observables with timescale.

2. Inter-methodological validation e.g. comparison to NMR spectroscopy, ITC, etc.

3. Intra-methodological validation e.g. comparison to other *in silico* methods.

4. Self-consistency across multiple independent perturbations e.g. constructing trends from ligand panels.

Grossfield et al. (2009) used the example of retinal ligand bound to dark-state rhodopsin to illustrate the contradictory results provided by simulations exploring different time scales [219]. A pertinent example examines a single ligand torsion responsible for the packing of the covalently bound ligand against a key tryptophan residue; an interaction crucial to the compound's ability to act as an inverse agonist. When simulation time was extended from 50 ns to 150 ns and then finally to 1600 ns, the conclusions regarding the predominant (gauche+, trans or gauche-) conformation of this simple localised quantity had to be continually modified because additional sampling transformed the topography upon which the previous conclusions had rested. On viewing the time series of torsional fluctuations at short time scales it is entirely possible that apparent trends are the result of localised fluctuations that rarely appear at longer timescales. Moreover, it is incredibly difficult to gauge whether a simulation has converged or not. Typical visual metrics may indicate convergence, but could actually be reporting a quasi-stable distribution (or plateau) whose value could potentially change again upon the discovery of regions of phase space previously inaccessible due to high energy barriers (§4.3.6). Thus, simulation software cannot *a priori* deliver the correct final Boltzmann weighted ratio of stable states (in the common case of imperfect sampling) that make up the ergodic distribution, because this is usually unknown. The best remedy against this is to run multiple long (> 1.0 μs) independent simulations (ideally N > 20) to maximise the probability of discovering new states and ensuring that their relative populations have stabilised [219]. The application of this measure to moderate sized (~150 residues) biological systems of interest is a computationally expensive proposition and few examples in the literature have achieved this level of sampling.

In order to know that correct converged results have been obtained, independent corroborating measurements are required. The obvious solution is cross-validation of MD results to experimental techniques such as NMR spectroscopy and ITC. However, there are issues with this cross-methodological approach as even these techniques are subject to issues of time scale and correlated errors. A study conducted by the Molecular Interactions Research Group of the Association of Biomolecular Resource Facilities (ABRF-MIRG'02) disseminated identical samples of BCA II protein and 4-carboxy-benzenesulfonamide ligand to a panel of expert ITC operators based in 14 different member laboratories. Surprisingly, errors in reported affinities and enthalpies covered a range that amounted to ~24% with reported errors severely underestimating the correct error value [253,500]. In the case of ITC, ΔG and TΔS are computed from measured K_a and ΔH values. A relatively large 20% error in the former translated to a mere 0.1 kcal/mol uncertainty in ΔG due to the logarithmic relationship in **eqn.1.5**, whilst errors in the latter were not mitigated in the same way and consequently amounted to 2.5 kcal/mol. As -TΔS is computed by subtracting ΔH from ΔG, the error in the enthalpy results in an equal and opposite shift in TΔS. In addition to the large correlated error to the entropy, this mechanism can result in the manifestation of spurious EEC effects [253]. Thus, the global entropic values used as a target for MD based analysis may potentially be erroneous.

In conjunction with cross-methodological checks, it is feasible, and indeed advisable to assess the internal consistency of multiple MD simulations describing a series of perturbations to the same protein. The correlations between multiple analyses (such as H-bonding, positional and orientational entropies) obtained from a *single* protein-ligand binding interaction could be dismissed as self-fulfilling because correlations between the different results could be the consequence of the rules governing the simulation. However, the inter-woven correlations between trends obtained from a panel of multiple independent binding interactions are more reliable, because any disagreement in the trends rapidly highlights inadequacies in hypothesis, method and sampling. Thus, if non-convergence is suspected, outliers can easily be identified by non-conformance and data from other members in the panel used to reconstruct the road to the true destination.

A2.2.1. Supplementary tables & figures

hex		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	1.38 ± 0.18	-1.13 ± 0.18
C1-C2	2.51 ± 0.00	1.54 ± 0.16	-0.97 ± 0.16
C2-C3	2.51 ± 0.00	1.46 ± 0.16	-1.05 ± 0.16
C3-C4	2.51 ± 0.00	1.58 ± 0.13	-0.93 ± 0.13
C4-C5	2.51 ± 0.00	1.64 ± 0.15	-0.87 ± 0.15
C5-C6	2.51 ± 0.00	1.82 ± 0.12	-0.69 ± 0.12

3c6		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	0.71 ± 0.22	-1.80 ± 0.22
C1-C2	2.51 ± 0.00	0.80 ± 0.24	-1.71 ± 0.24
C2-C3	2.51 ± 0.00	1.09 ± 0.20	-1.42 ± 0.20
C3-C4	2.51 ± 0.00	0.97 ± 0.13	-1.54 ± 0.13
C4-C5	2.51 ± 0.00	1.54 ± 0.14	-0.97 ± 0.14
C5-C6	2.51 ± 0.00	1.92 ± 0.07	-0.59 ± 0.07

hep		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	1.28 ± 0.15	-1.23 ± 0.15
C1-C2	2.51 ± 0.00	1.46 ± 0.13	-1.05 ± 0.13
C2-C3	2.51 ± 0.00	1.51 ± 0.08	-1.0 ± 0.08
C3-C4	2.51 ± 0.00	1.48 ± 0.1	-1.03 ± 0.1
C4-C5	2.51 ± 0.00	1.64 ± 0.04	-0.87 ± 0.04
C5-C6	2.51 ± 0.00	1.63 ± 0.05	-0.88 ± 0.05
C6-C7	2.51 ± 0.00	1.81 ± 0.05	-0.7 ± 0.05

3c7		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	0.57 ± 0.17	-1.94 ± 0.17
C1-C2	2.51 ± 0.00	0.75 ± 0.18	-1.76 ± 0.18
C2-C3	2.51 ± 0.00	0.99 ± 0.21	-1.52 ± 0.21
C3-C4	2.51 ± 0.00	0.78 ± 0.16	-1.73 ± 0.16
C4-C5	2.51 ± 0.00	1.39 ± 0.19	-1.12 ± 0.19
C5-C6	2.51 ± 0.00	1.79 ± 0.12	-0.72 ± 0.12
C6-C7	2.51 ± 0.00	1.81 ± 0.12	-0.70 ± 0.12

oct		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	1.05 ± 0.14	-1.46 ± 0.14
C1-C2	2.51 ± 0.00	1.06 ± 0.16	-1.45 ± 0.16
C2-C3	2.51 ± 0.00	1.3 ± 0.08	-1.21 ± 0.08
C3-C4	2.51 ± 0.00	1.0 ± 0.13	-1.51 ± 0.13
C4-C5	2.51 ± 0.00	1.27 ± 0.09	-1.24 ± 0.09
C5-C6	2.51 ± 0.00	1.06 ± 0.12	-1.45 ± 0.12
C6-C7	2.51 ± 0.00	1.31 ± 0.11	-1.2 ± 0.11
C7-C8	2.51 ± 0.00	1.36 ± 0.1	-1.15 ± 0.1

3c8		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	0.87 ± 0.16	-1.64 ± 0.16
C1-C2	2.51 ± 0.00	1.21 ± 0.10	-1.30 ± 0.10
C2-C3	2.51 ± 0.00	1.23 ± 0.09	-1.28 ± 0.09
C3-C4	2.51 ± 0.00	1.08 ± 0.10	-1.43 ± 0.10
C4-C5	2.51 ± 0.00	1.50 ± 0.10	-1.01 ± 0.10
C5-C6	2.51 ± 0.00	1.67 ± 0.08	-0.84 ± 0.08
C6-C7	2.51 ± 0.00	1.77 ± 0.07	-0.74 ± 0.07
C7-C8	2.51 ± 0.00	1.87 ± 0.08	-0.64 ± 0.08

non		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	1.06 ± 0.22	-1.45 ± 0.22
C1-C2	2.51 ± 0.00	1.09 ± 0.16	-1.42 ± 0.16
C2-C3	2.51 ± 0.00	1.25 ± 0.1	-1.26 ± 0.1
C3-C4	2.51 ± 0.00	1.0 ± 0.13	-1.51 ± 0.13
C4-C5	2.51 ± 0.00	1.26 ± 0.06	-1.25 ± 0.06
C5-C6	2.51 ± 0.00	1.04 ± 0.09	-1.47 ± 0.09
C6-C7	2.51 ± 0.00	1.16 ± 0.06	-1.35 ± 0.06
C7-C8	2.51 ± 0.00	1.24 ± 0.06	-1.27 ± 0.06
C8-C9	2.51 ± 0.00	1.54 ± 0.03	-0.97 ± 0.03

3c9		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.51 ± 0.00	1.25 ± 0.25	-1.26 ± 0.25
C1-C2	2.51 ± 0.00	1.49 ± 0.22	-1.02 ± 0.22
C2-C3	2.51 ± 0.00	1.15 ± 0.26	-1.36 ± 0.26
C3-C4	2.51 ± 0.00	0.80 ± 0.20	-1.71 ± 0.20
C4-C5	2.51 ± 0.00	1.19 ± 0.20	-1.32 ± 0.20
C5-C6	2.51 ± 0.00	1.18 ± 0.19	-1.33 ± 0.19
C6-C7	2.51 ± 0.00	1.33 ± 0.15	-1.18 ± 0.15
C7-C8	2.51 ± 0.00	1.33 ± 0.17	-1.18 ± 0.17
C8-C9	2.51 ± 0.00	1.57 ± 0.10	-0.94 ± 0.10

hex		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.16 ± 0.03	0.16 ± 0.03
C1-C2	0.00 ± 0.00	0.09 ± 0.03	0.09 ± 0.03
C2-C3	0.00 ± 0.00	0.12 ± 0.03	0.12 ± 0.03
C3-C4	0.00 ± 0.00	0.08 ± 0.01	0.08 ± 0.01
C4-C5	0.00 ± 0.00	0.09 ± 0.03	0.09 ± 0.03
C5-C6	0.00 ± 0.00	0.06 ± 0.01	0.06 ± 0.01

3c6		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.34 ± 0.08	0.34 ± 0.08
C1-C2	0.00 ± 0.00	0.32 ± 0.09	0.32 ± 0.09
C2-C3	0.00 ± 0.00	0.17 ± 0.07	0.17 ± 0.07
C3-C4	0.00 ± 0.00	0.27 ± 0.04	0.27 ± 0.04
C4-C5	0.00 ± 0.00	0.14 ± 0.04	0.14 ± 0.04
C5-C6	0.00 ± 0.00	0.05 ± 0.02	0.05 ± 0.02

hep		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.17 ± 0.04	0.17 ± 0.04
C1-C2	0.00 ± 0.00	0.09 ± 0.04	0.09 ± 0.04
C2-C3	0.00 ± 0.00	0.10 ± 0.01	0.10 ± 0.01
C3-C4	0.00 ± 0.00	0.12 ± 0.02	0.12 ± 0.02
C4-C5	0.00 ± 0.00	0.08 ± 0.01	0.08 ± 0.01
C5-C6	0.00 ± 0.00	0.09 ± 0.02	0.09 ± 0.02
C6-C7	0.00 ± 0.00	0.08 ± 0.02	0.08 ± 0.02

3c7		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.42 ± 0.05	0.42 ± 0.05
C1-C2	0.00 ± 0.00	0.33 ± 0.05	0.33 ± 0.05
C2-C3	0.00 ± 0.00	0.19 ± 0.08	0.19 ± 0.08
C3-C4	0.00 ± 0.00	0.36 ± 0.06	0.36 ± 0.06
C4-C5	0.00 ± 0.00	0.18 ± 0.05	0.18 ± 0.05
C5-C6	0.00 ± 0.00	0.06 ± 0.02	0.06 ± 0.02
C6-C7	0.00 ± 0.00	0.09 ± 0.02	0.09 ± 0.02

oct		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.30 ± 0.04	0.30 ± 0.04
C1-C2	0.00 ± 0.00	0.23 ± 0.05	0.23 ± 0.05
C2-C3	0.00 ± 0.00	0.11 ± 0.01	0.11 ± 0.01
C3-C4	0.00 ± 0.00	0.28 ± 0.05	0.28 ± 0.05
C4-C5	0.00 ± 0.00	0.10 ± 0.01	0.10 ± 0.01
C5-C6	0.00 ± 0.00	0.26 ± 0.04	0.26 ± 0.04
C6-C7	0.00 ± 0.00	0.17 ± 0.03	0.17 ± 0.03
C7-C8	0.00 ± 0.00	0.19 ± 0.03	0.19 ± 0.03

3c8		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.31 ± 0.04	0.31 ± 0.04
C1-C2	0.00 ± 0.00	0.16 ± 0.03	0.16 ± 0.03
C2-C3	0.00 ± 0.00	0.17 ± 0.05	0.17 ± 0.05
C3-C4	0.00 ± 0.00	0.22 ± 0.03	0.22 ± 0.03
C4-C5	0.00 ± 0.00	0.14 ± 0.04	0.14 ± 0.04
C5-C6	0.00 ± 0.00	0.11 ± 0.02	0.11 ± 0.02
C6-C7	0.00 ± 0.00	0.10 ± 0.01	0.10 ± 0.01
C7-C8	0.00 ± 0.00	0.09 ± 0.01	0.09 ± 0.01

non		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.28 ± 0.07	0.28 ± 0.07
C1-C2	0.00 ± 0.00	0.21 ± 0.05	0.21 ± 0.05
C2-C3	0.00 ± 0.00	0.12 ± 0.01	0.12 ± 0.01
C3-C4	0.00 ± 0.00	0.26 ± 0.06	0.26 ± 0.06
C4-C5	0.00 ± 0.00	0.11 ± 0.01	0.11 ± 0.01
C5-C6	0.00 ± 0.00	0.25 ± 0.04	0.25 ± 0.04
C6-C7	0.00 ± 0.00	0.17 ± 0.02	0.17 ± 0.02
C7-C8	0.00 ± 0.00	0.19 ± 0.02	0.19 ± 0.02
C8-C9	0.00 ± 0.00	0.08 ± 0.01	0.08 ± 0.01

3c9		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.22 ± 0.06	0.22 ± 0.06
C1-C2	0.00 ± 0.00	0.10 ± 0.05	0.10 ± 0.05
C2-C3	0.00 ± 0.00	0.24 ± 0.08	0.24 ± 0.08
C3-C4	0.00 ± 0.00	0.31 ± 0.08	0.31 ± 0.08
C4-C5	0.00 ± 0.00	0.19 ± 0.07	0.19 ± 0.07
C5-C6	0.00 ± 0.00	0.18 ± 0.06	0.18 ± 0.06
C6-C7	0.00 ± 0.00	0.15 ± 0.03	0.15 ± 0.03
C7-C8	0.00 ± 0.00	0.17 ± 0.06	0.17 ± 0.06
C8-C9	0.00 ± 0.00	0.10 ± 0.02	0.10 ± 0.02

Table.A2.2.1. Averaged rotational entropies and order parameters calculated from 6x 200 ns simulations for bound, free and "bound minus free". Errors are shown as standard errors.

hex		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.80 ± 0.18	-0.73 ± 0.18
C1-C2	2.53 ± 0.00	1.91 ± 0.16	-0.62 ± 0.16
C2-C3	2.53 ± 0.00	1.74 ± 0.16	-0.79 ± 0.16
C3-C4	2.53 ± 0.00	1.87 ± 0.13	-0.66 ± 0.13
C4-C5	2.53 ± 0.00	1.87 ± 0.15	-0.66 ± 0.15
C5-C6	2.53 ± 0.00	2.04 ± 0.12	-0.49 ± 0.12

3c6		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.12 ± 0.22	-1.41 ± 0.22
C1-C2	2.53 ± 0.00	1.17 ± 0.24	-1.36 ± 0.24
C2-C3	2.53 ± 0.00	1.45 ± 0.20	-1.08 ± 0.20
C3-C4	2.53 ± 0.00	1.38 ± 0.13	-1.15 ± 0.13
C4-C5	2.53 ± 0.00	1.95 ± 0.14	-0.58 ± 0.14
C5-C6	2.53 ± 0.00	2.19 ± 0.07	-0.34 ± 0.07

hex		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.07 ± 0.03	0.07 ± 0.03
C1-C2	0.00 ± 0.00	0.02 ± 0.03	0.02 ± 0.03
C2-C3	0.00 ± 0.00	0.09 ± 0.03	0.09 ± 0.03
C3-C4	0.00 ± 0.00	0.05 ± 0.01	0.05 ± 0.01
C4-C5	0.00 ± 0.00	0.05 ± 0.03	0.05 ± 0.03
C5-C6	0.00 ± 0.00	0.03 ± 0.01	0.03 ± 0.01

3c6		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.21 ± 0.08	0.21 ± 0.08
C1-C2	0.00 ± 0.00	0.20 ± 0.09	0.20 ± 0.09
C2-C3	0.00 ± 0.00	0.10 ± 0.07	0.10 ± 0.07
C3-C4	0.00 ± 0.00	0.12 ± 0.04	0.12 ± 0.04
C4-C5	0.00 ± 0.00	0.07 ± 0.04	0.07 ± 0.04
C5-C6	0.00 ± 0.00	0.02 ± 0.02	0.02 ± 0.02

hep		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.61 ± 0.15	-0.92 ± 0.15
C1-C2	2.53 ± 0.00	1.79 ± 0.13	-0.74 ± 0.13
C2-C3	2.53 ± 0.00	1.80 ± 0.08	-0.73 ± 0.08
C3-C4	2.53 ± 0.00	1.76 ± 0.10	-0.77 ± 0.10
C4-C5	2.53 ± 0.00	1.89 ± 0.04	-0.64 ± 0.04
C5-C6	2.53 ± 0.00	1.84 ± 0.05	-0.69 ± 0.05
C6-C7	2.53 ± 0.00	2.02 ± 0.05	-0.51 ± 0.05

3c7		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	0.90 ± 0.17	-1.63 ± 0.17
C1-C2	2.53 ± 0.00	1.05 ± 0.18	-1.48 ± 0.18
C2-C3	2.53 ± 0.00	1.47 ± 0.21	-1.06 ± 0.21
C3-C4	2.53 ± 0.00	1.36 ± 0.16	-1.17 ± 0.16
C4-C5	2.53 ± 0.00	1.91 ± 0.19	-0.62 ± 0.19
C5-C6	2.53 ± 0.00	2.20 ± 0.12	-0.33 ± 0.12
C6-C7	2.53 ± 0.00	2.14 ± 0.12	-0.39 ± 0.12

hep		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.11 ± 0.04	0.11 ± 0.04
C1-C2	0.00 ± 0.00	0.03 ± 0.04	0.03 ± 0.04
C2-C3	0.00 ± 0.00	0.06 ± 0.01	0.06 ± 0.01
C3-C4	0.00 ± 0.00	0.07 ± 0.02	0.07 ± 0.02
C4-C5	0.00 ± 0.00	0.04 ± 0.01	0.04 ± 0.01
C5-C6	0.00 ± 0.00	0.06 ± 0.02	0.06 ± 0.02
C6-C7	0.00 ± 0.00	0.04 ± 0.02	0.04 ± 0.02

3c7		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.32 ± 0.05	0.32 ± 0.05
C1-C2	0.00 ± 0.00	0.26 ± 0.05	0.26 ± 0.05
C2-C3	0.00 ± 0.00	0.05 ± 0.08	0.05 ± 0.08
C3-C4	0.00 ± 0.00	0.13 ± 0.06	0.13 ± 0.06
C4-C5	0.00 ± 0.00	0.09 ± 0.05	0.09 ± 0.05
C5-C6	0.00 ± 0.00	0.01 ± 0.02	0.01 ± 0.02
C6-C7	0.00 ± 0.00	0.06 ± 0.02	0.06 ± 0.02

oct		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.23 ± 0.14	-1.30 ± 0.14
C1-C2	2.53 ± 0.00	1.26 ± 0.16	-1.27 ± 0.16
C2-C3	2.53 ± 0.00	1.49 ± 0.08	-1.04 ± 0.08
C3-C4	2.53 ± 0.00	1.23 ± 0.13	-1.30 ± 0.13
C4-C5	2.53 ± 0.00	1.49 ± 0.09	-1.04 ± 0.09
C5-C6	2.53 ± 0.00	1.27 ± 0.12	-1.26 ± 0.12
C6-C7	2.53 ± 0.00	1.50 ± 0.11	-1.03 ± 0.11
C7-C8	2.53 ± 0.00	1.51 ± 0.10	-1.02 ± 0.10

3c8		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.23 ± 0.16	-1.30 ± 0.16
C1-C2	2.53 ± 0.00	1.60 ± 0.10	-0.93 ± 0.10
C2-C3	2.53 ± 0.00	1.63 ± 0.09	-0.90 ± 0.09
C3-C4	2.53 ± 0.00	1.54 ± 0.10	-0.99 ± 0.10
C4-C5	2.53 ± 0.00	1.88 ± 0.10	-0.65 ± 0.10
C5-C6	2.53 ± 0.00	1.97 ± 0.08	-0.56 ± 0.08
C6-C7	2.53 ± 0.00	2.02 ± 0.07	-0.51 ± 0.07
C7-C8	2.53 ± 0.00	2.11 ± 0.08	-0.42 ± 0.08

oct		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.28 ± 0.04	0.28 ± 0.04
C1-C2	0.00 ± 0.00	0.20 ± 0.05	0.20 ± 0.05
C2-C3	0.00 ± 0.00	0.10 ± 0.01	0.10 ± 0.01
C3-C4	0.00 ± 0.00	0.23 ± 0.05	0.23 ± 0.05
C4-C5	0.00 ± 0.00	0.05 ± 0.01	0.05 ± 0.01
C5-C6	0.00 ± 0.00	0.22 ± 0.04	0.22 ± 0.04
C6-C7	0.00 ± 0.00	0.12 ± 0.03	0.12 ± 0.03
C7-C8	0.00 ± 0.00	0.17 ± 0.03	0.17 ± 0.03

3c8		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.21 ± 0.04	0.21 ± 0.04
C1-C2	0.00 ± 0.00	0.08 ± 0.03	0.08 ± 0.03
C2-C3	0.00 ± 0.00	0.11 ± 0.05	0.11 ± 0.05
C3-C4	0.00 ± 0.00	0.08 ± 0.03	0.08 ± 0.03
C4-C5	0.00 ± 0.00	0.10 ± 0.04	0.10 ± 0.04
C5-C6	0.00 ± 0.00	0.07 ± 0.02	0.07 ± 0.02
C6-C7	0.00 ± 0.00	0.09 ± 0.01	0.09 ± 0.01
C7-C8	0.00 ± 0.00	0.06 ± 0.01	0.06 ± 0.01

non		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.37 ± 0.22	-1.16 ± 0.22
C1-C2	2.53 ± 0.00	1.35 ± 0.16	-1.18 ± 0.16
C2-C3	2.53 ± 0.00	1.46 ± 0.10	-1.07 ± 0.10
C3-C4	2.53 ± 0.00	1.23 ± 0.13	-1.30 ± 0.13
C4-C5	2.53 ± 0.00	1.43 ± 0.06	-1.10 ± 0.06
C5-C6	2.53 ± 0.00	1.20 ± 0.09	-1.33 ± 0.09
C6-C7	2.53 ± 0.00	1.29 ± 0.06	-1.24 ± 0.06
C7-C8	2.53 ± 0.00	1.34 ± 0.06	-1.19 ± 0.06
C8-C9	2.53 ± 0.00	1.62 ± 0.03	-0.91 ± 0.03

3c9		Entropy (k_B)	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	2.53 ± 0.00	1.63 ± 0.25	-0.90 ± 0.25
C1-C2	2.53 ± 0.00	1.94 ± 0.22	-0.59 ± 0.22
C2-C3	2.53 ± 0.00	1.55 ± 0.26	-0.98 ± 0.26
C3-C4	2.53 ± 0.00	1.24 ± 0.20	-1.29 ± 0.20
C4-C5	2.53 ± 0.00	1.53 ± 0.20	-1.00 ± 0.20
C5-C6	2.53 ± 0.00	1.39 ± 0.19	-1.14 ± 0.19
C6-C7	2.53 ± 0.00	1.54 ± 0.15	-0.99 ± 0.15
C7-C8	2.53 ± 0.00	1.49 ± 0.17	-1.04 ± 0.17
C8-C9	2.53 ± 0.00	1.71 ± 0.10	-0.82 ± 0.10

non		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.21 ± 0.07	0.21 ± 0.07
C1-C2	0.00 ± 0.00	0.16 ± 0.05	0.16 ± 0.05
C2-C3	0.00 ± 0.00	0.10 ± 0.01	0.10 ± 0.01
C3-C4	0.00 ± 0.00	0.22 ± 0.06	0.22 ± 0.06
C4-C5	0.00 ± 0.00	0.09 ± 0.01	0.09 ± 0.01
C5-C6	0.00 ± 0.00	0.22 ± 0.04	0.22 ± 0.04
C6-C7	0.00 ± 0.00	0.15 ± 0.02	0.15 ± 0.02
C7-C8	0.00 ± 0.00	0.18 ± 0.02	0.18 ± 0.02
C8-C9	0.00 ± 0.00	0.07 ± 0.01	0.07 ± 0.01

3c9		Order Parameters	
Bond Vec	**Free**	**Bound**	**Difference**
O1-C1	0.00 ± 0.00	0.16 ± 0.06	0.16 ± 0.06
C1-C2	0.00 ± 0.00	0.04 ± 0.05	0.04 ± 0.05
C2-C3	0.00 ± 0.00	0.19 ± 0.08	0.19 ± 0.08
C3-C4	0.00 ± 0.00	0.20 ± 0.08	0.20 ± 0.08
C4-C5	0.00 ± 0.00	0.13 ± 0.07	0.13 ± 0.07
C5-C6	0.00 ± 0.00	0.14 ± 0.06	0.14 ± 0.06
C6-C7	0.00 ± 0.00	0.12 ± 0.03	0.12 ± 0.03
C7-C8	0.00 ± 0.00	0.14 ± 0.06	0.14 ± 0.06
C8-C9	0.00 ± 0.00	0.08 ± 0.02	0.08 ± 0.02

Table.A2.2.2. Aggregate rotational entropies and order parameters calculated from 1x 1.2 μs simulations for bound, free and "bound minus free". Errors are shown as standard errors.

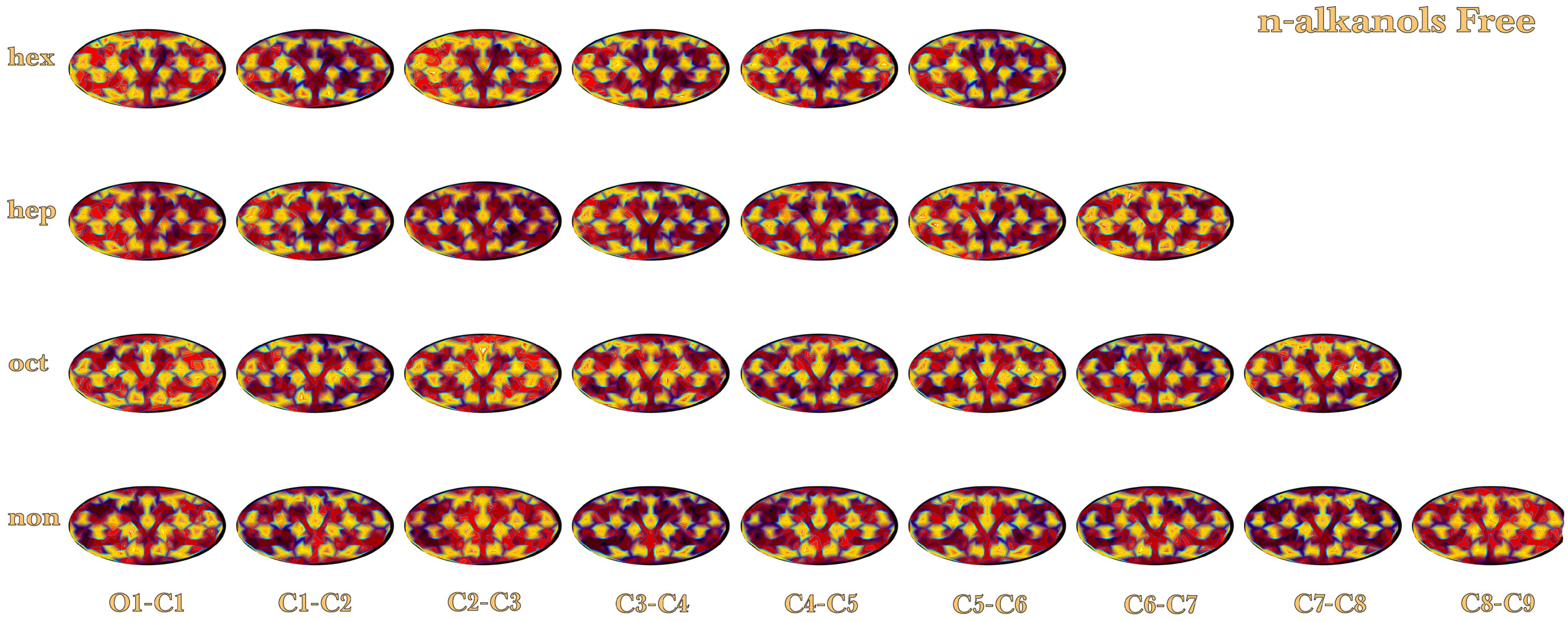


Fig.A2.2.1. Free n-alkanol panel: Hammer projection showing bond vector rotational motion on the surface of a sphere

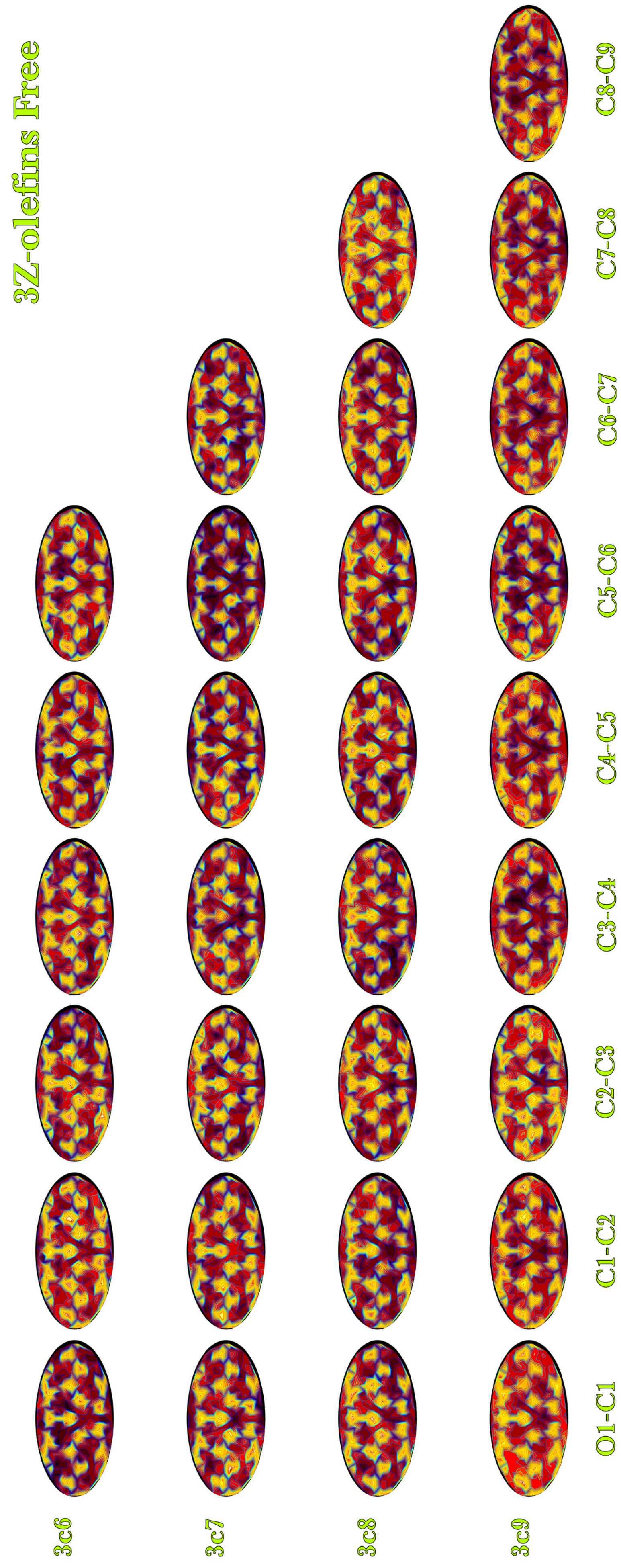


Fig.A2.2.2. Free 3Z-olefin panel: Hammer projection showing bond vector rotational motion on the surface of a sphere

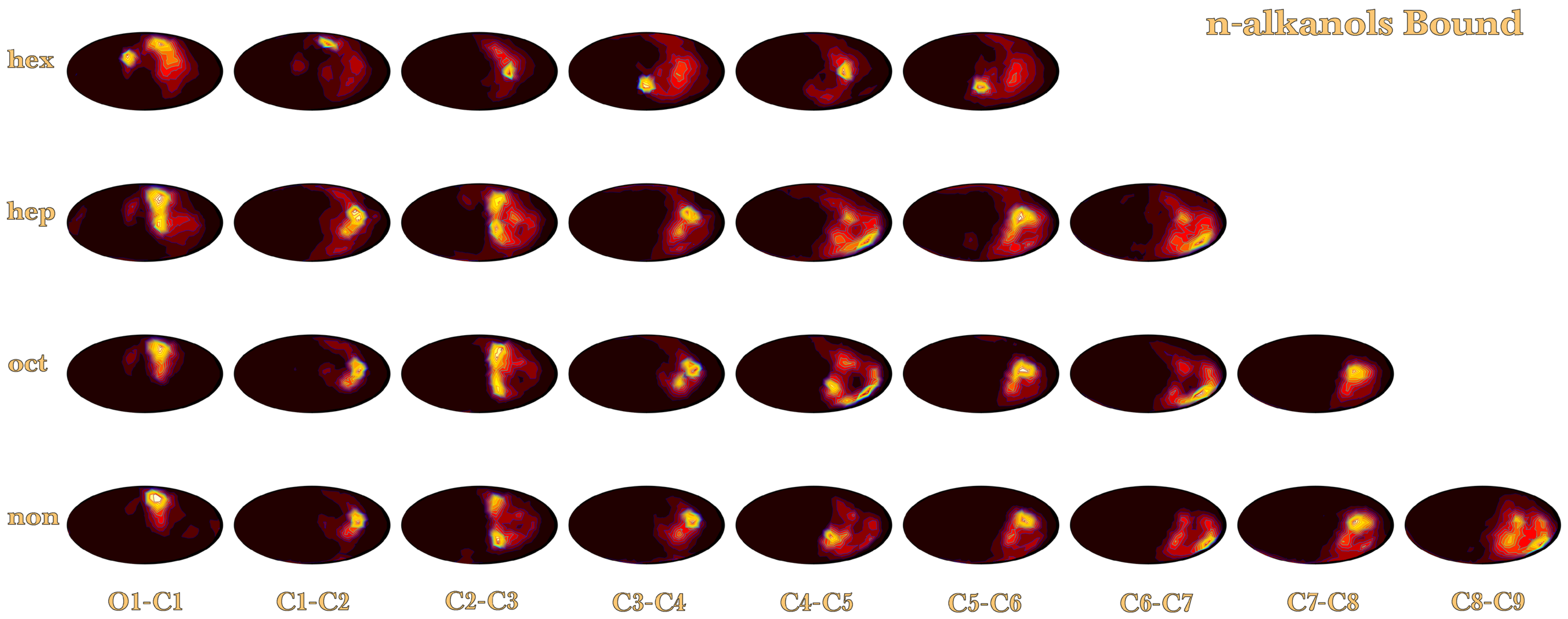


Fig.A2.2.3. Bound n-alkanol panel: Hammer projection showing bond vector rotational motion on the surface of a sphere.

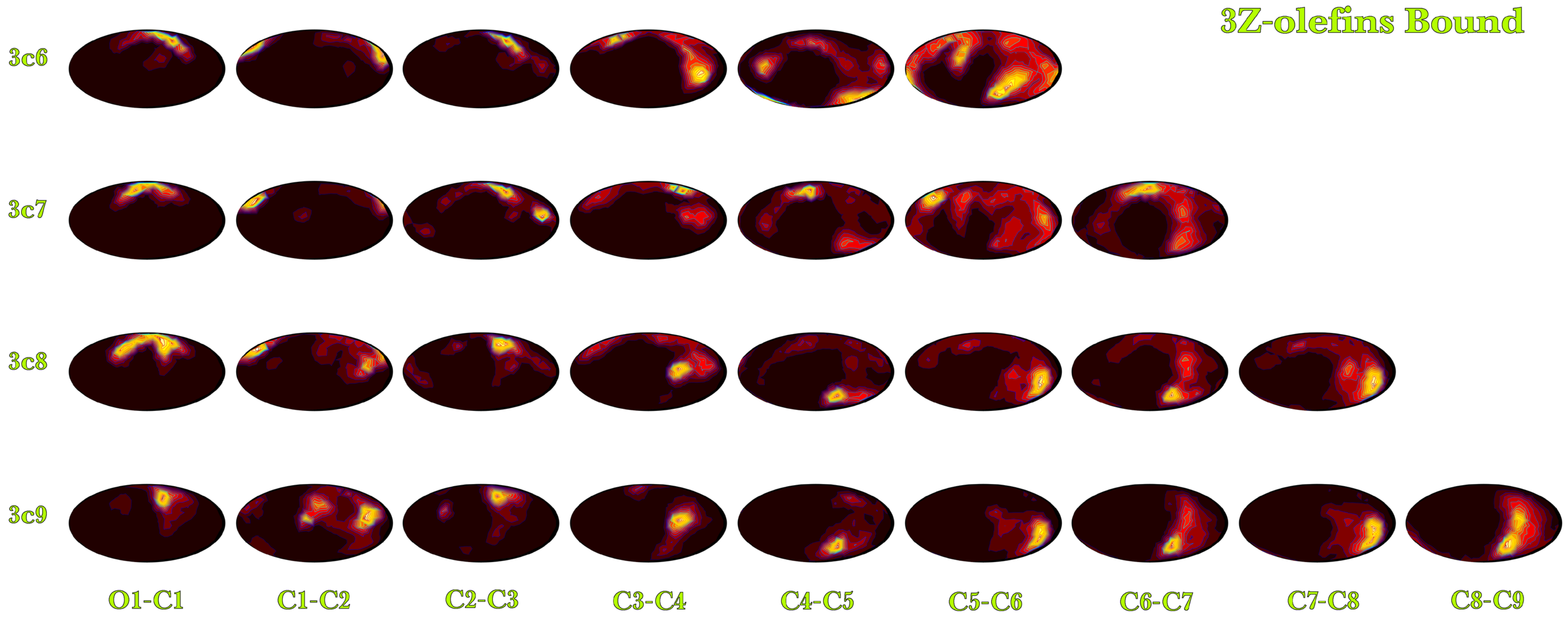


Fig.A2.2.4. Bound 3Z-olefin panel: Hammer projection showing bond vector rotational motion on the surface of a sphere.

Appendix A3

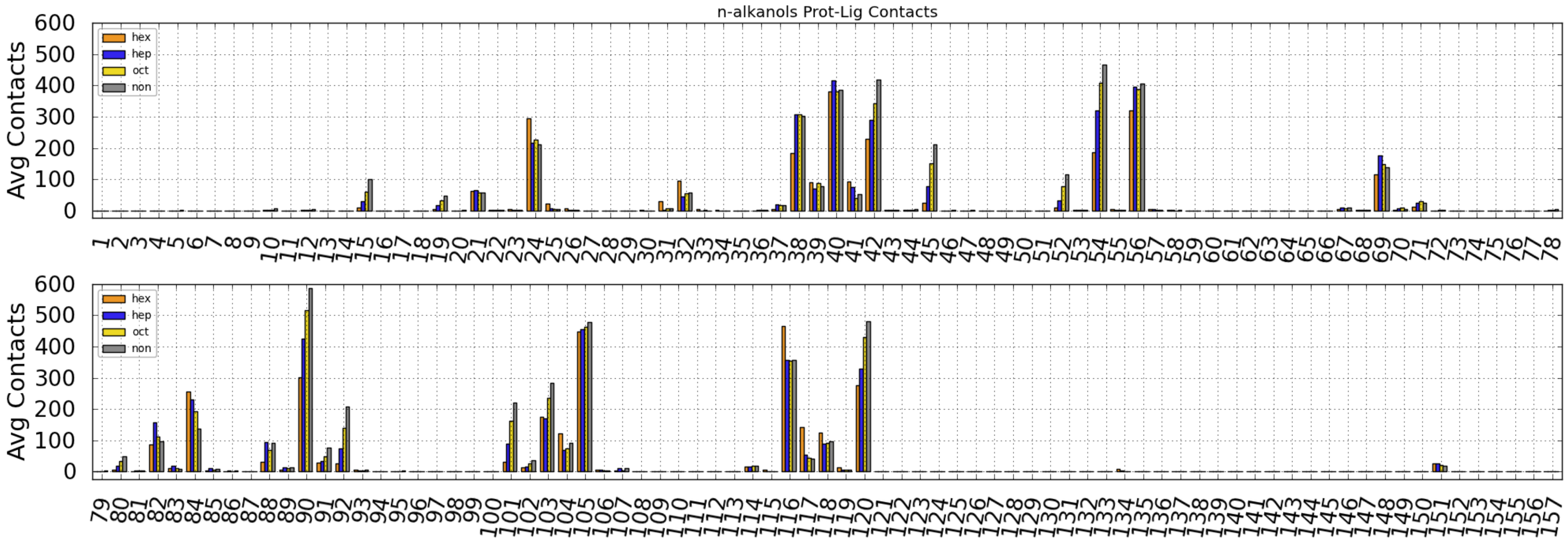


Fig.A3.1.1. n-alkanol panel: Average per-residue protein-ligand interfacial contacts obtained from 1.2 μs simulations. See §6.3.1.

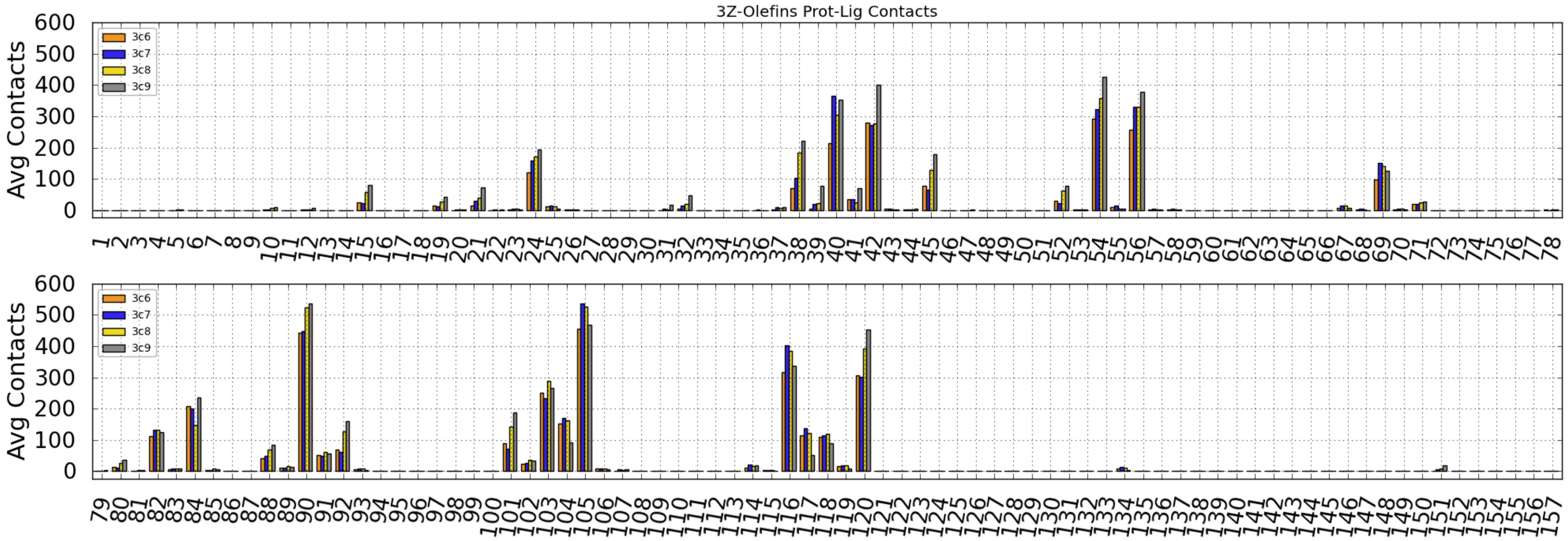


Fig.A3.1.2. 3Z-olefin panel: Average per-residue protein-ligand interfacial contacts obtained from 1.2 μs simulations. See §6.3.1.

Bibliography

(1) Drews, J. Drug Discovery: A Historical Perspective. *Science* 2000, *287* (5460), 1960–1964.

(2) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* 2004, *432* (7019), 855–861.

(3) Mandal, S.; Moudgil, M. ’nal; Mandal, S. K. Rational Drug Design. *Eur. J. Pharmacol.* 2009, *625* (1-3), 90–100.

(4) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discov.* 2003, *2* (5), 369–378.

(5) Waxman, D. J.; Strominger, J. L. Penicillin-Binding Proteins and the Mechanism of Action of Beta-Lactam Antibiotics. *Annu. Rev. Biochem.* 1983, *52* (1), 825–869.

(6) Chellappan, S.; Kiran Kumar Reddy, G. S.; Ali, A.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Kairys, V.; Fernandes, M. X.; Altman, M. D.; Tidor, B.; Rana, T. M.; Schiffer, C. A.; Gilson, M. K. Design of Mutation-Resistant HIV Protease Inhibitors with the Substrate Envelope Hypothesis. *Chem. Biol. Drug Des.* 2007, *69* (5), 298–313.

(7) Surleraux, D. L. N. G.; Tahri, A.; Verschueren, W. G.; Pille, G. M. E.; de Kock, H. A.; Jonckers, T. H. M.; Peeters, A.; De Meyer, S.; Azijn, H.; Pauwels, R.; de Bethune, M.-P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. T. P. Discovery and Selection of TMC114, a Next Generation HIV-1 Protease Inhibitor. *J. Med. Chem.* 2005, *48* (6), 1813–1822.

(8) Ghosh, A. K.; Chapsal, B. D.; Weber, I. T.; Mitsuya, H. Design of HIV Protease Inhibitors Targeting Protein Backbone: An Effective Strategy for Combating Drug Resistance. *Acc. Chem. Res.* 2008, *41* (1), 78–86.

(9) RCSB Protein Data Bank - RCSB PDB http://www.rcsb.org/pdb/home/home.do (accessed Oct 3, 2013).

(10) Whitesides, G. M.; Krishnamurthy, V. M. Designing Ligands to Bind Proteins. *Q. Rev. Biophys.* 2006, *38* (04), 385.

(11) Böhm, H.-J.; Klebe, G. What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? *Angew. Chem. Int. Ed. Engl.* 1996, *35* (22), 2588–2614.

(12) Mishra, K. P.; Ganju, L.; Sairam, M.; Banerjee, P. K.; Sawhney, R. C. A Review of High Throughput Technology for the Screening of Natural Products. *Biomed. Pharmacother.* 2008, *62* (2), 94–98.

(13) Zhu, Z.; Cuozzo, J. Review Article: High-Throughput Affinity-Based Technologies for Small-Molecule Drug Discovery. *J. Biomol. Screen.* 2009, *14* (10), 1157–1164.

(14) Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* 2000, *44* (1), 235–249.

(15) Oprea, T. I. Chemical Space Navigation in Lead Discovery. *Curr. Opin. Chem. Biol.* 2002, *6* (3), 384–389.

(16) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical

Space. *J. Comb. Chem.* 2001, *3* (2), 157–166.

(17) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Model.* 2003, *43* (1), 218–227.

(18) Earll, M. logppka Data http://www.raell.demon.co.uk/chem/logp/logppka.htm (accessed Feb 7, 2014).

(19) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* 1997, *23* (1), 3–25.

(20) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs. *J. Med. Chem.* 2003, *46* (7), 1250–1256.

(21) Price, N. C.; Price, N. C. *Principles and Problems in Physical Chemistry for Biochemists.*; Oxford University Press: Oxford; New York, 2001.

(22) Raffa, R. B. *Drug-Receptor Thermodynamics: Introduction and Applications*; Wiley: Chichester; New York, 2001.

(23) Smith, A. J. T.; Zhang, X.; Leach, A. G.; Houk, K. N. Beyond Picomolar Affinities: Quantitative Aspects of Noncovalent and Covalent Binding of Drugs to Proteins. *J. Med. Chem.* 2009, *52* (2), 225–233.

(24) Houk, K. N.; Leach, A. G.; Kim, S. P.; Zhang, X. Binding Affinities of Host–Guest, Protein–Ligand, and Protein–Transition-State Complexes. *Angew. Chem. Int. Ed.* 2003, *42* (40), 4872–4897.

(25) Vincent, J. P.; Lazdunski, M. Trypsin-Pancreatic Trypsin Inhibitor Association. Dynamics of the Interaction and Role of Disulfide Bridges. *Biochemistry (Mosc.)* 1972, *11* (16), 2967–2977.

(26) Tummino, P. J.; Copeland, R. A. Residence Time of Receptor–Ligand Complexes and Its Effect on Biological Function. *Biochemistry (Mosc.)* 2008, *47* (20), 5481–5492.

(27) Shuman, C. F.; Markgren, P.-O.; Hämäläinen, M.; Danielson, U. H. Elucidation of HIV-1 Protease Resistance by Characterization of Interaction Kinetics between Inhibitors and Enzyme Variants. *Antiviral Res.* 2003, *58* (3), 235–242.

(28) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug–target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* 2006, *5* (9), 730–739.

(29) Böhm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* 1994, *8* (3), 243–256.

(30) Williams, D. H.; Stephens, E.; O'Brien, D. P.; Zhou, M. Understanding Noncovalent Interactions: Ligand Binding Energy and Catalytic Efficiency from Ligand-Induced Reductions in Motion within Receptors and Enzymes. *Angew.*

Chem. Int. Ed. 2004, *43* (48), 6596–6616.

(31) Lambert, F. Entropy and the second law of thermodynamics http://2ndlaw.oxy.edu/entropy.html (accessed Feb 10, 2014).

(32) Holdgate, G. A.; Ward, W. H. Measurements of Binding Thermodynamics in Drug Discovery. *Drug Discov. Today* 2005, *10* (22), 1543–1550.

(33) Homans, S. W. Dynamics and Thermodynamics of Ligand–Protein Interactions. In *Bioactive Conformation I*; Peters, T., Ed.; Springer Berlin Heidelberg; Vol. 272, pp 51–82.

(34) Homans, S. W. Water, Water Everywhere – except Where It Matters? *Drug Discov. Today* 2007, *12* (13–14), 534–539.

(35) Ashworth, R. A.; Howe, G. B.; Mullins, M. E.; Rogers, T. N. Air-Water Partitioning Coefficients of Organics in Dilute Aqueous Solutions. *J. Hazard. Mater.* 1988, *18* (1), 25–36.

(36) Malham, R.; Johnstone, S.; Bingham, R. J.; Barratt, E.; Phillips, S. E. V.; Laughton, C. A.; Homans, S. W. Strong Solute—Solute Dispersive Interactions in a Protein—Ligand Complex. *J. Am. Chem. Soc.* 2005, *127* (48), 17061–17067.

(37) Sturtevant, J. M. Heat Capacity and Entropy Changes in Processes Involving Proteins. *Proc. Natl. Acad. Sci.* 1977, *74* (6), 2236–2240.

(38) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* 2005, *437* (7059), 640–647.

(39) Chandler, D. Hydrophobicity: Two Faces of Water. *Nature* 2002, *417* (6888), 491–491.

(40) Lazaridis, T. Solvent Size vs Cohesive Energy as the Origin of Hydrophobicity. *Acc. Chem. Res.* 2001, *34* (12), 931–937.

(41) Galamba, N. Water's Structure around Hydrophobic Solutes and the Iceberg Model. *J. Phys. Chem. B* 2013, *117* (7), 2153–2159.

(42) Dill, K. A.; Truskett, T. M.; Vlachy, V.; Hribar-Lee, B. Modeling Water, the Hydrophobic Effect, and Ion Solvation. *Annu. Rev. Biophys. Biomol. Struct.* 2005, *34* (1), 173–199.

(43) Cooper, A. Heat Capacity Effects in Protein Folding and Ligand Binding: A Re-Evaluation of the Role of Water in Biomolecular Thermodynamics. *Biophys. Chem.* 2005, *115* (2–3), 89–97.

(44) Loladze, V. V.; Ermolenko, D. N.; Makhatadze, G. I. Heat Capacity Changes upon Burial of Polar and Nonpolar Groups in Proteins. *Protein Sci. Publ. Protein Soc.* 2001, *10* (7), 1343–1352.

(45) Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The Thermodynamics of Protein–Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* 2008, *384* (4), 1002–1017.

(46) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The Maximal Affinity of Ligands. *Proc. Natl. Acad. Sci.* 1999, *96* (18), 9997–10002.

(47) Zhang, X.; Houk, K. N. Why Enzymes Are Proficient Catalysts: Beyond the Pauling Paradigm. *Acc. Chem. Res.* 2005, *38* (5), 379–385.

(48) Gilli, P.; Ferretti, V.; Gilli, G.; Borea, P. A. Enthalpy-Entropy Compensation in Drug-Receptor Binding. *J. Phys. Chem.* 1994, *98* (5), 1515–1518.

(49) Livnah, O.; Bayer, E. A.; Wilchek, M.; Sussman, J. L. Three-Dimensional Structures of Avidin and the Avidin-Biotin Complex. *Proc. Natl. Acad. Sci.* 1993, *90* (11), 5076–5080.

(50) Dunitz, J. D. Win Some, Lose Some: Enthalpy-Entropy Compensation in Weak Intermolecular Interactions. *Chem. Biol.* 1995, *2* (11), 709–712.

(51) Kim, K. H. Thermodynamic Aspects of Hydrophobicity and Biological QSAR. *J. Comput. Aided Mol. Des.* 2001, *15* (4), 367–380.

(52) Sharp, K. Entropy–enthalpy Compensation: Fact or Artifact? *Protein Sci.* 2001, *10* (3), 661–667.

(53) Cornish-Bowden, A. Enthalpy–entropy Compensation: A Phantom Phenomenon. *J. Biosci.* 2002, *27* (2), 121–126.

(54) SCORPIO http://scorpio.biophysics.ismb.lon.ac.uk/ (accessed Oct 6, 2013).

(55) Williams, D.; Westwell, M. Aspects of Weak Interactions. *Chem. Soc. Rev.* 1998, *27* (1), 57–64.

(56) Cooper, A.; Johnson, C. M.; Lakey, J. H.; Nöllmann, M. Heat Does Not Come in Different Colours: Entropy–enthalpy Compensation, Free Energy Windows, Quantum Confinement, Pressure Perturbation Calorimetry, Solvation and the Multiple Causes of Heat Capacity Effects in Biomolecular Interactions. *Biophys. Chem.* 2001, *93* (2), 215–230.

(57) Krug, R. R.; Hunter, W. G.; Grieger, R. A. Enthalpy-Entropy Compensation. 1. Some Fundamental Statistical Problems Associated with the Analysis of Van't Hoff and Arrhenius Data. *J. Phys. Chem.* 1976, *80* (21), 2335–2341.

(58) Freire, E. Do Enthalpy and Entropy Distinguish First in Class from Best in Class? *Drug Discov. Today* 2008, *13* (19-20), 869–874.

(59) Fersht, A. R.; Shi, J. P.; Knill-Jones, J.; Lowe, D. M.; Wilkinson, A. J.; Blow, D. M.; Brick, P.; Carter, P.; Waye, M. M. Y.; Winter, G. Hydrogen Bonding and Biological Specificity Analysed by Protein Engineering. *Nat. Lond.* 1985, *314*, 235–238.

(60) Lafont, V.; Armstrong, A. A.; Ohtaka, H.; Kiso, Y.; Mario Amzel, L.; Freire, E. Compensating Enthalpic and Entropic Changes Hinder Binding Affinity Optimization. *Chem. Biol. Drug Des.* 2007, *69* (6), 413–422.

(61) Cihlar, T.; He, G.-X.; Liu, X.; Chen, J. M.; Hatada, M.; Swaminathan, S.; McDermott, M. J.; Yang, Z.-Y.; Mulato, A. S.; Chen, X.; Leavitt, S. A.; Stray, K. M.; Lee, W. A. Suppression of HIV-1 Protease Inhibitor Resistance by Phosphonate-Mediated Solvent Anchoring. *J. Mol. Biol.* 2006, *363* (3), 635–647.

(62) Krishnamurthy, V. M.; Kaufman, G. K.; Urbach, A. R.; Gitlin, I.; Gudiksen,

K. L.; Weibel, D. B.; Whitesides, G. M. Carbonic Anhydrase as a Model for Biophysical and Physical-Organic Studies of Proteins and Protein–Ligand Binding. *Chem. Rev.* 2008, *108* (3), 946–1051.

(63) Krishnamurthy, V. M.; Bohall, B. R.; Semetey, V.; Whitesides, G. M. The Paradoxical Thermodynamic Basis for the Interaction of Ethylene Glycol, Glycine, and Sarcosine Chains with Bovine Carbonic Anhydrase II: An Unexpected Manifestation of Enthalpy/entropy Compensation. *J. Am. Chem. Soc.* 2006, *128* (17), 5802–5812.

(64) Stöckmann, H.; Bronowska, A.; Syme, N. R.; Thompson, G. S.; Kalverda, A. P.; Warriner, S. L.; Homans, S. W. Residual Ligand Entropy in the Binding of P-Substituted Benzenesulfonamide Ligands to Bovine Carbonic Anhydrase II. *J. Am. Chem. Soc.* 2008, *130* (37), 12420–12426.

(65) Donovan, J. W.; Ross, K. D. Increase in the Stability of Avidin Produced by Binding of Biotin. Differential Scanning Calorimetric Study of Denaturation by Heat. *Biochemistry (Mosc.)* 1973, *12* (3), 512–517.

(66) Green, N. M. Thermodynamics of the Binding of Biotin and Some Analogues by Avidin. *Biochem J* 1966, *101*, 774–780.

(67) Chilkoti, A.; Stayton, P. S. Molecular Origins of the Slow Streptavidin-Biotin Dissociation Kinetics. *J. Am. Chem. Soc.* 1995, *117* (43), 10622–10628.

(68) Lagona, J.; Mukhopadhyay, P.; Chakrabarti, S.; Isaacs, L. The Cucurbit[n]uril Family. *Angew. Chem. Int. Ed.* 2005, *44* (31), 4844–4870.

(69) Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L. A Synthetic Host-Guest System Achieves Avidin-Biotin Affinity by Overcoming Enthalpy–entropy Compensation. *Proc. Natl. Acad. Sci.* 2007, *104* (52), 20737–20742.

(70) Sharrow, S. D.; Vaughn, J. L.; Žídek, L.; Novotny, M. V.; Stone, M. J. Pheromone Binding by Polymorphic Mouse Major Urinary Proteins. *Protein Sci.* 2009, *11* (9), 2247–2256.

(71) Perez-Miller, S.; Zou, Q.; Novotny, M. V.; Hurley, T. D. High Resolution X-Ray Structures of Mouse Major Urinary Protein Nasal Isoform in Complex with Pheromones. *Protein Sci. Publ. Protein Soc.* 2010, *19* (8), 1469–1479.

(72) Timm, D. E.; Baker, L. J.; Mueller, H.; Zidek, L.; Novotny, M. V. Structural Basis of Pheromone Binding to Mouse Major Urinary Protein (MUP-I). *Protein Sci.* 2001, *10* (5), 997–1004.

(73) Abascal, J. L. F.; García Fernández, R.; MacDowell, L. G.; Sanz, E.; Vega, C. Ice: A Fruitful Source of Information about Liquid Water. *J. Mol. Liq.* 2007, *136* (3), 214–220.

(74) Kwak, J.; Grigsby, C. C.; Rizki, M. M.; Preti, G.; Köksal, M.; Josue, J.; Yamazaki, K.; Beauchamp, G. K. Differential Binding between Volatile Ligands and Major Urinary Proteins due to Genetic Variation in Mice. *Physiol. Behav.* 2012, *107* (1),

112–120.

(75) Hoffmann, F.; Musolf, K.; Penn, D. J. Freezing Urine Reduces Its Efficacy for Eliciting Ultrasonic Vocalizations from Male Mice. *Physiol. Behav.* 2009, *96* (4–5), 602–605.

(76) Hurst, J. L.; Beynon, R. J. Scent Wars: The Chemobiology of Competitive Signalling in Mice. *BioEssays* 2004, *26* (12), 1288–1298.

(77) Darwish Marie, A.; Veggerby, C.; Robertson, D. H. L.; Gaskell, S. J.; Hubbard, S. J.; Martinsen, L.; Hurst, J. L.; Beynon, R. J. Effect of Polymorphisms on Ligand Binding by Mouse Major Urinary Proteins. *Protein Sci.* 2001, *10* (2), 411–417.

(78) Beynon, R. J.; Hurst, J. L. Urinary Proteins and the Modulation of Chemical Scents in Mice and Rats. *Peptides* 2004, *25* (9), 1553–1563.

(79) Kwak, J.; Grigsby, C. C.; Preti, G.; Rizki, M. M.; Yamazaki, K.; Beauchamp, G. K. Changes in Volatile Compounds of Mouse Urine as It Ages: Their Interactions with Water and Urinary Proteins. *Physiol. Behav.* 2013, *120*, 211–219.

(80) Fukuhara, A.; Nakajima, H.; Miyamoto, Y.; Inoue, K.; Kume, S.; Lee, Y.-H.; Noda, M.; Uchiyama, S.; Shimamoto, S.; Nishimura, S.; Ohkubo, T.; Goto, Y.; Takeuchi, T.; Inui, T. Drug Delivery System for Poorly Water-Soluble Compounds Using Lipocalin-Type Prostaglandin D Synthase. *J. Controlled Release* 2012, *159* (1), 143–150.

(81) Gasymov, O. K.; Abduragimov, A. R.; Gasimov, E. O.; Yusifov, T. N.; Dooley, A. N.; Glasgow, B. J. Tear Lipocalin: Potential for Selective Delivery of Rifampin. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 2004, *1688* (2), 102–111.

(82) Schlehuber, S.; Skerra, A. Anticalins: Promising Tools for Clinical Diagnostics.

(83) Schlehuber, S.; Skerra, A. Lipocalins in Drug Discovery: From Natural Ligand-Binding Proteins to "anticalins." *Drug Discov. Today* 2005, *10* (1), 23–33.

(84) Gebauer, M.; Skerra, A. Chapter Seven - Anticalins: Small Engineered Binding Proteins Based on the Lipocalin Scaffold. In *Methods in Enzymology*; K. Dane Wittrup and Gregory L. Verdine, Ed.; Protein Engineering for Therapeutics, Part B; Academic Press, 2012; Vol. Volume 503, pp 157–188.

(85) Stivala, A.; Wybrow, M.; Wirth, A.; Whisstock, J. C.; Stuckey, P. J. Automatic Generation of Protein Structure Cartoons with Pro-Origami. *Bioinformatics* 2011, *27* (23), 3315–3316.

(86) Flower, D. R.; North, A. C. T.; Sansom, C. E. The Lipocalin Protein Family: Structural and Sequence Overview. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* 2000, *1482* (1–2), 9–24.

(87) Grzyb, J.; Latowski, D.; Strzałka, K. Lipocalins – a Family Portrait. *J. Plant Physiol.* 2006, *163* (9), 895–915.

(88) Patel, R. C.; Lange, D.; McConathy, W. J.; Patel, Y. C.; Patel, S. C. Probing the Structure of the Ligand Binding Cavity of Lipocalins by Fluorescence

Spectroscopy. *Protein Eng.* 1997, *10* (6), 621–625.

(89) Bingham, R. J.; Findlay, J. B. C.; Hsieh, S.-Y.; Kalverda, A. P.; Kjellberg, A.; Perazzolo, C.; Phillips, S. E. V.; Seshadri, K.; Trinh, C. H.; Turnbull, W. B.; Bodenhausen, G.; Homans, S. W. Thermodynamics of Binding of 2-Methoxy-3-Isopropylpyrazine and 2-Methoxy-3-Isobutylpyrazine to the Major Urinary Protein. *J. Am. Chem. Soc.* 2004, *126* (6), 1675–1681.

(90) Barratt, E.; Bingham, R. J.; Warner, D. J.; Laughton, C. A.; Phillips, S. E. V.; Homans, S. W. Van Der Waals Interactions Dominate Ligand—Protein Association in a Protein Binding Site Occluded from Solvent Water. *J. Am. Chem. Soc.* 2005, *127* (33), 11827–11834.

(91) Roy, J.; Laughton, C. A. Long-Timescale Molecular-Dynamics Simulations of the Major Urinary Protein Provide Atomistic Interpretations of the Unusual Thermodynamics of Ligand Binding. *Biophys. J.* 2010, *99* (1), 218–226.

(92) Berg, J. M.; Tymoczko, J. L.; Stryer, L.; Clarke, N. D. *Biochemistry*; Freeman and Company: New York, 2002.

(93) Ross, P. D.; Subramanian, S. Thermodynamics of Protein Association Reactions: Forces Contributing to Stability. *Biochemistry (Mosc.)* 1981, *20* (11), 3096–3102.

(94) Plyasunov, A. V.; Shock, E. L. Thermodynamic Functions of Hydration of Hydrocarbons at 298.15 K and 0.1 MPa. *Geochim. Cosmochim. Acta* 2000, *64* (3), 439–468.

(95) Israelachvili, J. N. *Intermolecular and Surface Forces*; Academic Press: Burlington, MA, 2011.

(96) Yang, D.; Kay, L. E. Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-Derived Order Parameters: Application to Protein Folding. *J. Mol. Biol.* 1996, *263* (2), 369–382.

(97) Homans, S. W. Probing the Binding Entropy of Ligand–Protein Interactions by NMR. *ChemBioChem* 2005, *6* (9), 1585–1591.

(98) Reetz, M. T.; Puls, M.; Carballeira, J. D.; Vogel, A.; Jaeger, K.-E.; Eggert, T.; Thiel, W.; Bocola, M.; Otte, N. Learning from Directed Evolution: Further Lessons from Theoretical Investigations into Cooperative Mutations in Lipase Enantioselectivity. *ChemBioChem* 2007, *8* (1), 106–112.

(99) Ong, J. L.; Loakes, D.; Jaroslawski, S.; Too, K.; Holliger, P. Directed Evolution of DNA Polymerase, RNA Polymerase and Reverse Transcriptase Activity in a Single Polypeptide. *J. Mol. Biol.* 2006, *361* (3), 537–550.

(100) Zídek, L.; Novotny, M. V.; Stone, M. J. Increased Protein Backbone Conformational Entropy upon Hydrophobic Ligand Binding. *Nat. Struct. Mol. Biol.* 1999, *6* (12), 1118–1121.

(101) Macek, P.; Novák, P.; K ížová, H.; Žídek, L.; Sklená , V. Molecular Dynamics Study of Major Urinary Protein–pheromone Interactions: A Structural Model for Ligand-Induced Flexibility Increase. *FEBS Lett.* 2006, *580* (2), 682–684.

(102) Macek, P.; Novák, P.; Žídek, L.; Sklená , V. Backbone Motions of Free and Pheromone-Bound Major Urinary Protein I Studied by Molecular Dynamics Simulation. *J. Phys. Chem. B* 2007, *111* (20), 5731–5739.

(103) Leach, A. R. *Molecular Modelling: Principles and Applications*; Prentice Hall: Harlow [etc.], 2001.

(104) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; Wiley: New York, 2001.

(105) Tuckerman, M. E.; Martyna, G. J. Understanding Modern Molecular Dynamics: Techniques and Applications. *J. Phys. Chem. B* 2000, *104* (2), 159–178.

(106) Kühne, T. D. Ab-Initio Molecular Dynamics. *ArXiv Prepr. ArXiv12015945* 2012.

(107) Söderhjelm, P.; Kongsted, J.; Ryde, U. Ligand Affinities Estimated by Quantum Chemical Calculations. *J. Chem. Theory Comput.* 2010, *6* (5), 1726–1737.

(108) Reddy, M. R.; Erion, M. D. Relative Solvation Free Energies Calculated Using an Ab Initio QM/MM-Based Free Energy Perturbation Method: Dependence of Results on Simulation Length. *J. Comput. Aided Mol. Des.* 2009, *23* (12), 837–843.

(109) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, 2012.

(110) Rappé, A. K.; Casewit, Carla J. *Molecular Mechanics across Chemistry*; University Science Books: Sausalito, Calif., 1997.

(111) Somoza, Ma. File:Morse-Potential.png. *Wikipedia, the free encyclopedia.*

(112) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* 2007, *36* (1), 21–42.

(113) Lim, T.; UX, A. The Relationship between Lennard-Jones (12-6) and Morse Potential. *Mol. Simul.* 2003, 615–617.

(114) Guvench, O.; MacKerell, A. D. Computational Evaluation of Protein–small Molecule Binding. *Curr. Opin. Struct. Biol.* 2009, *19* (1), 56–61.

(115) Vega, C.; Abascal, J. L. F.; Conde, M. M.; Aragones, J. L. What Ice Can Teach Us about Water Interactions: A Critical Comparison of the Performance of Different Water Models. *Faraday Discuss.* 2008, *141* (0), 251–276.

(116) File:Water models.svg - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/File:Water_models.svg (accessed Oct 6, 2013).

(117) Lopes, P. E. M.; Roux, B.; MacKerell, A. D. Molecular Modeling and Dynamics Studies with Explicit Inclusion of Electronic Polarizability: Theory and Applications. *Theor. Chem. Acc.* 2009, *124* (1-2), 11–28.

(118) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* 2010, *114* (8), 2549–2564.

(119) Aleksandrov, A.; Thompson, D.; Simonson, T. Alchemical Free Energy Simulations for Biological Complexes: Powerful but Temperamental …. *J. Mol.*

Recognit. 2009, n/a – n/a.

(120) Amaro, R. E.; Li, W. W. Emerging Methods for Ensemble-Based Virtual Screening. *Curr. Top. Med. Chem.* 2010, *10* (1), 3.

(121) Case, D. A.; Darden, T.; Cheatham III, T. E.; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R.; Zhang, W. *Amber 10 Users' Manual*; 2008.

(122) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* 1977, *23* (3), 327–341.

(123) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* 2008, *51* (7), 91–97.

(124) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J. Millisecond-Scale Molecular Dynamics Simulations on Anton. In *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*; IEEE, 2009; pp 1–11.

(125) NRBSC | Anton Request for Proposals http://www.nrbsc.org/anton_rfp/ (accessed Feb 3, 2014).

(126) Kirk, D.; Hwu, W. *Programming Massively Parallel Processors: A Hands-on Approach*; Morgan Kaufmann Elsevier: Burlington, Massachusetts, 2010.

(127) Uncanny valley - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Uncanny_valley (accessed Feb 3, 2014).

(128) Weblog - Arkanis Development http://arkanis.de/ (accessed Feb 3, 2014).

(129) File:Mori Uncanny Valley.svg - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/File:Mori_Uncanny_Valley.svg (accessed Feb 3, 2014).

(130) Double-precision floating-point format - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Double-precision_floating-point_format (accessed Feb 3, 2014).

(131) Single-precision floating-point format - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Single-precision_floating-point_format (accessed Feb 3, 2014).

(132) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* 2012, *8* (5), 1542–1555.

(133) Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without compromise–A Mixed Precision Model for GPU Accelerated Molecular Dynamics Simulations. *Comput. Phys. Commun.* 2013, *184* (2), 374–380.

(134) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular

Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* 2004, *120* (24), 11919.

(135) Pierce, L. C. T.; Salomon-Ferrer, R.; Augusto F. de Oliveira, C.; McCammon, J. A.; Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* 2012, *8* (9), 2997–3002.

(136) Barratt, E.; Bronowska, A.; Vondrášek, J.; erný, J.; Bingham, R.; Phillips, S.; Homans, S. W. Thermodynamic Penalty Arising from Burial of a Ligand Polar Group Within a Hydrophobic Pocket of a Protein Receptor. *J. Mol. Biol.* 2006, *362* (5), 994–1003.

(137) Steinbrecher, T. TUTORIAL A9: Thermodynamic Integration using soft core potentials. http://ambermd.org/tutorials/advanced/tutorial9/ (accessed Oct 7, 2013).

(138) AMBER TI Tutorials - Rizzo Lab http://ringo.ams.sunysb.edu/index.php/AMBER_TI_Tutorials (accessed Oct 7, 2013).

(139) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* 1993, *93* (7), 2395–2417.

(140) Grubmüller, H.; Ehrenhofer, N.; Tavan, P. Conformational Dynamics of Proteins: Beyond the Nanosecond Time Scale. In *Nonlinear Excitations in Biomolecules*; Peyrard, M., Ed.; Centre de Physique des Houches; Springer Berlin Heidelberg, 1995; pp 231–240.

(141) Blondel, A. Ensemble Variance in Free Energy Calculations by Thermodynamic Integration: Theory, Optimal "alchemical" Path, and Practical Solutions. *J. Comput. Chem.* 2004, *25* (7), 985–993.

(142) Hummer, G. Fast-Growth Thermodynamic Integration: Error and Efficiency Analysis. *J. Chem. Phys.* 2001, *114* (17), 7330.

(143) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear Scaling Schemes for Lennard-Jones Interactions in Free Energy Calculations. *J. Chem. Phys.* 2007, *127* (21), 214108.

(144) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* 1994, *222* (6), 529–539.

(145) Pitera, J. W.; van Gunsteren, W. F. A Comparison of Non-Bonded Scaling Approaches for Free Energy Calculations. *Mol. Simul.* 2002, *28* (1-2), 45–65.

(146) Wyczalkowski, M. A.; Vitalis, A.; Pappu, R. V. New Estimators for Calculating Solvation Entropy and Enthalpy and Comparative Assessments of Their Accuracy and Precision. *J. Phys. Chem. B* 2010, *114* (24), 8166–8180.

(147) Rodinger, T.; Pomès, R. Enhancing the Accuracy, the Efficiency and the Scope of Free Energy Simulations. *Curr. Opin. Struct. Biol.* 2005, *15* (2), 164–170.

(148) Meirovitch, H. Recent Developments in Methodologies for Calculating the

Entropy and Free Energy of Biological Systems by Computer Simulation. *Curr. Opin. Struct. Biol.* 2007, *17* (2), 181–186.

(149) Dupradeau, F. Y.; Pigache, A.; Zaffran, T.; Cieplak, P. RED Version 2.0 User's Manual and Tutorial. *Univ. Picardie Jules Verne Amiens Fr.* 2005.

(150) Macke, T. J.; Svrcek-Seiler, W. A.; Brown, R. A.; Kolossváry, I.; Bomble, Y. J.; Case, D. A.; Zhang, W.; Hou, T.; Schafmeister, C.; Ross, W. S. *AmberTools ver1.3 Users' Manual*; Ver, 2010.

(151) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* 2003, *24* (16), 1999–2012.

(152) Cerutti, D. S.; Duke, R.; Freddolino, P. L.; Fan, H.; Lybrand, T. P. A Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics. *J. Chem. Theory Comput.* 2008, *4* (10), 1669–1680.

(153) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* 2007, *3* (1), 26–41.

(154) Rapaport, D. C. *The Art of Molecular Dynamics Simulation*; Cambridge University Press: Cambridge, UK; New York, NY, 2004.

(155) Lawrenz, M.; Baron, R.; McCammon, J. A. Independent-Trajectories Thermodynamic-Integration Free-Energy Changes for Biomolecular Systems: Determinants of H5N1 Avian Influenza Virus Neuraminidase Inhibition by Peramivir. *J. Chem. Theory Comput.* 2009, *5* (4), 1106–1116.

(156) Kubo, M. M.; Gallicchio, E.; Levy, R. M. Thermodynamic Decomposition of Hydration Free Energies by Computer Simulation: Application to Amines, Oxides, and Sulfides. *J. Phys. Chem. B* 1997, *101* (49), 10527–10534.

(157) Smith, D. E.; Haymet, A. D. J. Free Energy, Entropy, and Internal Energy of Hydrophobic Interactions: Computer Simulations. *J. Chem. Phys.* 1993, *98*, 6445.

(158) Bevington, P. R.; Robinson, D. Keith. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill: Boston, 2003.

(159) Claxton, D. P.; Zou, P.; Mchaourab, H. S. Structure and Orientation of T4 Lysozyme Bound to the Small Heat Shock Protein -Crystallin. *J. Mol. Biol.* 2008, *375* (4), 1026–1039.

(160) Chia-en, A. C.; McLaughlin, W. A.; Baron, R.; Wang, W.; McCammon, J. A. Entropic Contributions and the Influence of the Hydrophobic Environment in Promiscuous Protein–protein Association. *Proc. Natl. Acad. Sci.* 2008, *105* (21), 7456–7461.

(161) Gao, M.; Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput Biol* 2013, *9* (10), e1003302.

(162) Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* 2013, *8* (6), e65894.

(163) Shimokhina, N.; Bronowska, A.; Homans, S. W. Contribution of Ligand Desolvation to Binding Thermodynamics in a Ligand–Protein Interaction. *Angew. Chem. Int. Ed.* 2006, *45* (38), 6374–6376.

(164) Irudayam, S. J.; Henchman, R. H. Entropic Cost of Protein—Ligand Binding and Its Dependence on the Entropy in Solution. *J. Phys. Chem. B* 2009, *113* (17), 5871–5884.

(165) Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration. *J. Chem. Theory Comput.* 2010, *6* (4), 1018–1027.

(166) Gao, C.; Park, M.-S.; Stern, H. A. Accounting for Ligand Conformational Restriction in Calculations of Protein-Ligand Binding Affinities. *Biophys. J.* 2010, *98* (5), 901–910.

(167) Raha, K.; Merz, K. M. Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein—Ligand Complexes. *J. Med. Chem.* 2005, *48* (14), 4558–4575.

(168) Pitzer, K. S.; Gwinn, W. D. Energy Levels and Thermodynamic Functions for Molecules with Internal Rotation I. Rigid Frame with Attached Tops. *J. Chem. Phys.* 2004, *10* (7), 428–440.

(169) Carver, J. P. Oligosaccharides: How Can Flexible Molecules Act as Signals? *Pure Appl. Chem.* 1993, *65*, 763–763.

(170) Searle, M. S.; Williams, D. H. The Cost of Conformational Order: Entropy Changes in Molecular Associations. *J. Am. Chem. Soc.* 1992, *114* (27), 10690–10697.

(171) Page, M. I.; Jencks, W. P. Entropic Contributions to Rate Accelerations in Enzymic and Intramolecular Reactions and the Chelate Effect. *Proc. Natl. Acad. Sci. U. S. A.* 1971, *68* (8), 1678–1683.

(172) Lundquist, J. J.; Toone, E. J. The Cluster Glycoside Effect. *Chem. Rev.* 2002, *102* (2), 555–578.

(173) Chia-en, A. C.; Chen, W.; Gilson, M. K. Ligand Configurational Entropy and Protein Binding. *Proc. Natl. Acad. Sci.* 2007, *104* (5), 1534–1539.

(174) Cram, D. J. Preorganization–from Solvents to Spherands. *Angew. Chem. Int. Ed. Engl.* 1986, *25* (12), 1039–1057.

(175) Khan, A. R.; Parrish, J. C.; Fraser, M. E.; Smith, W. W.; Bartlett, P. A.; James, M. N. G. Lowering the Entropic Barrier for Binding Conformationally Flexible Inhibitors to Enzymes†,‡. *Biochemistry (Mosc.)* 1998, *37* (48), 16839–16845.

(176) Morgan, B. P.; Holland, D. R.; Matthews, B. W.; Bartlett, P. A. Structure-Based

Design of an Inhibitor of the Zinc Peptidase Thermolysin. *J. Am. Chem. Soc.* 1994, *116* (8), 3251–3260.

(177) Benfield, A. P.; Teresk, M. G.; Plake, H. R.; DeLorbe, J. E.; Millspaugh, L. E.; Martin, S. F. Ligand Preorganization May Be Accompanied by Entropic Penalties in Protein–Ligand Interactions. *Angew. Chem. Int. Ed.* 2006, *45* (41), 6830–6835.

(178) Martin, S. F. Preorganization in Biological Systems: Are Conformational Constraints Worth the Energy? *Pure Appl. Chem.* 2007, *79* (2), 193–200.

(179) DeLorbe, J. E.; Clements, J. H.; Teresk, M. G.; Benfield, A. P.; Plake, H. R.; Millspaugh, L. E.; Martin, S. F. Thermodynamic and Structural Effects of Conformational Constraints in Protein—Ligand Interactions. Entropic Paradoxy Associated with Ligand Preorganization. *J. Am. Chem. Soc.* 2009, *131* (46), 16758–16770.

(180) Malham, R. W. *Dynamics and Thermodynamics of Protein-Ligand Interactions*; University of Leeds, 2012.

(181) Schafer, H.; Smith, L. J.; Mark, A. E.; van Gunsteren, W. F. Entropy Calculations on the Molten Globule State of a Protein: Side-Chain Entropies of Alpha-Lactalbumin. *Proteins Struct. Funct. Genet.* 2002, *46* (2), 215–224.

(182) Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K. Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation. *J. Chem. Phys.* 2007, *127* (2), 024107.

(183) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* 2008, *29* (10), 1605–1614.

(184) Meirovitch, H. Methods for Calculating the Absolute Entropy and Free Energy of Biological Systems Based on Ideas from Polymer Physics. *J. Mol. Recognit.* 2009.

(185) Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. Estimating Entropies from Molecular Dynamics Simulations. *J. Chem. Phys.* 2004, *120* (6), 2652.

(186) Brady, G. P.; Sharp, K. A. Entropy in Protein Folding and in Protein—protein Interactions. *Curr. Opin. Struct. Biol.* 1997, 7 (2), 215–221.

(187) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-Diffusive Systems. *Mol. Phys.* 1984, *51* (4), 1011–1028.

(188) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid–Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* 2011, 7 (8), 2638–2653.

(189) Baron, R. *Computational Drug Discovery and Design*; Springer, 2012.

(190) Carlsson, J.; Aqvist, J. Calculations of Solute and Solvent Entropies from

Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* 2006, *8* (46), 5385.

(191) G , N.; Scheraga, H. A. On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation. *Macromolecules* 1976, *9* (4), 535–542.

(192) Li, D.-W.; Brüschweiler, R. In Silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* 2009, *102* (11).

(193) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational Entropy in Molecular Recognition by Proteins. *Nature* 2007, *448* (7151), 325–329.

(194) Marlow, M. S.; Dogan, J.; Frederick, K. K.; Valentine, K. G.; Wand, A. J. The Role of Conformational Entropy in Molecular Recognition by Calmodulin. *Nat. Chem. Biol.* 2010, *6* (5), 352–358.

(195) Dolenc, J.; Baron, R.; Oostenbrink, C.; Koller, J.; van Gunsteren, W. F. Configurational Entropy Change of Netropsin and Distamycin upon DNA Minor-Groove Binding. *Biophys. J.* 2006, *91* (4), 1460–1470.

(196) Mammen, M.; Shakhnovich, E. I.; Whitesides, G. M. Using a Convenient, Quantitative Model for Torsional Entropy to Establish Qualitative Trends for Molecular Processes That Restrict Conformational Freedom. *J. Org. Chem.* 1998, *63* (10), 3168–3175.

(197) Cheluvaraja, S.; Meirovitch, H. Simulation Method for Calculating the Entropy and Free Energy of Peptides and Proteins. *Proc. Natl. Acad. Sci. U. S. A.* 2004, *101* (25), 9241–9246.

(198) General, I. J.; Dragomirova, R.; Meirovitch, H. New Method for Calculating the Absolute Free Energy of Binding: The Effect of a Mobile Loop on the Avidin/Biotin Complex. *J. Phys. Chem. B* 2011, *115* (1), 168–175.

(199) General, I. J.; Meirovitch, H. Relative Stability of the Open and Closed Conformations of the Active Site Loop of Streptavidin. *J. Chem. Phys.* 2011, *134* (2), 025104.

(200) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* 2001, *115* (14), 6289.

(201) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* 1981, *14* (2), 325–332.

(202) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. 1993, *2* (6), 617–621.

(203) Chang, C.-E.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Theory Comput.* 2005, *1* (5), 1017–1028.

(204) Killian, B. *Manual Killian, B. J. (2009). Algorithm for Computing Configurational ENTropy from Molecular Mechanics (beta)*; University of Maryland Biotechnology Institute, 2005.

(205) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-

Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. *J. Comput. Chem.* 2007, *28* (3), 655–668.

(206) Hensen, U.; Lange, O. F.; Grubmüller, H. Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach. *PLoS ONE* 2010, *5* (2), e9179.

(207) Hensen, U.; Grubmüller, H.; Lange, O. Adaptive Anisotropic Kernels for Nonparametric Estimation of Absolute Configurational Entropies in High-Dimensional Configuration Spaces. *Phys. Rev. E* 2009, *80* (1).

(208) Harrison, B. A.; Gierasch, T. M.; Neilan, C.; Pasternak, G. W.; Verdine, G. L. High-Affinity Mu Opioid Receptor Ligands Discovered by the Screening of an Exhaustively Stereodiversified Library of 1,5-Enediols. *J. Am. Chem. Soc.* 2002, *124* (45), 13352–13353.

(209) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* 2010, *12* (28), 7821.

(210) R.E.DD.B.: code F-90 http://q4md-forcefieldtools.org/REDDB/projects/F-90/ (accessed Apr 1, 2014).

(211) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers: Charge Derivation for DNA, RNA, and Proteins. *J. Comput. Chem.* 1995, *16* (11), 1357–1377.

(212) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. Application of RESP Charges to Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *J. Am. Chem. Soc.* 1993, *115* (21), 9620–9631.

(213) Kieseritzky, G.; Knapp, E.-W. Optimizing pKA Computation in Proteins with pH Adapted Conformations. *Proteins Struct. Funct. Bioinforma.* 2007, *71* (3), 1335–1348.

(214) Trylska, J. Computational Modelling of Protonation Equilibria and Reaction Mechanism of HIV-1 Protease.

(215) Trylska, J.; Ba\la, P.; Geller, M.; Grochowski, P. Molecular Dynamics Simulations of the First Steps of the Reaction Catalyzed by HIV-1 Protease. *Biophys. J.* 2002, *83* (2), 794–807.

(216) Won, H.; Kim, J. R.; Ko, K.; Won, Y. The Pka Shift of the Catalytic Aspartyl Dyad in the HIV-1 Protease Complexed with Hydroxyethylene Inhibitors. *Bull.-KOREAN Chem. Soc.* 2002, *23* (1), 27–28.

(217) Wittayanarakul, K.; Aruksakunwong, O.; Saen-oon, S.; Chantratita, W.; Parasuk, V.; Sompornpisut, P.; Hannongbua, S. Insights into Saquinavir Resistance in the G48V HIV-1 Protease: Quantum Calculations and Molecular Dynamic Simulations. *Biophys. J.* 2005, *88* (2), 867–879.

(218) Altman, D. G.; Bland, J. M. Standard Deviations and Standard Errors. *Bmj* 2005, *331* (7521), 903.

(219) Grossfield, A.; Zuckerman, D. M. Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. In *Annual Reports in Computational Chemistry*; Elsevier, 2009; Vol. 5, pp 23–48.

(220) Smith, L. *A Tutorial on Principal Components Analysis*; Wiley: New York, 2002.

(221) Skjaerven, L.; Martinez, A.; Reuter, N. Principal Component and Normal Mode Analysis of Proteins; a Quantitative Comparison Using the GroEL Subunit. *Proteins Struct. Funct. Bioinforma.* 2011, *79* (1), 232–243.

(222) Haider, S.; Parkinson, G. N.; Neidle, S. Molecular Dynamics and Principal Components Analysis of Human Telomeric Quadruplex Multimers. *Biophys. J.* 2008, *95* (1), 296–311.

(223) Jolliffe, I. . *Principle Component Analysis*, Second.; Springer, 2002.

(224) PCAsuite - Molecular Modelling & Bioinformatics Group http://mmb.pcb.ub.es/software/pcasuite/pcasuite.html (accessed Feb 26, 2014).

(225) NumPy – Numpy http://www.numpy.org/ (accessed Feb 26, 2014).

(226) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 2007, *9* (3), 90–95.

(227) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera–A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 2004, *25* (13), 1605–1612.

(228) ChemAxon. *MarvinSketch*; 2011.

(229) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* 2011, *51* (10), 2778–2786.

(230) Baron, R.; de Vries, A. H.; Hünenberger, P. H.; van Gunsteren, W. F. Configurational Entropies of Lipids in Pure and Mixed Bilayers from Atomic-Level and Coarse-Grained Molecular Dynamics Simulations. *J. Phys. Chem. B* 2006, *110* (31), 15602–15614.

(231) Solomons, T. W. G.; Fryhle, C. B.; Snyder, S. A. *Organic Chemistry*, 11 Unbnd edition.; John Wiley & Sons, 2013.

(232) Mo, Y. Computational Evidence That Hyperconjugative Interactions Are Not Responsible for the Anomeric Effect. *Nat. Chem.* 2010, *2* (8), 666–671.

(233) Becker, K. E.; Fichthorn, K. A. Accelerated Molecular Dynamics Simulation of the Thermal Desorption of N-Alkanes from the Basal Plane of Graphite. *J. Chem. Phys.* 2006, *125* (18), 184706.

(234) Thomas, L. L.; Christakis, T. J.; Jorgensen, W. L. Conformation of Alkanes in the Gas Phase and Pure Liquids. *J. Phys. Chem. B* 2006, *110* (42), 21198–21204.

(235) Ercolani, G. Comment on "Using a Convenient, Quantitative Model for Torsional Entropy To Establish Qualitative Trends for Molecular Processes That

Restrict Conformational Freedom." *J. Org. Chem.* 1999, *64* (9), 3350–3353.

(236) Sherman, A. One Hippopotami (Allan Sherman) http://tinyurl.com/cl9v5o (accessed Apr 1, 2014).

(237) Lefebvre, E.; Schiffer, C. A. Resilience to Resistance of HIV-1 Protease Inhibitors: Profile of Darunavir. *AIDS Rev.* 2008, *10* (3), 131.

(238) Ghosh, A. K.; Ramu Sridhar, P.; Kumaragurubaran, N.; Koh, Y.; Weber, I. T.; Mitsuya, H. Bis-Tetrahydrofuran: A Privileged Ligand for Darunavir and a New Generation of HIV Protease Inhibitors That Combat Drug Resistance. *ChemMedChem* 2006, *1* (9), 939–950.

(239) Ohtaka, H.; Freire, E. Adaptive Inhibitors of the HIV-1 Protease. *Prog. Biophys. Mol. Biol.* 2005, *88* (2), 193–208.

(240) Lazaridis, T.; Masunov, A.; Gandolfo, F. Contributions to the Binding Free Energy of Ligands to Avidin and Streptavidin. *Proteins Struct. Funct. Genet.* 2002, *47* (2), 194–208.

(241) Cerutti, D. S.; Trong, I. L.; Stenkamp, R. E.; Lybrand, T. P. Dynamics of the Streptavidin–Biotin Complex in Solution and in Its Crystal Lattice: Distinct Behavior Revealed by Molecular Simulations. *J. Phys. Chem. B* 2009, *113* (19), 6971–6985.

(242) Metcalfe, J. C.; Birdsall, N. J. M.; Feeney, J.; Lee, A. G.; Levine, Y. K.; Partington, P. 13C NMR Spectra of Lecithin Vesicles and Erythrocyte Membranes. *Nature* 1971, *233* (5316), 199–201.

(243) Lyerla, J. R.; McIntyre, H. M.; Torchia, D. A. A 13C Nuclear Magnetic Resonance Study of Alkane Motion. *Macromolecules* 1974, *7* (1), 11–14.

(244) Gedde, U. *Polymer Physics*; Springer, 1995.

(245) Shaw, M. T.; MacKnight, W. J. *Introduction to Polymer Viscoelasticity*; John Wiley & Sons, 2005.

(246) Sperling, L. H. *Introduction to Physical Polymer Science*; John Wiley & Sons, 2005.

(247) Glass transition - Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Glass_transition#The_glass_transition_in_specific_materials (accessed Feb 26, 2014).

(248) Baron, R.; Hünenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* 2009, *5* (12), 3150–3160.

(249) Baron, R.; van Gunsteren, W. F.; Hünenberger, P. H. Estimating the Configurational Entropy from Molecular Dynamics Simulations: Anharmonicity and Correlation Corrections to the Quasi-Harmonic Approximation. *Trends Phys Chem* 2006, *11*, 87–122.

(250) Carlsson, J.; Åqvist, J. Absolute and Relative Entropies from Computer Simulation with Applications to Ligand Binding. *J. Phys. Chem. B* 2005, *109* (13),

6448–6456.

(251) Finkelstein, A. V.; Janin, J. The Price of Lost Freedom: Entropy of Bimolecular Complex Formation. *Protein Eng.* 1989, *3* (1), 1–3.

(252) Morton, A.; Baase, W. A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry (Mosc.)* 1995, *34* (27), 8564–8575.

(253) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* 2007, *371* (4), 1118–1134.

(254) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get." *Structure* 2009, *17* (4), 489–498.

(255) Hong, L.; Zhang, X. C.; Hartsuck, J. A.; Tang, J. Crystal Structure of an in Vivo HIV-1 Protease Mutant in Complex with Saquinavir: Insights into the Mechanisms of Drug Resistance. *Protein Sci.* 2000, *9* (10), 1898–1904.

(256) Shen, C.-H.; Wang, Y.-F.; Kovalevsky, A. Y.; Harrison, R. W.; Weber, I. T. Amprenavir Complexes with HIV-1 Protease and Its Drug-Resistant Mutants Altering Hydrophobic Clusters: HIV Protease Mutants Altering Hydrophobic Clusters. *FEBS J.* 2010, *277* (18), 3699–3714.

(257) Hult, K.; Berglund, P. Enzyme Promiscuity: Mechanism and Applications. *Trends Biotechnol.* 2007, *25* (5), 231–238.

(258) Jensen, R. A. Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* 1976, *30* (1), 409–425.

(259) Chakraborty, S.; Rao, B. J. A Measure of the Promiscuity of Proteins and Characteristics of Residues in the Vicinity of the Catalytic Site That Regulate Promiscuity. *PLoS ONE* 2012, *7* (2), e32011.

(260) Nobeli, I.; Favia, A. D.; Thornton, J. M. Protein Promiscuity and Its Implications for Biotechnology. *Nat. Biotechnol.* 2009, *27* (2), 157–167.

(261) Schreiber, G.; Keating, A. E. Protein Binding Specificity versus Promiscuity. *Curr. Opin. Struct. Biol.* 2011, *21* (1), 50–61.

(262) Münz, M.; Hein, J.; Biggin, P. C. The Role of Flexibility and Conformational Selection in the Binding Promiscuity of PDZ Domains. *PLoS Comput Biol* 2012, *8* (11), e1002749.

(263) Feldmeier, K.; Höcker, B. Computational Protein Design of Ligand Binding and Catalysis. *Curr. Opin. Chem. Biol.* 2013, *17* (6), 929–933.

(264) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical Investigation into the Structural Complementarity of Natural Products and Synthetic Compounds. *Angew. Chem. Int. Ed.* 1999, *38* (5), 643–647.

(265) Basmadjian, C.; Zhao, Q.; Bentouhami, E.; Djehal, A.; Nebigil, C. G.; Johnson, R. A.; Serova, M.; de Gramont, A.; Faivre, S.; Raymond, E.; Desaubry, L. G. Cancer Wars: Natural Products Strike Back. *Front. Chem.* 2014, *2.*

(266) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate—DNA Helices. *J. Am. Chem. Soc.* 1998, *120* (37), 9401–9409.

(267) Vorobjev, Y. N.; Hermans, J. ES/IS: Estimation of Conformational Free Energy by Combining Dynamics Simulations with Explicit Solvent with an Implicit Solvent Continuum Model. *Biophys. Chem.* 1999, *78* (1–2), 195–205.

(268) Åqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand Binding Affinities from MD Simulations. *Acc. Chem. Res.* 2002, *35* (6), 358–365.

(269) Linear Interaction Energy: Method and Applications in Drug Design - Springer; Baron, R., Ed.; Methods in Molecular Biology; Springer New York, 2012.

(270) Gutiérrez-de-Terán, H.; \AAqvist, J. LIE: Method and Applications in Drug Design.

(271) Swanson, J. M.; Henchman, R. H.; McCammon, J. A. Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophys. J.* 2004, *86* (1), 67–74.

(272) Mark, A. E.; van Gunsteren, W. F. Decomposition of the Free Energy of a System in Terms of Specific Interactions: Implications for Theoretical and Experimental Studies. *J. Mol. Biol.* 1994, *240* (2), 167–176.

(273) Boresch, S.; Archontis, G.; Karplus, M. Free Energy Simulations: The Meaning of the Individual Contributions from a Component Analysis. *Proteins Struct. Funct. Bioinforma.* 1994, *20* (1), 25–33.

(274) Gohlke, H.; Kiel, C.; Case, D. A. Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes. *J. Mol. Biol.* 2003, *330* (4), 891–913.

(275) Brady, G. P.; Szabo, A.; Sharp, K. A. On the Decomposition of Free Energies. *J. Mol. Biol.* 1996, *263* (2), 123–125.

(276) Wu, X.; Wan, S.; Wang, G.; Jin, H.; Li, Z.; Tian, Y.; Zhu, Z.; Zhang, J. Molecular Dynamics Simulation and Free Energy Calculation Studies of Kinase Inhibitors Binding to Active and Inactive Conformations of VEGFR-2. *J. Mol. Graph. Model.* 2015, *56*, 103–112.

(277) Berhanu, W. M.; Masunov, A. E. Full Length Amylin Oligomer Aggregation: Insights from Molecular Dynamics Simulations and Implications for Design of Aggregation Inhibitors. *J. Biomol. Struct. Dyn.* 2014, *32* (10), 1651–1669.

(278) Moonrin, N.; Songtawee, N.; Rattanabunyong, S.; Chunsrivirot, S.; Mokmak, W.; Tongsima, S.; Choowongkomon, K. Understanding the Molecular Basis of EGFR Kinase domain/MIG-6 Peptide Recognition Complex Using Computational Analyses. *BMC Bioinformatics* 2015, *16* (1).

(279) Carra, C.; Saha, J.; Cucinotta, F. A. Theoretical Prediction of the Binding Free Energy for Mutants of Replication Protein A. *J. Mol. Model.* 2011, *18* (7), 3035–3049.

(280) Gao, C.; Grøtli, M.; Eriksson, L. A. Characterization of Interactions and Pharmacophore Development for DFG-out Inhibitors to RET Tyrosine Kinase. *J. Mol. Model.* 2015, *21* (7), 1–12.

(281) Wang, L.; Berne, B. J.; Friesner, R. A. Ligand Binding to Protein-Binding Pockets with Wet and Dry Regions. *Proc. Natl. Acad. Sci. U. S. A.* 2011, *108* (4), 1326–1330.

(282) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 2011, *51* (1), 69–82.

(283) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc.* 2001, *123* (22), 5221–5230.

(284) Srivastava, H. K.; Sastry, G. N. Molecular Dynamics Investigation on a Series of HIV Protease Inhibitors: Assessing the Performance of MM-PBSA and MM-GBSA Approaches. *J. Chem. Inf. Model.* 2012, *52* (11), 3088–3098.

(285) Amber Advanced Tutorials - Tutorial 3 - MM-PBSA - Introduction http://ambermd.org/tutorials/advanced/tutorial3/ (accessed Jun 27, 2015).

(286) Chen, J.-Z.; Myint, K.-Z.; Xie, X.-Q. New QSAR Prediction Models Derived from GPCR CB2-Antagonistic Triaryl Bis-Sulfone Analogues by a Combined Molecular Morphological and Pharmacophoric Approach. *SAR QSAR Environ. Res.* 2011, *22* (5-6), 525–544.

(287) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating Molecular Shape into the Alignment-Free GRid-INdependent Descriptors. *J. Med. Chem.* 2004, *47* (11), 2805–2815.

(288) Mekenyan, O. Dynamic QSAR Techniques: Applications in Drug Design and Toxicology. *Curr. Pharm. Des.* 2002, *8* (17), 1605–1621.

(289) Taylor, R. E.; Chen, Y.; Beatty, A.; Myles, D. C.; Zhou, Y. Conformation—Activity Relationships in Polyketide Natural Products: A New Perspective on the Rational Design of Epothilone Analogues. *J. Am. Chem. Soc.* 2003, *125* (1), 26–27.

(290) Carotenuto, A.; D'Ursi, A. M.; Mulinacci, B.; Paolini, I.; Lolli, F.; Papini, A. M.; Novellino, E.; Rovero, P. Conformation—Activity Relationship of Designed Glycopeptides as Synthetic Probes for the Detection of Autoantibodies, Biomarkers of Multiple Sclerosis. *J. Med. Chem.* 2006, *49* (17), 5072–5079.

(291) Schroeder, D. V. *An Introduction to Thermal Physics*, 1 edition.; Pearson, 2013.

(292) Lemons, D. S. *A Student's Guide to Entropy*; Cambridge University Press: Cambridge, 2013.

(293) Nash, L. K. *Elements of Statistical Thermodynamics*, 2nd edition.; Dover Publications

Inc.: Mineola, N.Y, 2006.

(294) Ben-Naim, A. *Entropy Demystified: The Second Law Reduced to Plain Common Sense*, Expanded edition.; World Scientific: New Jersey, 2008.

(295) Ben-Naim, A. *A Farewell To Entropy*; World Scientific Publishing Company: Hackensack, N.J, 2008.

(296) Lee, J. C. *Thermal Physics: Entropy and Free Energies*, 2 edition.; World Scientific Publishing Company, 2011.

(297) Kestin, J. *A Course in Thermodynamics. Volume 2.*, First edition.; Blaisdell, 1968.

(298) Hill, T. L. *An Introduction to Statistical Thermodynamics*, New edition edition.; Dover Publications Inc.: New York, 2003.

(299) Doty, P.; Myers, G. E. Low Molecular Weight Proteins. Thermodynamics of the Association of Insulin Molecules. *Discuss. Faraday Soc.* 1953, *13* (0), 51–58.

(300) Janin, J.; Chothia, C. Role of Hydrophobicity in the Binding of Coenzymes. *Biochemistry (Mosc.)* 1978, *17* (15), 2943–2948.

(301) Searle, M. S.; Williams, D. H.; Gerhard, U. Partitioning of Free Energy Contributions in the Estimation of Binding Constants: Residual Motions and Consequences for Amide-Amide Hydrogen Bond Strengths. *J. Am. Chem. Soc.* 1992, *114* (27), 10697–10704.

(302) Steinberg, I. Z.; Scheraga, H. A. Entropy Changes Accompanying Association Reactions of Proteins. *J. Biol. Chem.* 1963, *238* (1), 172–181.

(303) Zhou, H.-X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* 2009, *109* (9), 4092–4107.

(304) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* 1997, *72* (3), 1047–1069.

(305) Eckart, C. Some Studies Concerning Rotating Axes and Polyatomic Molecules. *Phys. Rev.* 1935, *47* (7), 552–558.

(306) Sayvetz, A. The Kinetic Energy of Polyatomic Molecules. *J. Chem. Phys.* 1939, *7* (6), 383–389.

(307) Williams, D. H.; Cox, J. P. L.; Doig, A. J.; Gardner, M.; Gerhard, U.; Kaye, P. T.; Lal, A. R.; Nicholls, I. A.; Salter, C. J.; Mitchell, R. C. Toward the Semiquantitative Estimation of Binding Constants. Guides for Peptide-Peptide Binding in Aqueous Solution. *J. Am. Chem. Soc.* 1991, *113* (18), 7020–7030.

(308) Privalov, P. L.; Tamura, A. Comments on the Comments. *Protein Eng.* 1999, *12* (3), 187–187.

(309) Dunitz, J. D. The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* 1994, *264* (5159), 670–670.

(310) Murray, C. W.; Verdonk, M. L. The Consequences of Translational and Rotational Entropy Lost by Small Molecules on Binding to Proteins. *J. Comput. Aided Mol. Des.* 2002, *16* (10), 741–753.

(311) Murphy, K. P.; Xie, D.; Thompson, K. S.; Amzel, L. M.; Freire, E. Entropy in Biological Binding Processes: Estimation of Translational Entropy Loss. *Proteins Struct. Funct. Bioinforma.* 1994, *18* (1), 63–67.

(312) Zaman, M. H.; Berry, R. S.; Sosnick, T. R. Entropic Benefit of a Cross-Link in Protein Association. *Proteins Struct. Funct. Bioinforma.* 2002, *48* (2), 341–351.

(313) Turnbull, W. B.; Precious, B. L.; Homans, S. W. Dissecting the Cholera Toxin–Ganglioside GM1 Interaction by Isothermal Titration Calorimetry. *J. Am. Chem. Soc.* 2004, *126* (4), 1047–1054.

(314) Yu, Y. B.; Lavigne, P.; Kay, C. M.; Hodges, R. S.; Privalov, P. L. Contribution of Translational and Rotational Entropy to the Unfolding of a Dimeric Coiled-Coil. *J. Phys. Chem. B* 1999, *103* (12), 2270–2278.

(315) Horton, N.; Lewis, M. Calculation of the Free Energy of Association for Protein Complexes. *Protein Sci.* 1992, *1* (1), 169–181.

(316) Doig, A. J.; Williams, D. H. Binding Energy of an Amide-Amide Hydrogen Bond in Aqueous and Nonpolar Solvents. *J. Am. Chem. Soc.* 1992, *114* (1), 338–343.

(317) Erickson, H. P. Co-Operativity in Protein-Protein Association: The Structure and Stability of the Actin Filament. *J. Mol. Biol.* 1989, *206* (3), 465–474.

(318) Baginski, M.; Fogolari, F.; Briggs, J. M. Electrostatic and Non-Electrostatic Contributions to the Binding Free Energies of Anthracycline Antibiotics to DNA1. *J. Mol. Biol.* 1997, *274* (2), 253–267.

(319) Amzel, L. M. Loss of Translational Entropy in Binding, Folding, and Catalysis. *Proteins Struct. Funct. Genet.* 1997, *28* (2), 144–149.

(320) Denisov, V. P.; Venu, K.; Peters, J.; Hörlein, H. D.; Halle, B. Orientational Disorder and Entropy of Water in Protein Cavities. *J. Phys. Chem. B* 1997, *101* (45), 9380–9389.

(321) Chothia, C.; Janin, J. Principles of Protein–protein Recognition. *Nature* 1975, *256* (5520), 705–708.

(322) Lee, M. S.; Olson, M. A. Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J.* 2006, *90* (3), 864–877.

(323) Luo, H.; Sharp, K. On the Calculation of Absolute Macromolecular Binding Free Energies. *Proc. Natl. Acad. Sci. U. S. A.* 2002, *99* (16), 10399–10404.

(324) Chang, C.-E.; Potter, M. J.; Gilson, M. K. Calculation of Molecular Configuration Integrals. *J. Phys. Chem. B* 2003, *107* (4), 1048–1055.

(325) Siebert, X.; Amzel, L. M. Loss of Translational Entropy in Molecular Associations. *Proteins Struct. Funct. Bioinforma.* 2004, *54* (1), 104–115.

(326) Lu, B.; Wong, C. F. Direct Estimation of Entropy Loss due to Reduced Translational and Rotational Motions upon Molecular Binding. *Biopolymers* 2005, *79* (5), 277–285.

(327) Ruvinsky, A. M. Role of Binding Entropy in the Refinement of Protein–ligand

Docking Predictions: Analysis Based on the Use of 11 Scoring Functions. *J. Comput. Chem.* 2007, *28* (8), 1364–1372.

(328) Zhang, J.; Lazaridis, T. Calculating the Free Energy of Association of Transmembrane Helices. *Biophys. J.* 2006, *91* (5), 1710–1723.

(329) Li, Z.; Lazaridis, T. The Effect of Water Displacement on Binding Thermodynamics: Concanavalin A. *J. Phys. Chem. B* 2005, *109* (1), 662–670.

(330) Li, Z.; Lazaridis, T. Thermodynamic Contributions of the Ordered Water Molecule in HIV-1 Protease. *J. Am. Chem. Soc.* 2003, *125* (22), 6636–6637.

(331) Li, Z.; Lazaridis, T. Thermodynamics of Buried Water Clusters at a Protein—Ligand Binding Interface. *J. Phys. Chem. B* 2006, *110* (3), 1464–1475.

(332) Tamura, A.; Privalov, P. L. The Entropy Cost of Protein Association. *J. Mol. Biol.* 1997, *273* (5), 1048–1060.

(333) Karplus, M.; Janin, J. Comment on:The Entropy Cost of Protein Association'. *Protein Eng.* 1999, *12* (3), 185–186.

(334) Jencks, W. P. Binding Energy, Specificity, and Enzymic Catalysis: The Circe Effect. In *Advances in Enzymology and Related Areas of Molecular Biology*; Meister, A., Ed.; John Wiley & Sons, Inc., 1975; pp 219–410.

(335) Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc. Natl. Acad. Sci.* 2011, *108* (39), 16247–16252.

(336) Tidor, B.; Karplus, M. The Contribution of Vibrational Entropy to Molecular Association. The Dimerization of Insulin. *J. Mol. Biol.* 1994, *238* (3), 405–414.

(337) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. Evaluation of the Configurational Entropy for Proteins: Application to Molecular Dynamics Simulations of an -Helix. *Macromolecules* 1984, *17* (7), 1370–1374.

(338) Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* 1957, *106* (4), 620.

(339) Minh, D. D. L.; Bui, J. M.; Chang, C.; Jain, T.; Swanson, J. M. J.; McCammon, J. A. The Entropic Cost of Protein-Protein Association: A Case Study on Acetylcholinesterase Binding to Fasciculin-2. *Biophys. J.* 2005, *89* (4), L25–L27.

(340) Hermans, J.; Wang, L. Inclusion of Loss of Translational and Rotational Freedom in Theoretical Estimates of Free Energies of Binding. Application to a Complex of Benzene and Mutant T4 Lysozyme. *J. Am. Chem. Soc.* 1997, *119* (11), 2707–2714.

(341) Schwarzl, S. M.; Tschopp, T. B.; Smith, J. C.; Fischer, S. Can the Calculation of Ligand Binding Free Energies Be Improved with Continuum Solvent Electrostatics and an Ideal-Gas Entropy Correction? *J. Comput. Chem.* 2002, *23* (12), 1143–1149.

(342) Grigoriev, F. V.; Luschekina, S. V.; Romanov, A. N.; Sulimov, V. B.; Nikitina,

E. A. Computation of Entropy Contribution to Protein-Ligand Binding Free Energy. *Biochem. Mosc.* 2007, *72* (7), 785–792.

(343) Mammen, M.; Shakhnovich, E. I.; Deutch, J. M.; Whitesides, G. M. Estimating the Entropic Cost of Self-Assembly of Multiparticle Hydrogen-Bonded Aggregates Based on the Cyanuric Acid⊙ Melamine Lattice. *J. Org. Chem.* 1998, *63* (12), 3821–3830.

(344) Gurney, R. W. *Ionic Processes In Solution - Primary Source Edition*; Nabu Press, 2014.

(345) Kauzmann, W. Some Factors in the Interpretation of Protein Denaturation. *Adv. Protein Chem.* 1959, *14*, 1–63.

(346) Tanford, C. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd Edition edition.; Wiley-Blackwell: New York, 1980.

(347) Holtzer, A. The "cratic Correction" and Related Fallacies. *Biopolymers* 1995, *35* (6), 595–602.

(348) Novotny, J.; Bruccoleri, R. E.; Saul, F. A. On the Attribution of Binding Energy in Antigen-Antibody Complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry (Mosc.)* 1989, *28* (11), 4735–4749.

(349) Henchman, R. H. Partition Function for a Simple Liquid Using Cell Theory Parametrized by Computer Simulation. *J. Chem. Phys.* 2003, *119* (1), 400.

(350) Rice, O. K. On Communal Entropy and the Theory of Fusion. *J. Chem. Phys.* 1938, *6* (8), 476–479.

(351) Kirkwood, J. G. Critique of the Free Volume Theory of the Liquid State. *J. Chem. Phys.* 1950, *18* (3), 380–382.

(352) Henchman, R. H. Free Energy of Liquid Water from a Computer Simulation via Cell Theory. *J. Chem. Phys.* 2007, *126* (6), 064504.

(353) Klefas-Stennett, M.; Henchman, R. H. Classical and Quantum Gibbs Free Energies and Phase Behavior of Water Using Simulation and Cell Theory†. *J. Phys. Chem. B* 2008, *112* (32), 9769–9776.

(354) Morton, A.; Matthews, B. W. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry (Mosc.)* 1995, *34* (27), 8576–8588.

(355) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, *12*, 2825–2830.

(356) Big O Notation. *Wikipedia, the free encyclopedia*; 2015.

(357) Hammer Projection – Basemap Matplotlib Toolkit 1.0.8 documentation http://matplotlib.org/basemap/users/hammer.html (accessed Jul 29, 2015).

(358) Furuti, C. Map Projections: Modified Azimuthal Projections http://www.progonos.com/furuti/MapProj/Normal/ProjMAz/projMAz.html (accessed Jul 29, 2015).

(359) Furuti, C. Map Projections: Preserving Areas http://www.progonos.com/furuti/

MapProj/Normal/CartProp/AreaPres/areaPres.html (accessed Jul 29, 2015).

(360) MacRaild, C. A.; Daranas, A. H.; Bronowska, A.; Homans, S. W. Global Changes in Local Protein Dynamics Reduce the Entropic Cost of Carbohydrate Binding in the Arabinose-Binding Protein. *J. Mol. Biol.* 2007, *368* (3), 822–832.

(361) Best, R. B.; Vendruscolo, M. Determination of Protein Structures Consistent with NMR Order Parameters. *J. Am. Chem. Soc.* 2004, *126* (26), 8090–8091.

(362) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. A Cavity-Containing Mutant of T4 Lysozyme Is Stabilized by Buried Benzene. *Nature* 1992, *355* (6358), 371–373.

(363) Mann, G.; Hermans, J. Modeling Protein-Small Molecule Interactions: Structure and Thermodynamics of Noble Gases Binding in a Cavity in Mutant Phage T4 Lysozyme L99A1. *J. Mol. Biol.* 2000, *302* (4), 979–989.

(364) Lazaridis, T.; Karplus, M. "New View" of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations. *Science* 1997, *278* (5345), 1928–1931.

(365) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. Lessons from the Lysozyme of Phage T4. *Protein Sci.* 2010, *19* (4), 631–641.

(366) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. Configurational Entropy in Protein-Peptide Binding. Computational Study of Tsg101 UEV Domain with an HIV-Derived PTAP Nonapeptide. *J. Mol. Biol.* 2009, *389* (2), 315–335.

(367) Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L.; Chen, W.; Moghaddam, S.; Gilson, M. K.; Kim, K.; Inoue, Y. A Synthetic Host-Guest System Achieves Avidin-Biotin Affinity by Overcoming Enthalpy-Entropy Compensation. *Proc. Natl. Acad. Sci. U. S. A.* 2007, *104* (52), 20737–20742.

(368) Syme, N. R.; Dennis, C.; Bronowska, A.; Paesen, G. C.; Homans, S. W. Comparison of Entropic Contributions to Binding in a "Hydrophilic" versus "Hydrophobic" Ligand-Protein Interaction. *J. Am. Chem. Soc.* 2010, *132* (25), 8682–8689.

(369) Merritt, E. A.; Kuhn, P.; Sarfaty, S.; Erbe, J. L.; Holmes, R. K.; Hol, W. G. J. The 1.25 Å Resolution Refinement of the Cholera Toxin B-Pentamer: Evidence of Peptide Backbone Strain at the Receptor-Binding site1. *J. Mol. Biol.* 1998, *282* (5), 1043–1059.

(370) Merritt, E. A.; Sarfaty, S.; Akker, F. V. D.; L'Hoir, C.; Martial, J. A.; Hol, W. G. J. Crystal Structure of Cholera Toxin B-Pentamer Bound to Receptor GM1 Pentasaccharide. *Protein Sci.* 1994, *3* (2), 166–175.

(371) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* 2006, *65* (3), 712–

725.

(372) Showalter, S. A.; Brüschweiler, R. Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field. *J. Chem. Theory Comput.* 2007, *3* (3), 961–975.

(373) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS ONE* 2012, *7* (2), e32131.

(374) Li, D.-W.; Brüschweiler, R. NMR-Based Protein Potentials. *Angew. Chem.* 2010, *122* (38), 6930–6932.

(375) Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* 2012, *8* (4), 1409–1414.

(376) Keutsch, F. N.; Saykally, R. J. Water Clusters: Untangling the Mysteries of the Liquid, One Molecule at a Time. *Proc. Natl. Acad. Sci.* 2001, *98* (19), 10533–10540.

(377) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* 2002, *99* (20), 12562–12566.

(378) Laio, A.; Gervasio, F. L. Metadynamics: A Method to Simulate Rare Events and Reconstruct the Free Energy in Biophysics, Chemistry and Material Science. *Rep. Prog. Phys.* 2008, *71* (12), 126601.

(379) *Free Energy Calculations*; Chipot, C., Pohorille, A., Eds.; Castleman, A. W., Toennies, J. P., Yamanouchi, K., Zinth, W., Series Eds.; Springer Series in CHEMICAL PHYSICS; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007; Vol. 86.

(380) Dellago, C.; Bolhuis, P. G. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*; Holm, P. C., Kremer, P. K., Eds.; Advances in Polymer Science; Springer Berlin Heidelberg, 2009; pp 167–233.

(381) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2011, *1* (5), 826–843.

(382) Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. Exploring Multiple Timescale Motions in Protein GB3 Using Accelerated Molecular Dynamics and NMR Spectroscopy. *J. Am. Chem. Soc.* 2007, *129* (15), 4724–4730.

(383) Borden, N.; Linklater, D. Hickam's Dictum. *West. J. Emerg. Med.* 2013, *14* (2), 164–164.

(384) NeuroLogica Blog » Occam's Razor vs Hickam's Dictum.

(385) Farrow, N. A.; Muhandiram, R.; Singer, A. U.; Pascal, S. M.; Kay, C. M.; Gish, G.; Shoelson, S. E.; Pawson, T.; Forman-Kay, J. D.; Kay, L. E. Backbone Dynamics of a Free and a Phosphopeptide-Complexed Src Homology 2 Domain Studied by 15N NMR Relaxation. *Biochemistry (Mosc.)* 1994, *33* (19), 5984–6003.

(386) Stivers, J. T.; Abeygunawardana, C.; Mildvan, A. S.; Whitman, C. P. 15N NMR

Relaxation Studies of Free and Inhibitor-Bound 4-Oxalocrotonate Tautomerase: Backbone Dynamics and Entropy Changes of an Enzyme upon Inhibitor Binding. *Biochemistry (Mosc.)* 1996, *35* (50), 16036–16047.

(387) Yu, L.; Zhu, C.-X.; Tse-Dinh, Y.-C.; Fesik, S. W. Backbone Dynamics of the C-Terminal Domain of Escherichia Coli Topoisomerase I in the Absence and Presence of Single-Stranded DNA. *Biochemistry (Mosc.)* 1996, *35* (30), 9661–9666.

(388) Sharrow, S. D.; Novotny, M. V.; Stone, M. J. Thermodynamic Analysis of Binding between Mouse Major Urinary Protein-I and the Pheromone 2- *sec* -Butyl-4,5-Dihydrothiazole †. *Biochemistry (Mosc.)* 2003, *42* (20), 6302–6309.

(389) Gasymov, O. K.; Abduragimov, A. R.; Glasgow, B. J. Excited Protein States of Human Tear Lipocalin for Low- and High-Affinity Ligand Binding Revealed by Functional AB Loop Motion. *Biophys. Chem.* 2010, *149* (1-2), 47–57.

(390) Shimamoto, S.; Yoshida, T.; Ohkubo, T. Ligand Recognition Mechanism of Lipocalin-type Prostaglandin D Sythase. *Pharm. Soc. Jpn.* 2011.

(391) Hajjar, E.; Perahia, D.; Débat, H.; Nespoulous, C.; Robert, C. H. Odorant Binding and Conformational Dynamics in the Odorant-Binding Protein. *J. Biol. Chem.* 2006, *281* (40), 29929–29937.

(392) Golebiowski, J.; Antonczak, S.; Fiorucci, S.; Cabrol-Bass, D. Mechanistic Events Underlying Odorant Binding Protein Chemoreception. *Proteins Struct. Funct. Bioinforma.* 2007, *67* (2), 448–458.

(393) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B* 2009, *113* (40), 13337–13346.

(394) Olano, L. R.; Rick, S. W. Hydration Free Energies and Entropies for Water in Protein Interiors. *J. Am. Chem. Soc.* 2004, *126* (25), 7991–8000.

(395) Baron, R.; Setny, P.; Andrew McCammon, J. Water in Cavity–Ligand Recognition. *J. Am. Chem. Soc.* 2010, *132* (34), 12091–12097.

(396) Setny, P.; Baron, R.; McCammon, J. A. How Can Hydrophobic Association Be Enthalpy Driven? *J. Chem. Theory Comput.* 2010, *6* (9), 2866–2871.

(397) Graziano, G. Comment on ?Entropy/enthalpy Compensation: Hydrophobic Effect, Micelles and Protein Complexes? By E. Fisicaro, C. Compari and A. Braibanti, Phys. Chem. Chem. Phys., 2004, 6, 4156. *Phys. Chem. Chem. Phys.* 2005, 7 (6), 1322.

(398) Setny, P.; Baron, R.; McCammon, J. A. Comment on "Molecular Driving Forces of the Pocket-Ligand Hydrophobic Association" by G. Graziano, Chem. Phys. Lett. 533 (2012) 95. *Chem. Phys. Lett.* 2013, *555*, 306–309.

(399) Graziano, G. Reply to the Comment by Setny, Baron and McCammon on the Article "Molecular Driving Forces of the Pocket-Ligand Hydrophobic Association", Chem. Phys. Lett. 533 (2012) 95. *Chem. Phys. Lett.* 2013, *555*, 310–311.

(400) Fersht, A. *Structure and Mechanism in Protein Science: Guide to Enzyme Catalysis and*

Protein Folding, 3rd Revised edition edition.; W.H.Freeman & Co Ltd: New York, 1999.

(401) Ubbink, M. The Courtship of Proteins: Understanding the Encounter Complex. *FEBS Lett.* 2009, *583* (7), 1060–1066.

(402) Kozakov, D.; Li, K.; Hall, D. R.; Beglov, D.; Zheng, J.; Vakili, P.; Schueler-Furman, O.; Paschalidis, I. C.; Clore, G. M.; Vajda, S. Encounter Complexes and Dimensionality Reduction in Protein–protein Association. *eLife* 2014, *3*, e01370.

(403) Held, M.; Metzner, P.; Prinz, J.-H.; Noé, F. Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations. *Biophys. J.* 2011, *100* (3), 701–710.

(404) Bla ejczyk, M.; Burdzy, K.; Góralski, G.; Bocquet, L.; others. Reduction of Dimensionality in a Diffusion Search Process and Kinetics of Gene Expression. *Phys. Stat. Mech. Its Appl.* 2000, *277* (1), 71–82.

(405) Ruvinsky, A. M.; Vakser, I. A. Chasing Funnels on Protein-Protein Energy Landscapes at Different Resolutions. *Biophys. J.* 2008, *95* (5), 2150–2159.

(406) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein Folding Funnels: A Kinetic Approach to the Sequence-Structure Relationship. *Proc. Natl. Acad. Sci.* 1992, *89* (18), 8721–8725.

(407) Wang, J.; Zheng, X.; Yang, Y.; Drueckhammer, D.; Yang, W.; Verkhivker, G.; Wang, E. Quantifying Intrinsic Specificity: A Potential Complement to Affinity in Drug Screening. *Phys. Rev. Lett.* 2007, *99* (19).

(408) Wang, J.; Verkhivker, G. M. Energy Landscape Theory, Funnels, Specificity, and Optimal Criterion of Biomolecular Binding. *Phys. Rev. Lett.* 2003, *90* (18).

(409) Getzoff, E. D.; Cabelli, D. E.; Fisher, C. L.; Parge, H. E.; Viezzoli, M. S.; Banci, L.; Hallewell, R. A. Faster Superoxide Dismutase Mutants Designed by Enhancing Electrostatic Guidance. *Nature* 1992, *358* (6384), 347–351.

(410) Swinney, D. C. Biochemical Mechanisms of Drug Action: What Does It Take for Success? *Nat. Rev. Drug Discov.* 2004, *3* (9), 801–808.

(411) Limongelli, V.; Bonomi, M.; Parrinello, M. Funnel Metadynamics as Accurate Binding Free-Energy Method. *Proc. Natl. Acad. Sci.* 2013, *110* (16), 6358–6363.

(412) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* 1935, *3* (5), 300.

(413) Kirkwood, J. G. *Theory of Liquids*, 1st Edition edition.; Routledge, 1968.

(414) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* 1954, *22* (8), 1420–1426.

(415) Landau, L. D.; Lifshitz, E. M. *Statistical Physics*, 1st edition.; Clarendon Press, 1938.

(416) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. Steered Molecular Dynamics. In *Computational Molecular*

Dynamics: Challenges, Methods, Ideas; Deuflhard, P., Hermans, J., Leimkuhler, B., Mark, A. E., Reich, S., Skeel, R. D., Eds.; Lecture Notes in Computational Science and Engineering; Springer Berlin Heidelberg, 1999; pp 39–65.

(417) Schlitter, J.; Engels, M.; Krüger, P. Targeted Molecular Dynamics: A New Approach for Searching Pathways of Conformational Transitions. *J. Mol. Graph.* 1994, *12* (2), 84–89.

(418) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* 1992, *13* (8), 1011–1021.

(419) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* 2011, *108* (25), 10184–10189.

(420) Wu, C.; Biancalana, M.; Koide, S.; Shea, J.-E. Binding Modes of Thioflavin-T to the Single-Layer -Sheet of the Peptide Self-Assembly Mimics. *J. Mol. Biol.* 2009, *394* (4), 627–633.

(421) Brannigan, G.; LeBard, D. N.; Hénin, J.; Eckenhoff, R. G.; Klein, M. L. Multiple Binding Sites for the General Anesthetic Isoflurane Identified in the Nicotinic Acetylcholine Receptor Transmembrane Domain. *Proc. Natl. Acad. Sci. U. S. A.* 2010, *107* (32), 14122–14127.

(422) Vanni, S.; Neri, M.; Tavernelli, I.; Rothlisberger, U. Predicting Novel Binding Modes of Agonists to Adrenergic Receptors Using All-Atom Molecular Dynamics Simulations. *PLoS Comput Biol* 2011, 7 (1), e1001053.

(423) Martí, M. A.; Lebrero, M. C. G.; Roitberg, A. E.; Estrin, D. A. Bond or Cage Effect: How Nitrophorins Transport and Release Nitric Oxide. *J. Am. Chem. Soc.* 2008, *130* (5), 1611–1618.

(424) Enkavi, G.; Tajkhorshid, E. Simulation of Spontaneous Substrate Binding Revealing the Binding Pathway and Mechanism and Initial Conformational Response of GlpT. *Biochemistry (Mosc.)* 2010, *49* (6), 1105–1114.

(425) Ahmad, M.; Gu, W.; Helms, V. Mechanism of Fast Peptide Recognition by SH3 Domains. *Angew. Chem. Int. Ed.* 2008, *47* (40), 7626–7630.

(426) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* 2011, *133* (24), 9181–9183.

(427) Troussicot, L.; Guillière, F.; Limongelli, V.; Walker, O.; Lancelin, J.-M. Funnel-Metadynamics and Solution NMR to Estimate Protein–Ligand Affinities. *J. Am. Chem. Soc.* 2015, *137* (3), 1273–1281.

(428) Valas, R.; Langmead, C. J. Classifying Protein Structural Dynamics via Residual Dipolar Couplings. 2004.

(429) Qiu, X.; Janson, C. A.; Blackburn, M. N.; Chhohan, I. K.; Hibbs, M.; Abdel-Meguid, S. S. Cooperative Structural Dynamics and a Novel Fidelity Mechanism

in Histidyl-tRNA Synthetases. *Biochemistry (Mosc.)* 1999, *38* (38), 12296–12304.

(430) Desamero, R.; Rozovsky, S.; Zhadin, N.; McDermott, A.; Callender, R. Active Site Loop Motion in Triosephosphate Isomerase: T-Jump Relaxation Spectroscopy of Thermal Activation†. *Biochemistry (Mosc.)* 2003, *42* (10), 2941–2951.

(431) Zhai, X.; Go, M. K.; O'Donoghue, A. C.; Amyes, T. L.; Pegan, S. D.; Wang, Y.; Loria, J. P.; Mesecar, A. D.; Richard, J. P. Enzyme Architecture: The Effect of Replacement and Deletion Mutations of Loop 6 on Catalysis by Triosephosphate Isomerase. *Biochemistry (Mosc.)* 2014, *53* (21), 3486–3501.

(432) Callender, R.; Dyer, R. B. Probing Protein Dynamics Using Temperature Jump Relaxation Spectroscopy. *Curr. Opin. Struct. Biol.* 2002, *12* (5), 628–633.

(433) Deng, H.; Zhadin, N.; Callender, R. Dynamics of Protein Ligand Binding on Multiple Time Scales: NADH Binding to Lactate Dehydrogenase†. *Biochemistry (Mosc.)* 2001, *40* (13), 3767–3773.

(434) Sassaman, C. Dynamics of a Lactate Dehydrogenase Polymorphism in the Wood Louse PORCELLIO SCABER Latr.: Evidence for Partial Assortative Mating and Heterosis in Natural Populations. *Genetics* 1978, *88* (3), 591–609.

(435) Masterson, J. E.; Schwartz, S. D. Changes in Protein Architecture and Subpicosecond Protein Dynamics Impact the Reaction Catalyzed by Lactate Dehydrogenase. *J. Phys. Chem. A* 2013, *117* (32), 7107–7113.

(436) Gasymov, O. K.; Abduragimov, A. R.; Glasgow, B. J. pH-Dependent Conformational Changes in Tear Lipocalin by Site-Directed Tryptophan Fluorescence. *Biochemistry (Mosc.)* 2010, *49* (3), 582–590.

(437) Eberini, I.; Baptista, A. M.; Gianazza, E.; Fraternali, F.; Beringhelli, T. Reorganization in Apo- and Holo- -Lactoglobulin upon Protonation of Glu89: Molecular Dynamics and pKa Calculations. *Proteins Struct. Funct. Bioinforma.* 2004, *54* (4), 744–758.

(438) Fogolari, F.; Moroni, E.; Wojciechowski, M.; Baginski, M.; Ragona, L.; Molinari, H. MM/PBSA Analysis of Molecular Dynamics Simulations of Bovine -Lactoglobulin: Free Energy Gradients in Conformational Transitions? *Proteins Struct. Funct. Bioinforma.* 2005, *59* (1), 91–103.

(439) Sakurai, K.; Konuma, T.; Yagi, M.; Goto, Y. Structural Dynamics and Folding of -Lactoglobulin Probed by Heteronuclear NMR. *Biochim. Biophys. Acta BBA - Gen. Subj.* 2009, *1790* (6), 527–537.

(440) Ragona, L.; Fogolari, F.; Catalano, M.; Ugolini, R.; Zetta, L.; Molinari, H. EF Loop Conformational Change Triggers Ligand Binding in -Lactoglobulins. *J. Biol. Chem.* 2003, *278* (40), 38840–38846.

(441) Fogolari, F.; Ragona, L.; Licciardi, S.; Romagnoli, S.; Michelutti, R.; Ugolini, R.; Molinari, H. Electrostatic Properties of Bovine -Lactoglobulin. *Proteins Struct. Funct. Bioinforma.* 2000, *39* (4), 317–330.

(442) Calderone, V.; Berni, R.; Zanotti, G. High-Resolution Structures of Retinol-Binding Protein in Complex with Retinol: pH-Induced Protein Structural Changes in the Crystal State. *J. Mol. Biol.* 2003, *329* (4), 841–850.

(443) Newcomer, M. E.; Ong, D. E. Plasma Retinol Binding Protein: Structure and Function of the Prototypic Lipocalin. *Biochim. Biophys. Acta BBA-Protein Struct. Mol. Enzymol.* 2000, *1482* (1), 57–64.

(444) Buttery, R. G.; Bomben, J. L.; Guadagni, D. G.; Ling, L. C. Volatilities of Organic Flavor Compounds in Foods. *J. Agric. Food Chem.* 1971, *19* (6), 1045–1048.

(445) Ernst, J. A.; Clubb, R. T.; Zhou, H. X.; Gronenborn, A. M.; Clore, G. M. Demonstration of Positionally Disordered Water within a Protein Hydrophobic Cavity by NMR. *Science* 1995, *267* (5205), 1813–1817.

(446) Shimizu, S. Estimating Hydration Changes upon Biomolecular Reactions from Osmotic Stress, High Pressure, and Preferential Hydration Experiments. *Proc. Natl. Acad. Sci. U. S. A.* 2004, *101* (5), 1195–1199.

(447) Filfil, R.; Ratavosi, A.; Chalikian, T. V. Binding of Bovine Pancreatic Trypsin Inhibitor to Trypsinogen: Spectroscopic and Volumetric Studies. *Biochemistry (Mosc.)* 2004, *43* (5), 1315–1322.

(448) Taulier, N.; Chalikian, T. V. Hydrophobic Hydration in Cyclodextrin Complexation. *J. Phys. Chem. B* 2006, *110* (25), 12222–12224.

(449) Persson, F.; Halle, B. Transient Access to the Protein Interior: Simulation versus NMR. *J. Am. Chem. Soc.* 2013, *135* (23), 8735–8748.

(450) Huggins, D. J. Quantifying the Entropy of Binding for Water Molecules in Protein Cavities by Computing Correlations. *Biophys. J.* 2015, *108* (4), 928–936.

(451) Lazaridis, T.; Karplus, M. Orientational Correlations and Entropy in Liquid Water. *J. Chem. Phys.* 1996, *105* (10), 4294–4316.

(452) Portman, K. L.; Long, J.; Carr, S.; Briand, L.; Winzor, D. J.; Searle, M. S.; Scott, D. J. Enthalpy/Entropy Compensation Effects from Cavity Desolvation Underpin Broad Ligand Binding Selectivity for Rat Odorant Binding Protein 3. *Biochemistry (Mosc.)* 2014, *53* (14), 2371–2379.

(453) Homans, S. W. Probing the Binding Entropy of Ligand–Protein Interactions by NMR. *ChemBioChem* 2005, *6* (9), 1585–1591.

(454) Homans, S. W. Water, Water Everywhere – except Where It Matters? *Drug Discov. Today* 2007, *12* (13–14), 534–539.

(455) Williams, D. H.; Stephens, E.; Zhou, M. Ligand Binding Energy and Catalytic Efficiency from Improved Packing within Receptors and Enzymes. *J. Mol. Biol.* 2003, *329* (2), 389–399.

(456) Park, S. J.; Borin, B. N.; Martinez-Yamout, M. A.; Dyson, H. J. The Client Protein p53 Forms a Molten Globule-like State in the Presence of Hsp90. *Nat. Struct. Mol. Biol.* 2011, *18* (5), 537–541.

(457) Wand, A. J. The Dark Energy of Proteins Comes to Light: Conformational Entropy and Its Role in Protein Function Revealed by NMR Relaxation. *Curr. Opin. Struct. Biol.* 2013, *23* (1), 75–81.

(458) Bronowska, A. Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design. In *Thermodynamics - Interaction Studies - Solids, Liquids and Gases*; Moreno Pirajn, J. C., Ed.; InTech, 2011.

(459) Nucci, N. V.; Pometun, M. S.; Wand, A. J. Mapping the Hydration Dynamics of Ubiquitin. *J. Am. Chem. Soc.* 2011, *133* (32), 12326–12329.

(460) Klebe, G. Applying Thermodynamic Profiling in Lead Finding and Optimization. *Nat. Rev. Drug Discov.* 2015, *14* (2), 95–110.

(461) Heller, G. T.; Sormanni, P.; Vendruscolo, M. Targeting Disordered Proteins with Small Molecules Using Entropy. *Trends Biochem. Sci.* 2015, *40* (9), 491–496.

(462) Prorok, M.; Castellino, F. J. Thermodynamics of Binding of Calcium, Magnesium, and Zinc to theN-Methyl-D-Aspartate Receptor Ion Channel Peptidic Inhibitors, Conantokin-G and Conantokin-T. *J. Biol. Chem.* 1998, *273* (31), 19573–19578.

(463) Zheng, L.; Krishnamoorthi, R.; Zolkiewski, M.; Wang, X. Distinct Ca2+ Binding Properties of Novel C2 Domains of Plant Phospholipase D and . *J. Biol. Chem.* 2000, *275* (26), 19700–19706.

(464) Tzeng, S.-R.; Kalodimos, C. G. Protein Activity Regulation by Conformational Entropy. *Nature* 2012, *488* (7410), 236–240.

(465) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity. *J. Am. Chem. Soc.* 1982, *104* (17), 4546–4559.

(466) Lipari, G.; Szabo, A. Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results. *J. Am. Chem. Soc.* 1982, *104* (17), 4559–4570.

(467) Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M. Ø.; Xu, H.; Trbovic, N.; Friesner, R. A.; Palmer, A. G.; Shaw, D. E. Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins †. *J. Phys. Chem. B* 2008, *112* (19), 6155–6158.

(468) Carr, H. Y.; Purcell, E. M. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. *Phys. Rev.* 1954, *94* (3), 630–638.

(469) Meiboom, S.; Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Rev. Sci. Instrum.* 1958, *29* (8), 688.

(470) Polyansky, A.; Zubac, R.; Zagrovic, B. Estimation of Conformational Entropy in Protein–Ligand Interactions: A Computational Perspective. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Methods in Molecular Biology; Springer New York, 2012; pp 327–353.

(471) Fenley, A. T.; Muddana, H. S.; Gilson, M. K. Entropy-Enthalpy Transduction

Caused by Conformational Shifts Can Obscure the Forces Driving Protein-Ligand Binding. *Proc. Natl. Acad. Sci.* 2012, *109* (49), 20006–20011.

(472) King, B. M.; Tidor, B. MIST: Maximum Information Spanning Trees for Dimension Reduction of Biological Data Sets. *Bioinformatics* 2009, *25* (9), 1165–1172.

(473) King, B. M.; Silver, N. W.; Tidor, B. Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation. *J. Phys. Chem. B* 2012, *116* (9), 2891–2904.

(474) Flower, D. R. The Lipocalin Protein Family: Structure and Function. *Biochem. J.* 1996, *318* (Pt 1), 1–14.

(475) North, A. C. T. Three-Dimensional Arrangement of Conserved Amino Acid Residues in a Superfamily of Specific Ligand-Binding Proteins. *Int. J. Biol. Macromol.* 1989, *11* (1), 56–58.

(476) Flower, D. R.; North, A. C. t.; Attwood, T. K. Structure and Sequence Relationships in the Lipocalins and Related Proteins. *Protein Sci.* 1993, *2* (5), 753–761.

(477) Zhou, Y.; Jiang, L.; Rui, L. Identification of MUP1 as a Regulator for Glucose and Lipid Metabolism in Mice. *J. Biol. Chem.* 2009, *284* (17), 11152–11159.

(478) Chamero, P.; Marton, T. F.; Logan, D. W.; Flanagan, K.; Cruz, J. R.; Saghatelian, A.; Cravatt, B. F.; Stowers, L. Identification of Protein Pheromones That Promote Aggressive Behaviour. *Nature* 2007, *450* (7171), 899–902.

(479) Adams, D. *The Hitchhiker's Guide to the Galaxy: 1/5*, Reprints edition.; Pan, 2009.

(480) Grimus, W. On the 100th Anniversary of the Sackur-Tetrode Equation. *ArXiv11123748 Cond-Mat Physicsphysics Physicsquant-Ph* 2011.

(481) Paños Expósito, F. J. The Sackur-Tetrode Equation and the Measure of Entropy. 2014.

(482) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* 1958, *181* (4610), 662–666.

(483) Krishnan, V.; Rupp, B. Macromolecular Structure Determination: Comparison of X-Ray Crystallography and NMR Spectroscopy. In *eLS*; John Wiley & Sons, Ltd, 2001.

(484) Tzakos, A. G.; Grace, C. R. R.; Lukavsky, P. J.; Riek, R. Nmr Techniques for Very Large Proteins and Rnas in Solution. *Annu. Rev. Biophys. Biomol. Struct.* 2006, *35* (1), 319–342.

(485) Robustelli, P.; Stafford, K. A.; Palmer, A. G. Interpreting Protein Structural Dynamics from NMR Chemical Shifts. *J. Am. Chem. Soc.* 2012, *134* (14), 6365–6374.

(486) Wyman, J.; Allen, D. W. The Problem of the Heme Interactions in Hemoglobin and the Basis of the Bohr Effect. *J. Polym. Sci.* 1951, 7 (5), 499–518.

(487) Garman, E. F. Developments in X-Ray Crystallographic Structure Determination of Biological Macromolecules. *Science* 2014, *343* (6175), 1102–1108.

(488) Collier, G.; Ortiz, V. Emerging Computational Approaches for the Study of Protein Allostery. *Arch. Biochem. Biophys.* 2013, *538* (1), 6–15.

(489) Frauenfelder, H.; Petsko, G. A.; Tsernoglou, D. Temperature-Dependent X-Ray Diffraction as a Probe of Protein Structural Dynamics. *Nature* 1979, *280* (5723), 558–563.

(490) Yuan, Z.; Bailey, T. L.; Teasdale, R. D. Prediction of Protein B-Factor Profiles. *Proteins Struct. Funct. Bioinforma.* 2005, *58* (4), 905–912.

(491) Weik, M.; Colletier, J.-P. Temperature-Dependent Macromolecular X-Ray Crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 2010, *66* (4), 437–446.

(492) Neutze, R.; Moffat, K. Time-Resolved Structural Studies at Synchrotrons and X-Ray Free Electron Lasers: Opportunities and Challenges. *Curr. Opin. Struct. Biol.* 2012, *22* (5), 651–659.

(493) Miller, R. J. D. Femtosecond Crystallography with Ultrabright Electrons and X-Rays: Capturing Chemistry in Action. *Science* 2014, *343* (6175), 1108–1116.

(494) Carroll, L. *The Annotated Alice: The Definitive Edition Upd Sub Edition by Lewis Carroll*, 4th ed.; Tenniel, J., Ed.; W. W. Norton & Company, 2000.

(495) Ben-Zvi, P. Lewis Carroll and the Search for Non-Being by Pinhas Ben-Zvi http://www.the-philosopher.co.uk/alice.htm (accessed Apr 23, 2015).

(496) Palmer, J. A. *Parmenides and Presocratic Philosophy*; Oxford University Press: Oxford [England] ; New York, 2009.

(497) Kemeny, J. G. *Philosopher Looks at Science*; Van Nost.Reinhold,U.S.: New York, 1959.

(498) Perusco, L.; Michael, K. The Importance of Scenarios in Evaluating the Socio-Ethical Implications of Location-Based Services. 2006.

(499) Morgan, T. *Literary Outlaw: The Life and Times of William S. Burroughs*, Reprint edition.; W. W. Norton & Company: New York, 2012.

(500) Myszka, D. G.; Abdiche, Y. N.; Arisaka, F.; Byron, O.; Eisenstein, E.; Hensley, P.; Thomson, J. A.; Lombardo, C. R.; Schwarz, F.; Stafford, W.; others. The ABRF-MIRG'02 Study: Assembly State, Thermodynamic, and Kinetic Analysis of an Enzyme/inhibitor Interaction. *J. Biomol. Tech. JBT* 2003, *14* (4), 247.

(501) Martin, M. Martin McKenna. The Grasping Goblin! http://www.martinmckenna.net/article.asp?article=73 (accessed Jan 1, 2016).