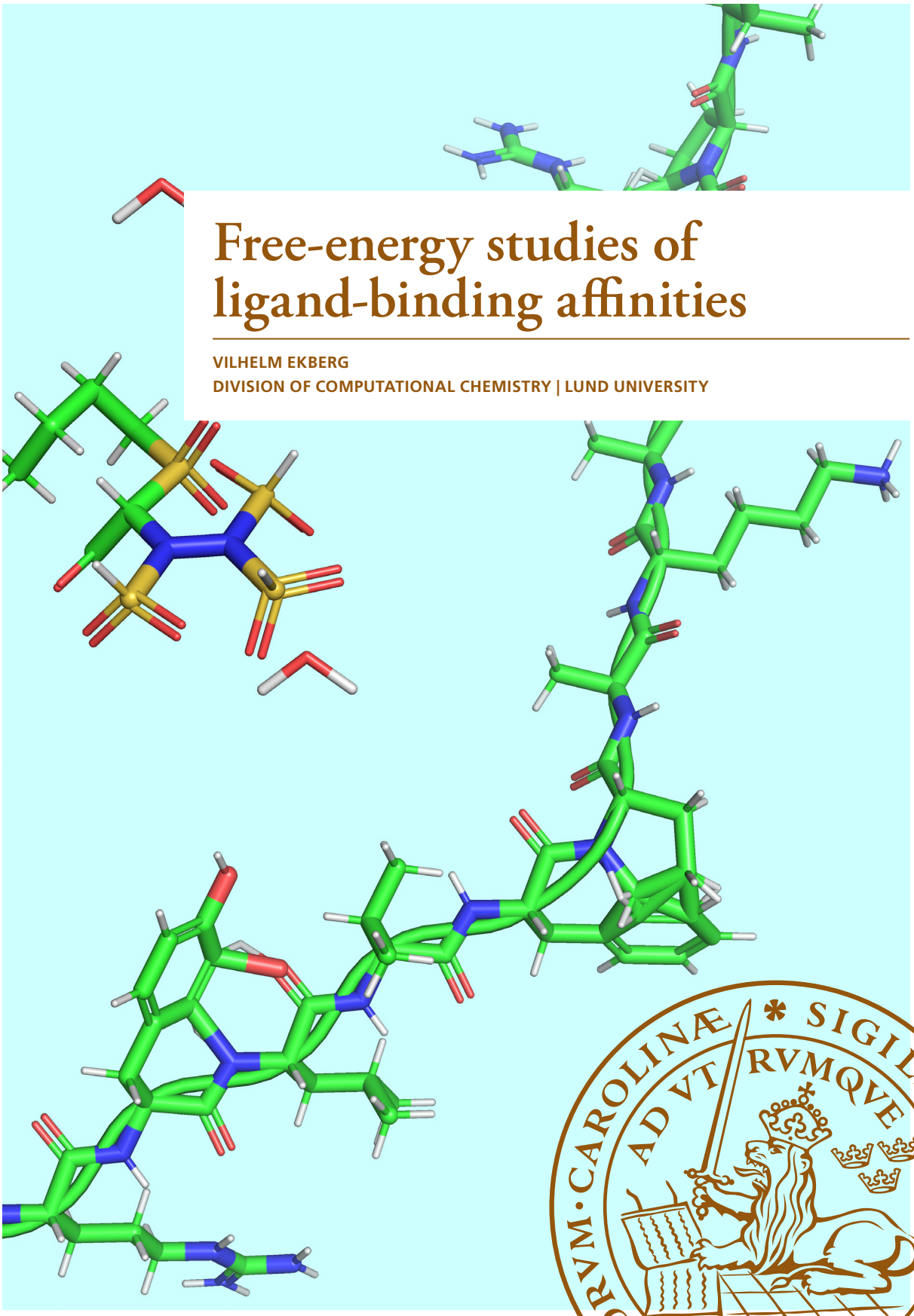
Free-energy studies of ligand-binding affinities

Ekberg, Vilhelm

2023



Free-energy studies of ligand-binding affinities

VILHELM EKBERG
DIVISION OF COMPUTATIONAL CHEMISTRY | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

ISBN: 978-91-7422-962-2

Division of Computational Chemistry
Department of Chemistry
Faculty of Science
Lund University

Free-energy studies of ligand-binding affinities

Free-energy studies of ligand-binding affinities

by Vilhelm Ekberg

LUND
UNIVERSITY

Thesis for the degree of Doctor
Thesis advisor: Prof. Ulf Ryde
Faculty opponent: Prof. Peter Coveney

To be presented, with the permission of the Division of Computational Chemistry of Lund University, for public criticism at KC:A on Tuesday, the 26th of September 2023 at 13:00.

Organization
LUND UNIVERSITY
Department of Chemistry
Box 124
SE–221 00 LUND
Sweden

Document name
DOCTORAL DISSERTATION
Date of disputation
2023-09-26
Sponsoring organization

Author(s)
Vilhelm Ekberg

Title and subtitle
Free-energy studies of ligand-binding affinities

Abstract
In drug discovery, it is of utmost importance to accurately calculate the free energies of binding ligands to various protein targets, such as enzymes and receptors. We have assessed and used computational tools for this aim, most of them based on molecular dynamics (MD) simulations. We mostly used molecular mechanics (MM) in order to model the protein–ligand interactions, which is more approximate than quantum-mechanical (QM) methods, but necessary to reduce the computational cost when doing calculations on protein–ligand systems, which often contain tens of thousand of atoms.
In one study of a large set of protein–ligand complexes, we tried to improve the free energies of binding by using MD simulations with QM-derived charges, which sometimes led to improved results, but not always. We also ran QM/MM simulations on casein-kinase 2 (CK_2), where the ligand and a few surrounding residues were treated at the QM level, and the rest of the system at the MM level. However, those results were unsatisfying. Furthermore, it is important and challenging to accurately model the large entropic contribution to ligand-binding free energies. This entropy largely stems from the fluctuation of the protein and ligand. We tried to estimate this entropy with methods based on fluctuations of interaction energies. We also saw how a combination of theoretical and experimental methods can shed light on phenomena like entropy–entropy compensation and halogen bonding. Additionally, we compared how MD and grand-canonical Monte Carlo (GCMC) can be used to assess dynamics and thermodynamics of protein–ligand binding for both buried and solvent-exposed binding sites.

Key words
Free energy perturbation, drug design, ligand-binding affinity, entropy, molecular mechanics, molecular dynamics

Classification system and/or index terms (if any)

Supplementary bibliographical information
Language
English

ISSN and key title
ISBN
978-91-7422-962-2 (print)
978-91-7422-963-9 (pdf)

Recipient's notes
Number of pages
154
Price
Security classification

Signature _____ Date 2023-08-09

Free-energy studies of ligand-binding affinities

by Vilhelm Ekberg

LUND
UNIVERSITY

Cover illustration front: Figure by Vilhelm Ekberg.

Funding information: The thesis work was financially supported by the Swedish Research Council.

Division of Computational Chemistry, Department of Chemistry

isbn: 978-91-7422-962-2 (print)
isbn: 978-91-7422-963-9 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2023

Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Contents

List of publications iii
Publications not included in this thesis v
Acknowledgements vi
Abbreviations vii
Populärvetenskaplig sammanfattning på svenska viii

1 Introduction **1**
1.1 Drug discovery and development 1
1.2 Thermodynamics of protein–ligand interactions 2
1.3 Different types of protein–ligand interactions 3
1.4 Protein systems 5
1.4.1 Ferritin 5
1.4.2 Galectin-3C 6
1.4.3 Casein-kinase 2 6

2 Modelling of molecular systems **9**
2.1 Quantum mechanical methods 9
2.1.1 Hartree–Fock theory 9
2.1.2 Density-functional theory 11
2.2 Molecular-mechanics methods 12
2.3 QM/MM methods 13

3 Sampling **17**
3.1 Molecular dynamics 17
3.2 Monte Carlo 18

4 Calculating free energies of ligand binding **21**
4.1 Free energy perturbation 22
4.2 Free energies at the QM/MM level 23
4.3 End-point methods 24

5 Thermodynamics of biomolecular systems **27**
5.1 Entropy of protein–ligand complexes 27
5.2 Thermodynamics of the solvent 28
5.2.1 Inhomogeneous solvation theory 28

5.2.2 Grid inhomogeneous solvation theory 30

6 Summary of papers **33**
6.1 Paper I 33
6.2 Paper II 35
6.3 Paper III 36
6.4 Paper IV 37
6.5 Paper V 39
6.6 Paper VI 41

7 Conclusions and Outlook **43**

References **45**

Scientific publications **51**
Author contributions 51
Paper I: Attempts to improve alchemical relative binding free-energy simulations by quantum-mechanical charges 53
Paper II: Entropy–Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C 63
Paper III: Comparison of Grand Canonical and Conventional Molecular Dynamics Simulation Methods for Protein-Bound Water Networks . 83
Paper IV: Halogen Bond Interactions and Solvation in Protein–Ligand Binding: Progressive Changes in Binding Thermodynamics Across a Series of Halogen-Substituted Ligands 99
Paper V: On the Use of Interaction Entropy and Related Methods to Estimate Binding Entropies 113
Paper VI: QM/MM binding-affinity calculations in proteins with the reference-potential approach 129

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Attempts to improve alchemical relative binding free-energy simulations by quantum-mechanical charges**
V. Ekberg, L. Cao, M. M. Ignjatović, M. A. Olsson, U. Ryde
Manuscript

II **Entropy–Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C**
J. Wallerstein, **V. Ekberg**, M. M. Ignjatović, R. Kumar, O. Caldararu, K. Peterson, S. Wernersson, U. Brath, H. Leffler, E. Oksanen, D. T. Logan, U. J. Nilsson, U. Ryde, M. Akke
JACS Au, 1(4), 484-500, **2021**

III **Comparison of Grand Canonical and Conventional Molecular Dynamics Simulation Methods for Protein-Bound Water Networks**
V. Ekberg, M. L. Samways, M. M. Ignjatović, J. W. Essex, U. Ryde
ACS Phys. Chem. Au, 2(3), 247-259, **2022**

IV **Halogen Bond Interactions and Solvation in Protein–Ligand Binding: Progressive Changes in Binding Thermodynamics Across a Series of Halogen-Substituted Ligands**
M. L. Verteramo, M. M. Ignjatović, R. Kumar, S. Wernersson, **V. Ekberg**, J. Wallerstein, G. Carlström, V. Chadimová, H. Leffler, F. Zetterberg, D.T. Logan, U. Ryde, M. Akke, U. J. Nilsson
Manuscript

V **On the Use of Interaction Entropy and Related Methods to Estimate Binding Entropies**
V. Ekberg, U. Ryde
J. Chem. Theory Comput., 17(8), 5379-5391, **2021**

VI **QM/MM binding-affinity calculations in proteins with the reference-potential approach**
V. Ekberg, M. Wang, M. M. Ignjatović, U. Ryde
Manuscript

All papers are published with open access.

Publications not included in this thesis

VII **Exploring ligand dynamics in protein crystal structures with ensemble refinement**
O. Caldararu, **V. Ekberg**, D. T. Logan, E. Oksanen, U. Ryde
Acta Crystallogr. D, 77(8), 1099-1115, **2021**

VIII **Force-field problems to reproduce hydrogen-bond networks of ligands binding to proteins**
V. Ekberg, U. Ryde
Manuscript

Acknowledgements

I wish to start by acknowledging my supervisor **Ulf**, without whom I would never have gotten the chance to make this thesis, and from whom I learned how to develop as a computational scientist over the course of my PhD studies. It has been four very substantial and exciting years under your supervision. I also want to acknowledge my co-supervisor **Mikael**, with whom we made some really intriguing research on galectin-3C.

Big thanks also to the **Division of Computational Chemistry**, and to my fellow **group members**, both current and former, in particular. This thesis would not have been made if not for the long-time, terrific company of my former office mate **Justin**, the wisdom and amusing anecdotes of my current office mate **Marcos**, the friendliness of **Magne**, the humor and knowledge of **Kristoffer**, a nice stay at a conference in Gothenburg with **Hao** and all the interesting and fun chats with **Ernst**, **Joel**, **Eric**, **Victor**, **Simon**, **Mickaël** and my first office mate **Erik**.

Lastly, I thank my **parents** and **siblings**, for generously supporting me and encouraging me in my endeavour over these four years.

Abbreviations

AMBER	Assisted Model Building with Energy Refinement
BAR	Bennett acceptance ratio
C2	second-order cumulant approximation
CK2	casein-kinase 2
DFT	density-functional theory
FEP	free energy perturbation
GAFF	general AMBER force field
GCMC	grand-canonoical Monte Carlo
GCI	grand-canonocial integration
GIST	grid inhomogeneous solvation theory
HF	Hartree–Fock
IE	interaction entropy
IST	inhomogeneous solvation theory
ITC	isothermal titration calorimetry
LJ	Lennard-Jones
LCAO	linear combination of atomic orbitals
MC	Monte Carlo
MD	molecular dynamics
MBAR	multistate Bennett acceptance ratio
MM	molecular mechanics
MMPBSA	molecular mechanics Poisson–Boltzmann surface area
MMGBSA	molecular mechanics generalised Born surface area
NM	normal-mode
NMR	nuclear magnetic resonance
PBC	periodic boundary conditions
QM	quantum-mechanical
RESP	restrained electrostatic potential
RMSD	root-mean-square deviation
ROI	region of interest
RPQS	reference-potential with QM/MM sampling
SASA	solvent-accessible surface area
TI	thermodynamic integration
VDW	Van der Waals

Populärvetenskaplig sammanfattning på svenska

Läkemedelsindustrin är omfattande, och att tillverka ett läkemedel innefattar många olika steg, alltifrån att först hitta en tänkbar kandidat, vilket kan göras genom att studera vad andra forskare har gjort eller medicinsk historia, att studera biverkningar, testa stora bibliotek av tänkbara läkemedelskandidater, och utföra olika former av datorstudier. Man kan uföra test på vävnad eller enskilda celler, s.k. *in vitro*, eller djurförsök, *in vivo*, vilket såklart innebär olika etiska utmaningar. Experimentella tekniker, såsom att använda röntgenstrålning för att få fram kristallstrukturer, är ofta av intresse. När man har en bra läkemedelskandidat, så ska dess kemiska struktur finslipas experimentellt och testas i flera olika kliniska studier med successivt större grupper av patienter. Allt som allt tar det ofta mer än 10 år att tillverka ett läkemedel idag, och det kostar miljardtals kronor. Man bör därför, så gott det går, försöka försäkra sig om att den läkemedelskandidat man går vidare med verkligen kommer att bli framgångsrik. Det är här som datorer kommer in i bilden.

Den mänskliga kroppen är väldigt komplex, så datorer kommer nog aldrig att kunna ersätta kliniska studier, iallafall inte helt. Det som vi fokuserar på i den här avhandlingen är att analysera hur olika sorters läkemedelskandidater binder till olika proteiner, till exempel enzymer eller receptorer som finns i den mänskliga kroppen. Det är när läkemedlet binder till dessa mål som de får en effekt i kroppen, och det är därför som datorer är ett oumbärligt komplement till experimentella metoder när det kommer till läkemedelstillverkning. Ju starkare bindning, desto bättre läkemedelskandidat, som vi för enkelhetens skull kommer att kalla för *ligand* i fortsättningen. Att bedöma hur bra ligander binder handlar om att utvärderar något som kallas för *fria energier*, vilket i sin tur består av en energisk del, kallad *entalpi*, samt en del som mäter hur stor *entropin* är. Entalpin består av många olika bidrag, bland annat attraktion mellan laddningar med samma tecken och repulsion mellan laddningar med olika tecken. Entropin kommer av det faktum att atomer och molekyler, stora som små, är fria att röra sig i tid och rum, vilket innebär att både proteiner och ligander har en viss flexibilitet. Man kan säga att att entropin är ett mått på oordningen i systemet, och att utvärdera entropin är ofta svårt.

För att modellera ett system med ett protein och en ligand, så använder man sig ofta av något som kallas för *molekylmekanik* (MM). Här betraktar man alla atomer som små bollar, som hålls ihop av fjädrar. Metoden är inte lika noggrann som *kvantmekanik* (QM), där man betraktar elektronerna explicit, men mycket snabbare; med MM kan man behandla tiotusentals atomer medan QM metoder bara kan hantera hundratals atomer. När man simulerar dynamiken i systemet, dvs. hur de olika atomerna och molekylerna rör sig i tid och rum, används en teknik som kallas för *molekyldynamik* (MD). Den är baserad på Newtons rörelseekvationer.

Vi har testat olika metoder för att beräkna fria energier för ligandbindning; dessa resultat jämförs alltid med experimentella resultat. Vi har testat att använda laddningar beräknade med QM i MM-simulationer, vilket fungerade ibland. Vi kunde också genom att kombinera experimentella och teoretiska metoder få en inblick i väldigt intressanta fenomen, som hur halogener påverkar ligandbindningen, samt hur olika sorters entropi kan kompensera varandra. Sammanfattningsvis är dessa typer av beräkningar fortsatt väldigt utmanande, även om det samtidigt finns mycket potential samt en del lovande tendenser.

Chapter 1

Introduction

1.1 Drug discovery and development

Drug discovery is a long and complex process, encompassing several stages, such as choosing a drug target, finding a 'lead compound', improving pharmaceutical properties, and carrying out clinical trials. The whole process, from start to end, may take more than 10 years and cost more than 4 billion dollars. Drugs are manufactured to target a broad range of diseases, such as depression, cancer, flu, cancer, migraine, etc. The first step is to identify a drug target, which can be a receptor, enzyme or nucleic acid. One strives for optimizing the target specificity and selectivity, in order to ensure that only the target in question is affected, and unwanted side effects are minimsed.

In early stages of drug development, when trying to identify a lead compound, a large number of compounds is usually synthesized and tested *in vitro* (on isolated cells or tissues, etc.), or *in vivo* (on animals). Testing on humans is not done at this stage, and *in vitro* tests are often preferred, as they are faster and do not affect animals. One may start from a natural ligand, screen natural materials (plants, marine world, venoms, etc.), assess existing drugs, consider medical folklore, or do a high-throughput screening. Nowadays, determining the atomistic structure of a compound is made easy by X-ray crystallography, cryogenic electron microscopy and nuclear magnetic resonance (NMR) spectroscopy. Synthesizing drugs, however, is often costly, so it is helpful to start from a natural product or a commercially available compound, and then modify it by conventional synthetic methods.

A cruical step is to identify important ligand–target interactions, which may be inferred from the molecular structure. For example, hydroxyl groups are important for

hydrogen bonds, so replacing them with a methyl group will weaken or perhaps even destroy that bond. Amides commonly form hydrogen bonds and aromatic rings are often involved in Van der Waals (VDW) interactions.

Additionally, it is important to understand how drugs are metabolised in the body, Polar drugs are usually secreted by the kidneys, and non-polar drugs are converted to more polar derivatives in the liver. Drug metabolites may be harmful, though, and it is important to test the metabolites on humans and animals. In general, toxicity studies *in vitro* and *in vivo* must be done before proceeding with clinical test phases I–III involving successively larger groups of people. Phase I involves tests on healthy volunteers, to see if the drug has any serious side effects. Phase II involves tests on a small group of patients, in order to test if the drug has the desired effect, and adjust the dose level. In Phase III, the drug is tested on a much larger group of patients. In order to determine whether the drug is truly effective, placebo groups may be involved, and the drug may be compared to other available treatments. The dose level is further optimized in phase III, and the larger sample of patients allows for more side effects to be identified. If these clinical phases are successful, it is time to start marketing the drug. Once the drug is released to the marker, the drug is continuously monitored in phase IV, where rare side effects may be found several years after the drug has been introduced to the market. [1]

1.2 Thermodynamics of protein–ligand interactions

Most cases in this thesis involve a ligand (L) binding to a protein target (P):

$$P + L \rightleftharpoons PL \tag{1.1}$$

This is a reversible reaction that continues until it reaches equilibrium. From the equilibrium concentrations, one may define a binding constant K_b (units of M^{-1})

$$K_b = \frac{[PL]}{[P][L]} \tag{1.2}$$

The free energy (ΔG) is the driving force for the process in Eq. 1.1; it must be negative for the binding of the ligand to the protein to be spontaneous. The free energy is related to the binding constant via the Gibbs relationship:

$$\Delta G = -RT \ln K_b \tag{1.3}$$

where T is the absolute temperature and R is the universal gas constant (8.314 J/mol/K)

The free energy is related to the enthalpy (ΔH) and entropy (ΔS) via the following equation:

$$\Delta G = \Delta H - T\Delta S \tag{1.4}$$

ΔH measures the change in enthalpy when the ligand binds to the protein. It is the net result from the formation and disruption of many individual interactions between the protein and the ligand, and between the ligand and solvent, as well as reorganization of solvent molecules near the protein and ligand surfaces.

ΔS is the entropy of the binding process. It may be divided into several contributions:

$$\Delta S = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t} \tag{1.5}$$

Where ΔS_{solv} comes from the changes in solvent entropy, such as release of solvent upon binding of the ligand, which often results in a positive entropy contribution. ΔS_{conf} reflects the change in conformational freedom of the ligand and protein during the binding process and $\Delta S_{r/t}$ accounts for the loss of rotational and translational entropy of the ligand and protein during the binding process; since two entities are turned into one in the binding reaction, this contribution is negative.

There are different models for describing the binding of a ligand to a protein. The *lock-and-key* model suggests that the protein and ligand are both rigid, and may only bind to each other if the ligand provides a perfect fit. However, the *induced fit* model accounts for the case when the protein and ligand do not perfectly fit each other. In this model, the protein and ligand are allowed to undergo conformational changes during the binding process. Furthermore, in the *conformational selection* model, the protein is found in several conformations in equilibrium with each other. The ligand binds to the most favourable conformation, and shifts the equilibrium towards that conformation.[2]

1.3 Different types of protein–ligand interactions

A ligand can interact with a protein via various intermolecular interactions. When non-polar molecules are in close proximity, there are temporary fluctuations in the electron density, leading to the formation of induced dipoles. These induced dipoles fluctuate in a coordinated way and give rise to attractive Van der Waals (VDW) interactions, and the larger the species, the stronger the interactions. These interactions are also referred to as *London forces*.

For polar molecules, attractive *dipole–dipole interactions* may also occur. This is due to the attraction between groups with a surplus of opposite charge density, for in-

stance between the δ^- dipole of an O atom to the δ^+ dipole of a C atom. Here, the permanent dipoles align in an antiparallel fashion. For polar molecules, VDW interactions are also in effect in addition to the dipole–dipole interactions. A permanent dipole may also interact with a non-polar group via so-called *dipole–induced dipole interactions*.

When a hydrogen atom is bound to O, S, N or F, the polarization of the bond is especially pronounced, which allows for a particularly strong form of dipole–dipole interaction, known as *hydrogen bonding*. Hydrogen bonding is important for the tertiary and quarternary structure of proteins. When two peptide units are close to each other, the carbonyl oxygen of one unit forms a hydrogen bond to the proton on the amide nitrogen of the second unit.

Water is a common solvent in biomolelcular interactions; the molecules often form hydrogen bonds with the protein and the ligand. This may result in the formation of water bridges between the protein and ligand. Water also gives rise to dielectric screening, since the polar water molecules screen the solvated protein–ligand complex from surrounding electrostatic interactions interactions.[3,4] Often, some water molecules are bound to the protein at fixed locations, while most water molecules move freely in the solution. Another consequence of the hydrogen bonding is the *hydrophobic effect*, in which the hydrophobic parts of amphiphilic molecules cluster together to minimize the contacts with water molecules.[5]

In this thesis, *halogen bonds* are also explored; it a non-covalent interaction between a donor halogen atom (Cl, Br or I) and an acceptor Lewis base. A halogen atom generally has an $s^2p_x^2p_y^2p_z^1$ configuration, where the electron-poor z-axis, i.e. δ^+, forms a so-called σ-hole. The neighbouring lone pairs form an electronegative belt, which may serve as a hydrogen bond acceptor perpendicular to the σ-hole. The extent of the σ-hole increases as the size and polarizability of the halogen atom increases (Cl < Br < I), F only has an insignificant σ-hole.[6,7,8]

In addition, there may be *ionic interactions*, which occur between charged species, such as carboxylate and ammonium ions, as well as *ion–dipole interactions* between permanent dipoles and ions. There is also the *cation–π interaction* between an electron-rich π system, such as benzene or ethylene, and a cation. The cation–π interaction has a strength comparable to the hydrogen bond in solution, and plays an important role in protein structure, molecular recognition and enzyme catalysis.[9,10]

In protein–ligand binding, it is not just the interactions between the protein and the ligand that are important. Equally important are the interactions that are present before the formation of the protein–ligand complex, such as the interactions between the protein and the solvent, as well as between the ligand and the solvent. It is the sum of the changes in all of these interactions that determine the net enthalpy change of

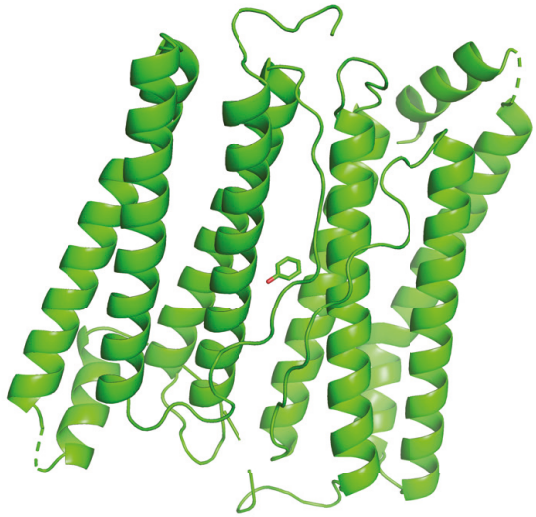


Figure 1.1: Picture of a dimer of ferritin with a bound phenol ligand (monomer PDB ID: 3F39 [12], dimer made with Maestro [13]). Made in Pymol.

the binding process. And moreover, if the net enthalpy change is negative, the process may still not be spontaneous if there is a significant entropic penalty (often around 40–60 kJ/mol) [11] associated with the binding.

1.4 Protein systems

In this section, attention will be devoted to three of the protein systems studied in this thesis: ferritin, galectin-3C and casein-kinase 2 (CK2).

1.4.1 Ferritin

Ferritin (fig. 1.1) is a 24-mer protein which may bind iron, and thereby function as a iron reservoir in the body, preventing iron levels in the cell from becoming toxic. Each subunit has a molecular mass of roughly 20 kDa. The 24-mer forms a hollow, roughly spherical structre, and the additional, smaller cavities at the dimer interfaces are potential targets of anesthetics. [12]

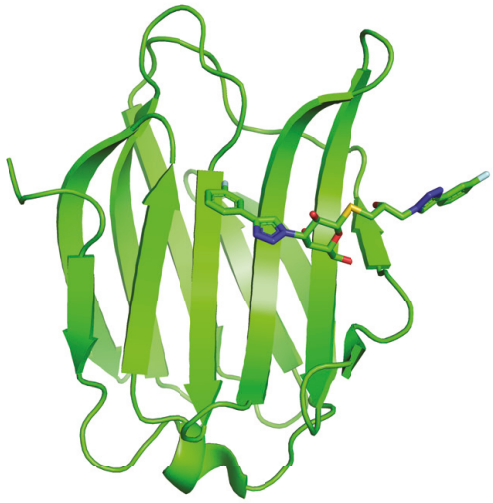


Figure 1.2: Picture of galectin-3C in complex with (2R)-2-hydroxy-3-(4-(3-fluorophenyl)-1H-1,2,3-triazol-1-yl)-propyl) 2,4,6-tri-O-acetyl-3-deoxy-3-(4-(3-fluorophenyl)-1H-1,2,3-triazol-1-yl)-1-thio-β-d-galactopyranoside (PDB ID: 6QGF[14]). Made in Pymol.

1.4.2 Galectin-3C

Galectin-3 contains several subunits, but we will focus on the carbohydrate-recognition domain, called galectin-3C. The protein is a member of the galectin family of mammalian lectins, which are carbohydrate-binding proteins. Galectin-3C (fig. 1.2) has a solvent-exposed binding site, placed in a shallow groove, with water molecules bridging between the ligand and protein. Galectins are attractive targets for treating cancer and inflammations due to their role in cell growth, cell differentiation, cell-cycle regulation, signaling and apoptosis.[14]

1.4.3 Casein-kinase 2

Kinases play a central role in signal transduction in cells by catalyzing the transfer of the γ phosphate group of ATP to serine, threonine or tyrosine residues of various protein substrates. In the human genome, more than 500 different kinases are encoded. Kinases are potential targets for drugs aiming to combat tumor growth. The catalytic subunits of CK2 (α and α^-) are active either alone or in the form of a heterotetrameric holoenzyme (fig. 1.3).[15]

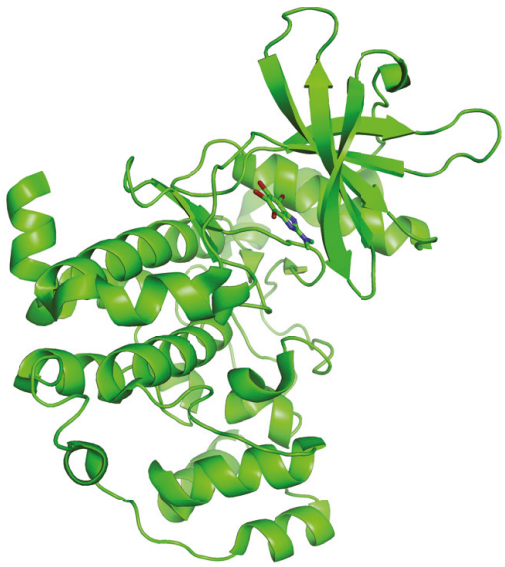


Figure 1.3: Picture of CK2 in complex with a tetrabromo-benzimidazole inhibitor (PDB ID: 1ZOE[15]). Made in Pymol.

Chapter 2

Modelling of molecular systems

In this chapter, we will look into some different ways of modelling a molecular system. When it comes to modelling protein–ligand interactions, there are relatively simple methods such as docking.[16] With this method one can roughly analyze how the ligand fits into the binding site, but the method cannot accurately account for the protein flexibility.[17] Therefore, more advanced methods are needed in order to improve the accuracy. We will assess both quantum-mechanical (QM) methods as well as the less accurate but faster molecular-mechanics (MM) method.

2.1 Quantum mechanical methods

2.1.1 Hartree–Fock theory

Hartree–Fock (HF) theory starts from the time-independent Schrödinger equation:

$$\mathbf{H}\Psi = E\Psi \tag{2.1}$$

where $\mathbf{H}$ is the Hamiltonian operator, Ψ is the wavefuncion and E is the energy of the state in question. HF is an example of an *ab initio* method, meaning that it does not involve any experimental data. The HF method is based on the approximation that each electron is affected by the average interactions from all the other electrons; this means that *electron correlation*, i.e. that space–time variations in the interactions, are neglected. More approximations are also involved for the HF method. The *Born–Oppenheimer* approximation accounts for the fact that the atom nuclei are much heavier than the electrons, meaning that the electrons will move much faster

and immediately adapt to any changes in the nuclei positions. Relativistic effects are ignored as well.

The next step is to introduce *Slater determinants*, where ϕ are one-electron wave functions, while $1, 2, ..., N$ represent electronic coordinates. We use one-electron functions that are the product of a spatial molecular orbital (ψ) and a spin function (α or β). So, for N electrons and N molecular orbitals, the Slater determinant is as follows:

$$\Phi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \dots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \dots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \cdots & \phi_N(N) \end{vmatrix} \; ; \; \langle \phi_i | \phi_j \rangle = \delta_{ij} \tag{2.2}$$

Here, we use the Dirac notation to denote integrals, and the expression to the right of the semi-colon denotes that the orbitals are orthonormal.

We will consider each molecular orbital to be a linear combination of basis functions, which are orbitals centered on the atoms. This is expressed in the form of a linear combination of atomic orbitals (LCAO):

$$\psi_i = \sum_{\alpha} c_{i\alpha} \chi_\alpha \tag{2.3}$$

where χ_α are the atomic orbitals (such as Gaussian-type orbitals, for instance), which together form the so-called basis set, and the $c_{i\alpha}$ coefficients are determined via an iterative procedure explained below.

Since the wave function of an electron depends on all the other electrons, due to the mutual interactions, the HF energy must be computed via an iterative procedure. If we neglect electron correlation, we may assume that the trial wave function is a single determinant. By seeking to minimize the energy by the *variational principle*:

$$E = \frac{\langle \Psi | \mathbf{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \tag{2.4}$$

the HF equations may be obtained. For a closed-shells system, the practical starting point is the *Roothaan–Hall* equations:

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{2.5}$$

$$F_{\alpha\beta} = \langle \chi_\alpha | \mathbf{f} | \chi_\beta \rangle \tag{2.6}$$

$$S_{\alpha\beta} = \langle \chi_\alpha | \chi_\beta \rangle \tag{2.7}$$

where the **S** matrix accounts for the overlap between basis functions, and the Fock matrix, **F**, contains the Fock matrix elements. The Fock operator, **f**, for electron i is given as:

$$\mathbf{f}_i = \mathbf{h}_i + \sum_{j=1}^{m} (\mathbf{J}_j - \mathbf{K}_j) \tag{2.8}$$

where

$$\mathbf{h}_i = -\frac{1}{2}\nabla_i^2 - \sum_{k=1}^{N} \frac{Z_k}{|\mathbf{R}_k - \mathbf{r}_i|} \tag{2.9}$$

where $\mathbf{h}_i$ is the one-electron Hamiltonian operator describing the kinetic energy of electron i, as well as attraction to all the nuclei. Also, m is the total number of electrons, N is the total number of nuclei, Z_k is the atomic number of nucleus k and $\mathbf{R}_k$ and $\mathbf{r}_i$ are the positional vectors of nucleus k and electron i, respectively. $\mathbf{J}_j$ is the Coulomb operator and describes the electron–electron repulsion, while the exchange operator $\mathbf{J}_j$ gives the energy of exchanging two electrons.

In the *self-consistent field* method, the Roothaan–Hall equations serve as a starting point. One starts off by guessing the coefficients in **C**, then forming and diagonalizing the **F** matrix. Next, one calculates new coefficients, constructs a new Fock matrix from them, and this cycle is carried on until convergence is achieved to within some tolerance.

In order to reduce the computational effort, one may use *semi-empirical methods*. Here, only valence electrons are treated explicitly and the combined repulsion from the nuclei and core electrons may be modelled by functions. A minimum basis set, i.e. one basis function for each atomic orbital, is used. Furthermore, many of the integrals involved in solving the HF equation are made into parameters based on experimental data or calculations.

Semi-empirical functions can make use of *hydrogen-bond correction*. Here the hydrogen-bond energy is computed from a function dependent on the distance and angle between the hydrogen and acceptor atom, as well as a damping function to correct the short- and long-range behavior.[18] Furthermore, the absence of electron correlation in semi-empirical methods also mean that VDW interactions cannot be modelled accurately unless a *dispersion correction* is also enforced, which is also based on distance- and angle-dependent functions.[19]

2.1.2 Density-functional theory

At the heart of density-functional theory (DFT) is the Hohenberg–Kohn theorem, which states that the ground-state electronic energy may be determined from the

electron density ρ.[20] This means that a functional exists for computing the energy from the electron density, but the issue is that the exact form of the functional is not known. In orbital-free DFT, one starts by dividing the energy functional into the kinetic energy, $T[\rho]$, electron–nuclei attraction, $E_{ne}[\rho]$, and electron–electron repulsion, $E_{ee}[\rho]$.

One big problem in DFT is finding a good representation of the kinetic energy. In Kohn–Sham theory, orbitals are re-introuced, and the functional depicting the kinetic energy is divided into one part that can be computed exactly, and a correction term. This gives for the DFT energy:

$$E_{DFT}[\rho] = T_S[\rho] + E_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \tag{2.10}$$

where $T_S[\rho]$ is the exact kinetic energy functional for non-interacting electrons. $J[\rho]$ is the Coulomb functional, which is computed classically, just like $E_{ne}[\rho]$. In Kohn–Sham theory, the problem instead reduces to finding an approximate form of the exchange–correlation functional, $E_{xc}[\rho]$.[16]

One problem with DFT is that the energy is a functional of the single-particle electron density, meaning that two-particle interactions cannot be distinguished from from self-interactions. This gives rise to the so-called *self-interaction error*, which does not exist for wave function methods, such as HF. In practice, however, the error is somewhat reduced for DFT when the exchange–correlation functional is optimized to reproduce experimental data. Another option is to introduce HF-exchange to the exchange–correlation functional,[21] giving rise to so-called hybrid functionals.

2.2 Molecular-mechanics methods

When modelling protein–ligand interactions, quantum-mechanical methods are often too expensive. Instead classical methods in the form of MM are typically used, as they allow for tens of thousands of atoms to be modelled. The MM software that we primarily used in this thesis is Assisted Model Building with Energy Refinement (AMBER).[22] Flexibility of both the protein and ligand can be accounted for with MM methods, and explicit solvent effects may be included.[23]

The typical form for a biomolecular force field in MM theory is composed of the bonded terms

$$E_{bond} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{torsions} K_\phi[1 + \cos(n\phi - \delta)] \tag{2.11}$$

which account for fluctuations of of the bond distance (b), bond angle (θ) and dihedral angle (ϕ). The stretch and bend terms are modelled as harmonic oscillators, with associated force constants (K_b, K_θ) and equilibrium values (b_0, θ_0) values. The dihedral term is a sum of cosine functions with amplitudes K_Φ, periodicity $n = 1, 2, 3...$, and phases δ.

The non-bonded terms are described by sums over all pairs of atom–atom interactions:

$$E_{nb} = \sum_{i<j} 4\epsilon_{ij} \left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_r \epsilon_0 r_{ij}} \tag{2.12}$$

which treats VDW interactions in the form of the Lennard-Jones (LJ) 6–12 potential, and electrostatics in the form of Coulombic interactions. Here, ϵ_{ij} is the depth of the LJ potential-energy well, r_{ij} is the interatomic distance, and σ_{ij} is the interatomic distance at which the LJ potential energy is zero. The partial charges of the respective atoms are given by q_i and q_j, ϵ_r is the relative permittivity of the medium and ϵ_0 is the permittivity of vacuum. This point-charge model does not include polarisation, which means that the force field charges strictly should be re-derived for each dielectric medium.[23,24]

The computation of the non-bonded terms takes up most of the time in MM calculations. All the parameters in Eqs. 2.11 – 2.12 should be optimized in order to reproduce experimental results as accurately as possible. One set of parameters is obtained for one set of molecules, and applying these parameters to another type of molecule is not recommended. The sets of parameters in MM methods are referred to as *force field*.[23] One drawback of MM methods is the neglect of electronic effects, which means that neither bond breaking nor bond formation can be modelled.[25]

In the thesis, we used the ff14SB force field for proteins,[26] while some different water models (TIP3P, TIP4P-Ew and OPC)[27,28,29] were used.

2.3 QM/MM methods

One way to combine the accuracy of QM with the speed of MM is the so-called QM/MM method. Here, a small part of the protein–ligand complex, typically the ligand and perhaps some resides interacting with it, is treated at the QM level (system 1), while the rest of the protein is treated at the MM level (system 2) (fig. 2.1).[30]

In AMBER, an additive scheme is used to compute the total QM/MM energy, E_{eff}:

$$E_{eff} = E_{QM1} + E_{MM2} + E_{QM1/MM2} \tag{2.13}$$

meaning that the energies of the QM and MM regions, E_{QM1} and E_{MM2}, are computed separately, with the interaction between the regions accounted for by the term $E_{QM/MM}$.[22,31]

Alternatively, a subtractive scheme can be used, where an MM energy is calculated for both regions (E_{MM12}) and added to the E_{QM1} energy. The MM energy of system 1 (E_{MM1}) is then subtracted to avoid double-counting:[31]

$$E_{eff} = E_{QM1} + E_{MM12} - E_{MM1} \tag{2.14}$$

By default, AMBER uses an *electronic embedding* scheme, in which a set of point charges, one for each MM atom, is included in the QM calculations, so that they polarise the QM electron density. If there are any covalent bonds between systems 1 and 2, AMBER truncates the QM system with atoms (hydrogen by default), the so-called *link-atom* approach.[22]

The additive method requires special MM software that can run QM/MM simulations, in which no interactions are omitted or double-counted, meaning that the user must select which MM terms to include. One advantage of the additive method is that the atoms in the QM region require no MM parameters. As for the subtractive method, standard QM and MM software may be used, and it also allow various link-atoms corrections to be introduced, although it introduces more MM parameters and often requires considerable effort. The two schemes should produce identical results if properly implemented, especially if there are no link-atoms.[31]

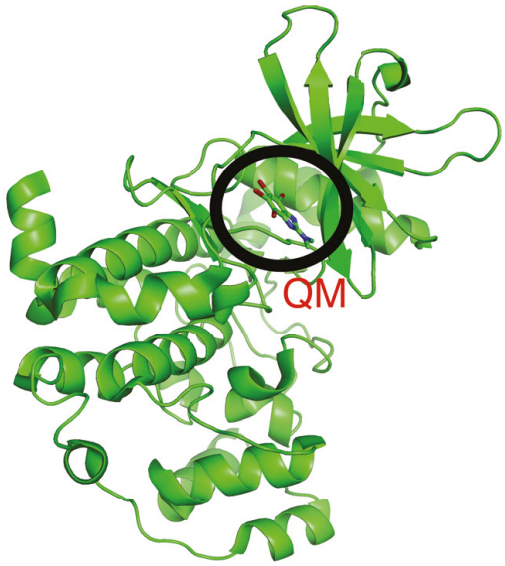


Figure 2.1: Illustration of the QM/MM method, with the QM region highlighted (PDB ID: 1ZOE [15]).

Chapter 3

Sampling

We have in the previous chapter assessed various methods to compute the potential energy of a protein–ligand system. These energy functions depend on the coordinates of the molecules in question, which vary over time. Both the protein and ligand are flexible and may adapt numerous different conformations. This highlights the need to efficiently and thoroughly sample the configurational space. We will assess two different techniques for that in this chapter.

3.1 Molecular dynamics

At the heart of molecular dynamics (MD) lies classical mechanics, which is derived from Newton's second law of motion

$$\mathbf{F} = m\mathbf{a}(t) = m\frac{d^2\mathbf{r}(t)}{dt^2} \tag{3.1}$$

which relates the acceleration, $\mathbf{a}(t)$, second derivative of the position $\mathbf{r}(t)$ with respect to time, of a particle with mass m to the force $\mathbf{F}$ experienced by the particle. The force is related to the potential energy function U via:

$$\mathbf{F} = -\frac{dU)}{d\mathbf{r}(t)} \tag{3.2}$$

Equation 3.1 is a second-order differential equation, which may be solved iteratively by first assigning initial positions and velocities, $\mathbf{v}$(t), to all particles. Over a small time step Δt (0.5–1 fs), the positions are updated according to a second-order truncation

of a Taylor expansion of the position function:

$$\mathbf{r}(t + \Delta t) \approx \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{\mathbf{a}(t)(\Delta t)^2}{2} \tag{3.3}$$

After obtaining the new positions of all atoms, a new set of energies and forces may be calculated. By repeating this procedure over and over, we obtain a trajectory showing how the positions and velocities of all the atoms vary over time.

MD simulations are usually quite demanding in terms of computational resources. One way to make the process more efficient is to use the SHAKE algorithm, where all bonds involving hydrogen atoms are constrained to their equilibrium values.[32] This enables for a longer time step of 2 fs. One can also ignore the non-bonded interactions beyond a certain cut-off distance, which is shorter for VDW interactions than for electrostatics.

Another problem with molecular simulations is that only a rather small number of atoms can be simulated (typically 10^4-10^6 atoms), whereas real systems contain $\sim 10^{23}$ atoms. A way to solve this problem is to simulate infinite systems by using periodic boundary conditions (PBC), in which an infinite amount of copies of the simulation box extends in all directions, so that a particle that leaves the box on one side enters it on the opposite side. For PBC simulations, long-range electrostatic interactions can be computed with Ewald summation.[33]

3.2 Monte Carlo

Monte Carlo (MC) methods make use of random sampling. One starts with a configuration of particles and randomly perturbs the coordinates of a particle, by for instance translating or rotating it.[16] In the Metropolis method,[34] the new set of coordinates is accepted if the resulting energy change, ΔU, is negative, and otherwise it is accepted if:

$$p < e^{-\Delta U / k_B T} \tag{3.4}$$

where p is a random number between 0 and 1. To allow for a reasonable acceptance ratio, the step size should not be too large. By conducting the above procedure N times, an ensemble average of the system X may be obtained:

$$\langle X \rangle \approx \frac{1}{N} \sum_{i=1}^{N} X_i \tag{3.5}$$

In this thesis, we are using a more sophisticated MC technique, the grand-canonical Monte Carlo (GCMC) method, as implement by Essex and coworkers.[35] This is useful for solvated protein–ligand systems with buried binding sites, where the exchange

of water molecules in the binding site with the surroundings is slow. In this method, a region of interest (ROI), e.g. the binding site, communicates with an ideal-gas bulk region of water molecules. In addition to ordinary MC moves, attempts are also made to insert or delete water molecule from the ROI (containing N water molecules) with acceptance probabilities given by:

$$P_{insert} = \min\left[1, \frac{1}{N+1} e^{B} e^{-\Delta U/k_B T}\right] \tag{3.6}$$

$$P_{delete} = \min\left[1, N e^{-B} e^{-\Delta U/k_B T}\right] \tag{3.7}$$

Here, B is the Adams parameter,[35,36] given by:

$$B = \frac{\mu}{k_B T} + \ln\left(\frac{V_{ROI}}{\lambda^3}\right) \tag{3.8}$$

where μ is the chemical potential, V_{ROI} is the volume of the ROI, and λ is the thermodynamic de Broglie wavelength of water. For a particle of mass m, the thermodynamic de Broglie wavelength is given by:[37]

$$\lambda = \frac{1}{\sqrt{2\pi m k_B T}} \tag{3.9}$$

Performing GCMC simulations at different Adams values will result in a titration curve for the average number of inserted water molecules as a function of B. From this curve, one can compute the binding free energy of when going from N_i to N_f water molecules in the ROI:

$$\begin{aligned}\Delta G_{bind}(N_i \to N_f)/k_B T = {} & N_f B_f - N_i B_i \\ & - (N_f - N_i)\left[\frac{u'_{solv}}{k_B T} + \ln\left(\frac{V_{ROI}}{V^0}\right)\right] - \int_{B_i}^{B_f} N(B) dB\end{aligned} \tag{3.10}$$

where B_i to B_f are the Adams parameters that on average produces N_i to N_f water molecules at equilibrium, respectively, u'_{solv} is the excess chemical potential of water, and V^0 is the standard-state volume of water. This technique 3.10 is referred to as grand-canonical integration (GCI). In order to obtain $N(B)$, the GCMC titration data is fitted to a logistic formula:

$$N(B) = \sum_{i=1}^{k} \frac{n_i}{1 + e^{\omega_{0i} - \omega_i B}} \tag{3.11}$$

where k, n_i, ω_{0i} and ω_i are fitted parameters.

Chapter 4

Calculating free energies of ligand binding

Many of the problems in this thesis concern the calculation of the difference in free energy when binding two different ligands, L_0 and L_1, to a protein. The starting point in this case is the *thermodynamic cycle* in fig. 4.1,[38] from which we infer that the difference in free energies of the binding processes:

$$P + L_0 \rightarrow PL_0 \; ; \; \Delta G_{bind}(L_0) \tag{4.1}$$

and

$$P + L_1 \rightarrow PL_1 \; ; \; \Delta G_{bind}(L_1) \tag{4.2}$$

equals the free energy difference between the processes:

$$PL_0 \rightarrow PL_1 \; ; \; \Delta G_{bound,L_0 \rightarrow L_1} \tag{4.3}$$

and

$$P + L_0 \rightarrow P + L_1 \; ; \; \Delta G_{free,L_0 \rightarrow L_1} \tag{4.4}$$

i.e, we can calculate the relative free energy, $\Delta\Delta G_{L_0 \rightarrow L_1}$, of binding the ligands L_0 and L_1 to the protein as:

$$\Delta\Delta G_{L_0 \rightarrow L_1} = G_{bind}(L_1) - \Delta G_{bind}(L_0) = \Delta G_{bound,L_0 \rightarrow L_1} - \Delta G_{free,L_0 \rightarrow L_1} \tag{4.5}$$

The processes 4.3 and 4.4 are the ones that we simulate in this thesis, i.e. we run simulations where we transform one ligand into another, either when it is bound to protein or when it is free in solution. So, how do we then calculate the free energy differences between two states?

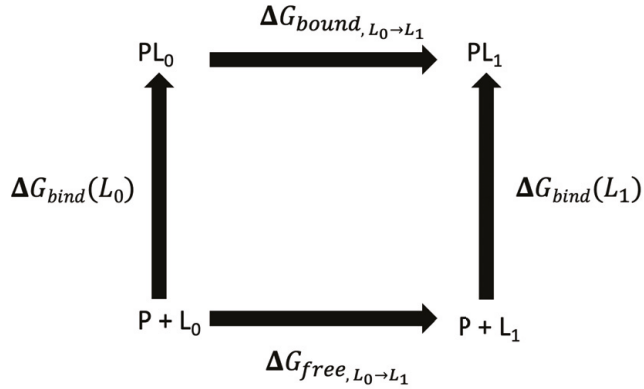


Figure 4.1: illustration of the thermodynamic cycle of protein–ligand binding.

4.1 Free energy perturbation

The free energy perturbation (FEP) formalism refers to various methods of computing free energy differences, but will start by discussing the original method: the *Zwanzig equation* from 1954.[39] If we look at two states *A* and *B*, with potential energy functions U_A and U_B, the Zwanzig equation (also called the exponential average) gives:

$$\Delta G_{A \to B} = -k_B T ln \left\langle e^{-(U_B - U_A)/k_B T)} \right\rangle_A \tag{4.6}$$

which is based on a ensemble average over configurations sampled for state A. T is the absolute temperature, k_B is the Boltzmann constant and ΔG is the Helmholtz free energy difference between the states *A* and *B*. The exponential average in Eq. 4.6 is poorly conditioned and will not converge if the free energy difference between the two states is too large. Therefore, the transformation between the states *A* and *B* is typically divided into several intermediate states, employing a mixing parameter λ:

$$U(\lambda) = (1 - \lambda)U_A + \lambda U_B \; ; \; \lambda \in [0, 1] \tag{4.7}$$

Thus, we pass through several intermediate states as we alchemically transform state *A* into *B*. The convergence is determined by the overlap between the neighbouring states. In order to assess the convergence, we have employed in this thesis the Wu and Kofke bias measure (II), which measures how much of the phase space of the the respective states overlaps with each other.[40]

Another common method to estimate free energy differences when alchemically transforming one ligand into another is thermodynamic integration (TI).[41] Here, the free energy difference is obtained by integrating ensemble-averaged derivative of the

potential-energy function with respect to λ:

$$\Delta F = \int_0^1 \left\langle \frac{\delta U(\lambda)}{\delta \lambda} \right\rangle_\lambda d\lambda \tag{4.8}$$

Convergence problems due to large energy changes will be seen in the form of big leaps in the derivative, which is one advantage of TI.

Another often used way of computing free energy differences is the Bennett acceptance ratio (BAR) method:[42,43]

$$e^{-(\Delta F - C)/k_B T} = \frac{\langle f[(U_B - U_A - C)/k_B T]\rangle_A}{\langle f[(U_A - U_B - C)/k_B T]\rangle_B} \tag{4.9}$$

where the Fermi function $f(x)$ is given by

$$f(x) = \frac{1}{1 + e^{x/k_B T}} \tag{4.10}$$

and the constant C is calculated in an iterative procedure that goes on until the ensemble averages in 4.9 are equal. Note that Eq. 4.9 involves simulations of both the *A* and *B* states. The BAR method gives the lowest variance among all estimators of the free energy.[43]

It is possible to generalize the BAR method to the case with multiple states, which is the case with alchemically transforming one ligand to another using mixing parameter λ. In the multistate Bennett acceptance ratio (MBAR) method, data from simulations of all λ states are used to compute the free energy difference.[44]

4.2 Free energies at the QM/MM level

In this thesis, we primarily compute ligand-binding free energies at the the MM level, but it is in principle also possible to calculate the free energy difference in 4.5 at the QM/MM level. However, running simulations of protein–ligand complexes at the QM/MM level is very computationally demanding, and such full QM/MM-FEP studies are few and usually only involve the ligand in the QM region, treated with semi-empirical methods.[45,46,47,48]

An alternative is to employ the reference-potential with QM/MM sampling (RPQS) method,[48] which is based on work by Warshel and Gao.[49,50,51] Here, a full FEP is first done at the MM level, and the energies are then corrected to the QM/MM level at the end-points (see thermodynamic cycle in fig. 4.2). The transformation

MM→QM/MM is done via intermediate states, using the mixing parameter Λ for the MM energy function, U_{MM}, and QM/MM energy function, $U_{QM/MM}$:

$$U(\Lambda) = (1-\Lambda)U_{MM} + \Lambda U_{QM/MM} \; ; \; \Lambda \in [0,1] \tag{4.11}$$

So, for each state (s = bound/free), the QM/MM corrected energy is given by:

$$\Delta G^{QM/MM}_{s,L_0\to L_1} = \Delta G^{MM}_{s,L_0\to L_1} - \Delta G^{MM\to QM/MM}_{s,L_0} + \Delta G^{MM\to QM/MM}_{s,L_1} \tag{4.12}$$

which gives for the relative binding free energy of two ligands:

$$\Delta\Delta G^{QM/MM}_{L_0\to L_1} = \Delta G^{QM/MM}_{bound,L_0\to L_1} - \Delta G^{QM/MM}_{free,L_0\to L_1} \tag{4.13}$$

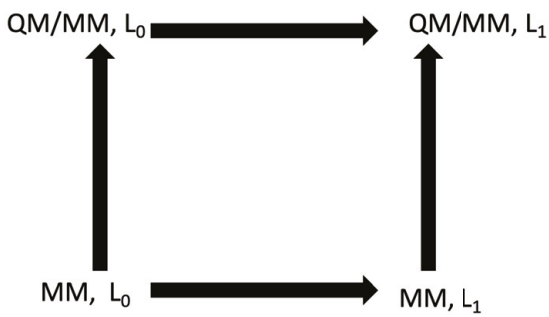


Figure 4.2: Illustration of the thermodynamic cycle used in the RPQS method. The cycle is employed both for the protein–ligand complex and the free ligand in solution.

4.3 End-point methods

In this thesis, we also compute free energies with methods that utilize only the physical end states of the complex, protein and ligand. These are the molecular mechanics Poisson–Boltzmann or generalised Born surface area (MM/PBSA and MM/GBSA) methods.[52,53] First, molecular dynamics (MD) simulations are run, after which water molecules are stripped off and the free energy is approximated by:

$$\Delta G_{bind} = \langle \Delta E_{el} \rangle + \langle \Delta E_{VDW} \rangle + \langle \Delta G_{sol} \rangle + \langle \Delta G_{SASA} \rangle - \langle T\Delta S_{NM} \rangle \tag{4.14}$$

where ΔE_{el} is the electrostatic energy and ΔE_{VDW} is the VDW energy, both calculated with a MM force field. ΔG_{sol} is the solvation free-energy calculated either from the Poisson–Boltzmann equation or by the generalised Born approach and ΔG_{SASA}

is the non-polar solvation free energy, estimated from the solvent-accessible surface area (SASA). Finally, a *normal-mode* (NM) analysis of vibrational frequencies yields ΔS_{NM}, the translational, rotational, and vibrational entropies. The brackets denote that all terms are calculated as averages over a number of snapshots from the MD simulations.

Each energy term in 4.14 is obtained by taking the difference between the complex (RL), the free receptor (R) and the ligand (L):

$$\Delta E = E(RL) - E(R) - E(L) \tag{4.15}$$

The last two terms on the right side of 4.15 are computed by stripping off the ligand or the receptor from the snapshots taken from the simulations of the complex.

Chapter 5

Thermodynamics of biomolecular systems

The entropic contribution to the free energy of protein–ligand binding, as specified in Eq. 1.3, is significant. It can arise from various sources, as discussed in chapter 1.2: the changes in solvent entropy, such as the release of solvent upon binding of the ligand, the change in conformational freedom of the ligand and protein and the loss of rotational and translational entropy of the ligand and protein during the binding proceess. It is therefore of utmost importance to be able to estimate this part, and we will discuss various methods of doing so in this section.

5.1 Entropy of protein–ligand complexes

As mentioned in chapter 3.3, one may compute the entropic contribution to the binding free energy (equation 4.14) by the NM method, which performs a harmonical analysis of vibrational frequencies.[16] This method is quite time-consuming,[54,55], which stems from the cost of diagonalizing the Hessian matrix (the second derivatives of the potential energy function with respect to the coordinates).[16] It is consequently of interest to find cheaper methods to compute entropies.

In 2016, Zhang and coworkers suggested the interaction entropy method (IE),[56] as a means of estimating the entropy of the binding process:

$$-T\Delta S_{IE} = RT\ln\left\langle e^{(\Delta E_{IE}-\langle\Delta E_{IE}\rangle)/RT}\right\rangle \tag{5.1}$$

where R is the gas constant and $\Delta E_{IE} = \Delta E_{el} + \Delta E_{VDW}$, the non-bonded

interaction energy. The latter energy terms may be estimated from MM/PBSA or MM/GBSA methods; the IE method estimates the entropy from the fluctuations of these energies.

Minh and coworkers suggested an even cheaper method of computing entropies in 2018, the so-called second-order cumulant approximation (C2) method.[57] The C2 method is based on the expression of the binding free energy as an exponential average of ΔE_{IE}, which is then expanded:

$$\Delta G_{IE} = \Delta H_{IE} - T\Delta S_{IE} = RT\ln e^{\langle \Delta E_{IE} \rangle / RT} = \langle \Delta E_{IE} \rangle + \frac{\sigma_{IE}^2}{2RT} + \ldots \quad (5.2)$$

where the standard deviation, σ_{IE}, is computed for E_{IE} over all snapshots. As can be seen from the expression above, an approximation for the binding entropy can be found in the second-order cumulant approximation term:

$$-T\Delta S_{C2} = \frac{\sigma_{IE}^2}{2RT} \quad (5.3)$$

Since exponential averaging is not used to estimate the entropy in 5.3, it is more numerically stable than the IE method. In paper V, we compare the IE and C2 methods and their convergence.

5.2 Thermodynamics of the solvent

Water plays a significantly role in biomolecular reactions. Of course, the protein, the ligand and the protein–ligand complex are completely surrounded by water molecules. Several of the water molecules will interact with the ligand and protein residues in the binding site, even if it is buried within the protein. Here, we will see how to calculate the contribution of the solvent thermodynamics for a ROI, in this case the binding site.

5.2.1 Inhomogeneous solvation theory

Inhomogeneous solvation theory (IST) was developed for analysing the thermodynamics of solvent (in our case water) molecules in MD simulations, making use of statistical thermodynamics.[58,59] The method calculates the energies and entropies by integrating correlation functions representing the rotational and translational degrees of freedom. The general starting point of IST is that the solvation entropy has both solute–water (sw) and water–water (ww) contributions:

$$\Delta S_{solv} = \Delta S_{sw} + \Delta S_{ww} \quad (5.4)$$

The water–water entropy term may be ignored, which gives the following formula for the entropy:

$$\Delta S_{solv} \approx \Delta S_{sw} \equiv \frac{-k_B \rho^0}{8\pi^2} \int g_{sw}(\mathbf{r},\omega) \ln\, g_{sw}(\mathbf{r},\omega) d\mathbf{r} d\omega \tag{5.5}$$

where ρ^0 is the number density of bulk solvent; $g_{sw}(\mathbf{r},\omega)$ is the solute–water pair-correlation function in the solute frame of reference, $\mathbf{r}$ is defined as the location of a water oxygen relative to the solute and ω is the Euler angles in the solute frame of reference; the factor of $1/(8\pi^2)$ normalizes the orientational integrals. Since $g_{sw}(\mathbf{r},\omega)$ equals unity for bulk density and a uniform orientational distribution, the solvation entropy vanishes in the bulk. Therefore, all contributions to the solvation entropy come from regions occupied by water, which is why the water–water entropy term may be ignored.

The solute-water entropy term, ΔS_{sw}, may be split into translational and orientational terms by rewriting $g_{sw}(\mathbf{r},\omega)$ as the product of a translational distribution function, $g_{sw}(\mathbf{r})$, and an orientational distribution function, conditioned on the position, $g_{sw}(\omega|\mathbf{r})$:

$$\Delta S_{sw} = \Delta S_{sw}^{trans} + \Delta S_{sw}^{orient} \tag{5.6}$$

This gives:

$$\Delta S_{sw}^{trans} \equiv -k_B \rho^0 \int g_{sw}(\mathbf{r}) \ln\, g_{sw}(\mathbf{r}) d\mathbf{r} \tag{5.7}$$

and

$$\Delta S_{sw}^{orient} \equiv \rho^0 \int g_{sw}(\mathbf{r}) S^{\omega}(\mathbf{r}) d\mathbf{r} \tag{5.8}$$

where

$$S^{\omega}(\mathbf{r}) \equiv \frac{-k_B}{8\pi^2} \int g_{sw}(\omega|\mathbf{r}) \ln\, g_{sw}(\omega|\mathbf{r}) d\omega \tag{5.9}$$

and $g_{sw}(\mathbf{r}) \equiv \rho(\mathbf{r})/\rho^0$ and $g_{sw}(\omega|\mathbf{r}) \equiv \rho(\omega|\mathbf{r})/\rho^0_\omega = 8\pi^2 \rho(\omega|\mathbf{r})$, where $\rho(\mathbf{r})$ and $\rho(\omega|\mathbf{r})$ are Boltzmann probability densities.

Likewise, the solvation energy is also divided into solute–water and water–water terms:

$$\Delta E_{solv} = \Delta E_{sw} + \Delta E_{ww} \tag{5.10}$$

The water–water energy term may be ignored for the same reasons as for the water–water entropy term. We have for solute–water:

$$\Delta E_{sw} = \rho^0 \int g_{sw}(\mathbf{r}) \Delta E_{sw}(\mathbf{r}) d\mathbf{r} \tag{5.11}$$

$$\Delta E_{sw}(\mathbf{r}) \equiv \frac{1}{8\pi^2} g_{sw}(\omega|\mathbf{r}) U_{sw}(\mathbf{r},\omega) d\omega \tag{5.12}$$

where $U_{sw}(\omega,\mathbf{r})$ is the solute–water interaction potential.

5.2.2 Grid inhomogeneous solvation theory

In this thesis, we worked with the discrete version of IST, called grid inhomogeneous solvation theory (GIST).[59] One major concern is that water molecules interacting with the protein–ligand complex often exchange with bulk water molecules during the simulations. This may be addressed by clustering water molecules that are close together in time and space,[60] but then only water molecules with high occupancies are thus considered. Another way of dealing with this was proposed by Gilson and coworkers in the GIST method, namely to assign energies and entropies to 3-dimensional voxels instead of individual water molecules. These voxels collectively make up a 3D grid spanning the ROI. In order to assign the voxels during the simulation in a meaningful way, it is necessary to restrain the solute throughout the simulations.

We use k to index the voxels, which gives for the total translational entropy of a ROI:

$$\Delta S_{sw}^{ROI,trans} \approx \sum_{k \in ROI} \Delta S_{sw}^{trans}(\mathbf{r}_k) \tag{5.13}$$

$$\Delta S_{sw}^{trans}(\mathbf{r}_k) \approx k_B \rho^0 V_k g(\mathbf{r}_k) \ln g(\mathbf{r}_k) \tag{5.14}$$

$$g(\mathbf{r}_k) = \frac{N_k}{\rho^0 V_k N_{frame}} \tag{5.15}$$

where ρ^0 is the number density of bulk water, $\mathbf{r}_k$ is the position vector for voxel k, N_k is the total number of water molecules for all frames within voxel k, V_k is the volume of that voxel and N_{frame} is the number of frames in the simulation. Here, it is assumed that the correlation function $g(\mathbf{r}_k)$ is uniform within each voxel.

As for the orientational entropy, we have:

$$\Delta S_{sw}^{ROI,orient} \approx \sum_{k \in ROI} \Delta S_{sw}^{orient}(\mathbf{r}_k) \tag{5.16}$$

$$\Delta S_{sw}^{orient}(\mathbf{r}_k) \approx \rho^0 V_k g(\mathbf{r}_k) S^{\omega}(\mathbf{r}_k) \tag{5.17}$$

$$S^{\omega}(\mathbf{r}_k) = \frac{-k_B}{N_k} \left[\gamma + \sum_{i=1}^{N_k} \ln g(\omega_i | \mathbf{r}_k) \right] \tag{5.18}$$

where γ is Euler's constant and $g(\omega_i|\mathbf{r}_k)$ is the value of the orientational distribution for water i, which is approximated as:

$$g(\omega_i | \mathbf{r}_k) = \frac{8\pi^2}{V_i^{\omega}} \tag{5.19}$$

$$V_i^{\omega} = \frac{4\pi(\Delta\omega_i)^3}{3} \tag{5.20}$$

For each of the N_k water molecules in voxel k, in any trajectory frame, one finds the shortest angular distance $\Delta\omega_i$ to any other water molecule in the voxel. The angular distance between water molecules i and j is:

$$\Delta\omega_i \equiv \left[(\phi_i - \phi_j)^2 + (\cos\theta_i - \cos\theta_j)^2 + (\psi_i - \psi_j)^2\right]^{1/2} \tag{5.21}$$

with (ϕ, cosθ, ψ) being the Euler angles in the solute frame of reference.

As for the energy, the solute–water term is:

$$\Delta E_{sw}^{ROI} = \sum_{k \in ROI} \Delta E_{sw}(\mathbf{r}_k) \tag{5.22}$$

where E_{sw}^{ROI} is the total solute–water interaction energy in voxel k, averaged over all simulation frames. The water–water energy term is:

$$\Delta E_{ww}^{ROI} = \sum_{k \in ROI} \Delta E_{ww}(\mathbf{r}_k) - \frac{1}{2} \sum_{k \in ROI} \sum_{l \in ROI} \Delta E_{ww}(\mathbf{r}_k, \mathbf{r}_l) \tag{5.23}$$

where the total water–water interaction energies for the respetive voxels, $\Delta E_{ww}(\mathbf{r}_k)$ and $\Delta E_{ww}(\mathbf{r}_k, \mathbf{r}_l)$ are averaged over all simulation frames. Finally, the solvation free energy of voxel k can be expressed as follows:

$$\Delta G(\mathbf{r}_k) = \Delta E_{sw}(\mathbf{r}_k) + \Delta E_{ww}(\mathbf{r}_k) - T\Delta S_{sw}^{trans}(\mathbf{r}_k) - T\Delta S_{sw}^{orient}(\mathbf{r}_k) \tag{5.24}$$

Chapter 6

Summary of papers

6.1 Paper I

It is important in drug development to estimate the relative affinity of small molecules binding to biomacromolecules. In 2014, in a previous study, we calculated 91 relative binding affinities with TI, encompassing 107 ligands and ten proteins, using the AMBER ff09 for the protein and general AMBER force field (GAFF) for the ligands. The mean absolute deviations from experimental results were 1.6–10.5 kJ/mol and the maximum errors were 3–23 kJ/mol. In this FEP study, which was done with MM methods, we try to improve these results in different ways. The results show how challenging it is to consistently improve the results of FEP calculations.

First, we try updated force fields (FF14SB/GAFF2/OPC) and include the entire protein and all subunits of multimeric proteins as well as all cofactors and modifications. This gave improved results for five proteins, among them HIV-PT and fXa, but worse results for four proteins. For COX2, we successfully introduced new van der Waals parameters for the sulfonamide group, in order to avoid the formation of internal hydrogen bonds.

Second, we try to improve the charges by fitting them to 20 snapshots of a MD simulation of the protein–ligand complex. We tested two different approaches. The ligand was either placed in vacuum (Vac), or had its charges polarised by the surrounding protein and solvent (Ptch). For five proteins, the Vac results were better than the results obtained with the restrained electrostatic potential (RESP) charges, whereas the results deteriorated for four proteins. The Vac results were generally better than the Ptch results, although the opposite was true for three proteins.

We also measured the root-mean-square deviation (RMSD) of the ligand and the hydrogen-bond pattern relative to the crystal-structure conformation. For most proteins, the RMSD was low and the MD simulations did not deviate much from the crystal structure. However, for three proteins (COX2, HIV-PT and GP), the hydrogen bonds changed significantly during the MD simulations, which we tried to counteract by employing restraints for the hydrogen bonds. However, since this led to worse calculated binding affinites, it is possible that the crystal structure may not completely describe the actual conformation in the protein.

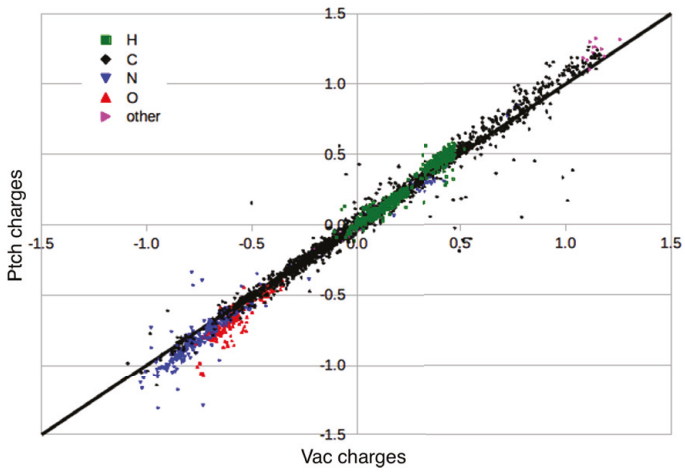


Figure 6.1: Correlation between the Vac and Ptch charges for the ligands, colour-coded after the element (the last group include S, with strongly positive charges, as well as F, Cl and Br with slightly negative charges).

6.2 Paper II

This paper focused on a congeneric series of ligands with a fluorophenyltriazole moiety, where the fluorine substituent is put in different ortho, meta, and para positions (denoted O, M, and P). The purpose was to study how the binding affinity of this ligand to galectin-3C is affected by small changes in the ligand structure. The total entropy of binding was split into contributions from protein, ligand, and solvent. This question was scrutinized with a variety of of experimental methods, isothermal titration calorimetry (ITC), X-ray crystallography and NMR relaxation, as well as theoretical methods in the form of MD simulations and GIST calculations.

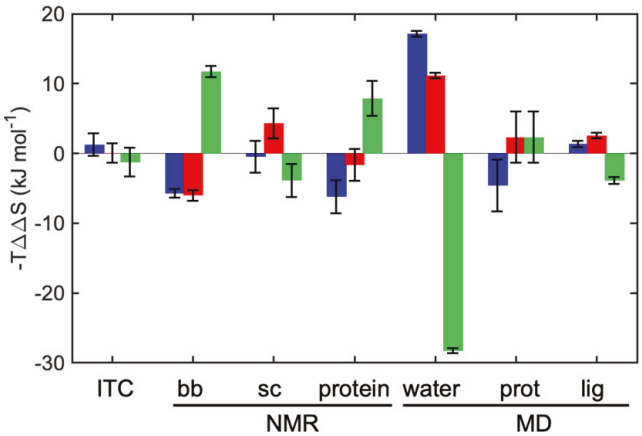


Figure 6.2: Entropy contributions to the differential binding of ligands M (blue), P (red), and O (green) to galectin-3C. The colored bars represent intercomplex differences in entropy; hence, a negative value corresponds to a favorable entropic contribution to binding for the specified complex, relative to the other two complexes. ITC reports the total entropy of binding. NMR reports estimates of the conformational entropy of the backbone (bb) and methyl-bearing side chains (sc). MD reports the conformational entropy of the protein (prot) and ligand (lig), and the solvation entropy determined by GIST (water). Error bars indicate ±1 standard deviation. (Figure taken from ref. 61.)

The three ligands have similar free energies of binding, which is expected from their structural similarities. However, the O ligand has less favorable binding enthalpy than the M and P ligands, which have similar enthalpies. This can be attributed to a smaller number of interactions between fluorine atom and surrounding protein for the O-complex. Also, the O-complex has lower conformational entropy than the other complexes, as suggested by NMR and ensemble-refined X-ray diffraction data. The GIST calculations showed that the O-bound complex has a less unfavorable solvation entropy compared to the other two complexes (fig. 6.2). This indicates that changes in ligand conformational entropy and water entropy compensates for changes in protein conformational entropy, i.e. an entropy–entropy compensation.[61]

6.3 Paper III

In this study we analyzed the role of water in biomolecular processes. We focused on two proteins: a ferritin dimer with a buried binding site and galectin-3C which has a solvent-exposed binding site. The methods included both dynamic simulations, by means of GCMC, MD and GCMC/MD, as well as the study of thermodynamics by means of GCMC and GIST.

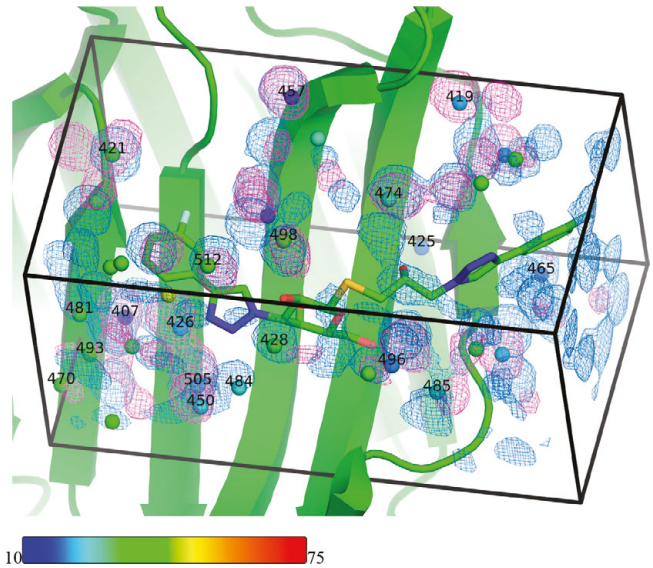


Figure 6.3: Density maps comparing the water sampling observed using constrained MD (AMBER, marine) or GCMC (ProtoMS, magenta) for the R ligand in complex with galectin-3C. The protein and ligand are shown with the crystallographic coordinates, and the experimental water sites are colored according to their temperature factors (scale shown at the bottom of the figure). Water molecules that make hydrogen bonds with the protein or the ligand are marked with residue numbers. The density maps are contoured at an isovalue of 0.6. (Figure taken from ref. 62.)

We saw that GCMC/MD equilibrates faster than regular MD when running simulations on ferritin with the buried binding site. All methods could reproduce crystal-water molecules quite well for ferritin, although GCMC/MD is more consistent between simulations that were started with and without water molecules present in the buried binding site at the beginning. However, for galectin-3C with an open binding site, MD is preferable to GCMC (fig. 6.3).

The solvation free energies for GIST and GCMC are not comparable, since they use different reference states. Enforcing restraints in the MD simulations improves the precision of the calculated thermodynamic GIST quantities, but also changes them profoundly.[62]

6.4 Paper IV

We studied the thermodynamics of a series of a series of complexes between galectin-3C and β-D-thiogalactopyranoside ligands with a meta-substituted phenyl group: H, F, Cl, Br, or I. This was done both with experimental methods, including ITC, competitive fluorescence polarization, X-ray crystallography and NMR spectroscopy, and theoretical methods, including QM calculations, MD simulations and GIST calculations.

ITC reveals that the favorable binding enthalpy increases from H to I, while the entropic penalty increases from H to F and then from Br to I. The ligands containing F, Cl, and Br have similar binding thermodynamic profiles. X-ray crystallography show that all ligands bind in very similar poses, with small variations of the length of the halogen bond to the carbonyl oxygen of Gly182 for F to I (3.1–3.3 Å). However, as for the water molecule binding to the halogen atom, the X–O distance increases from 3.1 Å for F to 3.6 Å for Br; for I the water is displaced.

QM calculations and ITC studies together suggest that the halogen bond between the ligand halogen and the Gly182 carbonyl oxygen is a significant contributor to the binding enthalpy. From the GIST results, it could be inferred that the solvation thermodynamics has contributions from water molecules across the entire binding site. Altogether, this highlights the importance of contributions of both the direct halogen bond interactions and solvation thermodynamics.

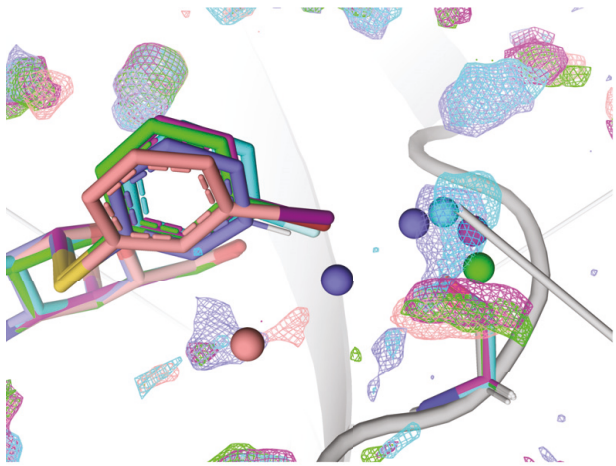


Figure 6.4: Superposition of the crystal structures and the water densities from the MD simulations for the five galectin-3C–ligand complexes with extra point parameters for Cl, Br and I, focused on the variable part of the ligands. The isodensity level is five times the bulk density. The variable water molecules in the crystal structures are shown as balls. The structures and densities are color coded: **H** (slate), **F** (cyan), **Cl** (magenta), **Br** (green) and **I** (salmon pink).

Attempts were made to improve the MM description of the halogen bonds, by adding an extra point charge outside the halogen atom representing the positive σ-hole. However, we were unable to find any parameters that successfully reproduced both the structure (fig. 6.4) and entropies derived from experiments. We also observed that FEP calculations on MD simulations gave more accurate results if an extra point charge was not used in the model.

6.5 Paper V

In this study, we used the IE and C2 methods to estimate the entropies of many different protein–ligand complexes. Both methods estimate entropies from the interaction energies (ΔE_{IE}), which is the sum of electrostatic and VDW energies. The interacting energies were calculated by applying the MM/GBSA method to the end-states of MD simulations. While the IE method is based on an exponential average of fluctuations in ΔE_{IE}, the C2 entropy is directly proportional to the square of the standard deviation of the interaction energies (σ_{IE}).

For σ_{IE} < 16 kJ/mol, the IE and C2 entropies agree quite well (to within 15 kJ/mol), although the difference between the two methods increases for larger σ_{IE}, the IE entropy consistently being smaller than the C2 entropy. Furthermore, by using stochastic simulations, we showed that for σ_{IE} > 15 kJ/mol, it is almost impossible to converge the IE entropies (fig. 6.5). The C2 method could on the other hand be converged for σ_{IE} up to 150 kJ/mol (fig. 6.5), although the entropies are most likely overestimated for σ_{IE} > 25 kJ/mol.

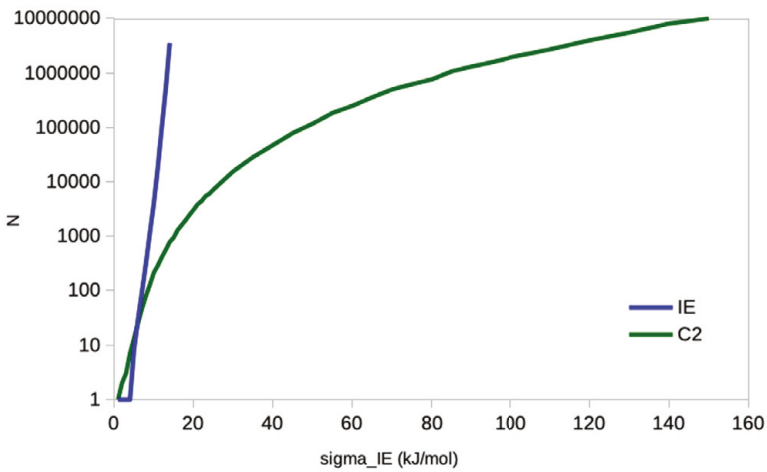


Figure 6.5: Number of snapshots (N) needed to converge the IE and C2 entropies within 4 kJ/mol of the analytical result with 95% confidence, assuming that the ΔE_{IE} energies follow a Gaussian distribution. (Figure taken from ref. 63.)

Zhang and coworkers proposed that the problem with poorly converged entropies can be solved by using a very dense sampling frequency of 10 fs in the MD simulations. Our results indicate that this is too dense, giving strongly correlated energies. This did not improve the calculated entropies, but gave statistical inefficiencies of 3–40 for lysozyme and ferritin. They also suggested that σ_{IE} could be reduced by enforcing restraints on the protein, but it also reduced the entropy as well as affected dynamics

in general. Enforcing a cutoff, i.e. ignoring energies $> 3\sigma_{IE}$ when computing IE entropies, is also questionable; for large σ_{IE}, the lowest values of ΔE_{IE} dominate the calculations of the IE entropies. Ignoring those values would distort the calculated entropies.

By utilizing block-averaging, we showed that for lysozyme (σ_{IE} = 6 kJ/mol), the C2 entropies are independent of the sample size, which is not the case for the IE entropies. For ferritin (σ_{IE} = 13 kJ/mol, both IE and C2 entropies depend strongly on the sample size. This may be explained by the non-Gaussian distribution of the ΔE_{IE} energies.

To conclude, we recommend that σ_{IE} should always be reported when using the IE and C2 methods. Entropies should be calculated with both methods and compared. Their dependence on the sample size should also be assessed.[63]

6.6 Paper VI

The purpose of this study was to analyse the binding of inhibitors of three halogenated inhibitors (Cl, Br and I) to CK2 and two synthetic disaccharides to galectin-3C (fig. 6.6). This was done by running MD simulations, and then employing with MM FEP calculations to extract relative binding free energies. For comparison, QM/MM simulations were also run in an attempt to improve the free energies, by adding QM/MM level corrections to the endpoints of the MM simulations, the so-called RPQS method. It was also analysed how the size of the QM region affected the RPQS QM/MM results, for CK2 by first including only the ligand in the QM region and later both the ligand and a small part of the protein. For both CK2 and galectin-3C, the QM region was simulated at the DFTB3 level of theory.

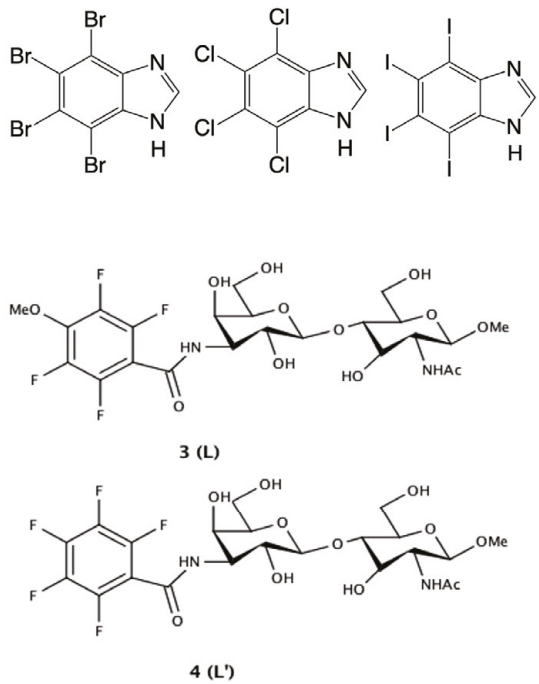


Figure 6.6: The ligands used in this study, a) TBB, TCB and TIB for CK2 and b) **3** and **4** for gal3.

For CK2, the MM FEP results correlated very well with the experimental results. This was surprising, due the inability of the point-charge model in a force field to accurately model the σ-hole of Cl, Br and I. Also surprisingly, the RPQS QM/MM results were worse, especially for the large QM region. This highlights the complexity of QM/MM simulations, such as modelling the interactions between the QM and MM regions (an electrostatic embedding scheme is used by default in AMBER, as well as in our simulations). In addition, when including part of the protein in the QM

region, covalent bonds must be cut and and saturated hydrogen atoms, adding further approximations. Moreover, although DFTB3 is one of most advanced Hamiltonians in the AMBER QM/MM implementation, it does not include halogen corrections. For CK2, we also ran a full FEP at the QM/MM level, which gave similar results as for RPQS for the small QM region, but worse results for the larger QM region.

For galectin-3C, RPQS QM/MM was tested with some different QM regions: the full ligand, the substituted benzene ring of the ligand and the substituted benzene ring of the ligand together with a few residues closest to the *para*-substituent. However, none of these approaches resulted in better binding affinities than with MM FEP, which did not accurately reproduce the experimental results to start with.

Chapter 7

Conclusions and Outlook

In this thesis, we have used computational methods to analyze the dynamics and thermodynamics of protein–ligand systems. It is a challenging undertaking, as the systems are very large and thus require more approximate methods than QM, and it is also very difficult to assess the important entropic part of the thermodynamics.

We have shown that QM-derived charges may improve MM relative free energies of protein–ligand binding, but far from always. Compuational approaches such as MD and GIST can also be combined with experimental methods, such as NMR, to provide a more detailed picture of various phenomena, such as entropy–entropy compensation and halogen bonding. We concluded that combining GCMC with MD in GCMC/MD sped up the sampling of buried binding sites, and we saw that both GCMC and GIST could be used to assess the solvent thermodynamics. We used both IE and C2 methods to estimate entropies of protein–ligand systems, but were forced to conclude that those type of calculations are often difficult. We also used QM/MM simulations to estimate binding free energies, but with less satisfying results.

In the thesis, we have explored challenges for the pharmaceutical industry. Although we saw some slightly promising methods, it is still very hard to reproduce experimental results to within acceptable accuracy. MM-FEP calculations may be improved by trying other force fields, such as ff15ipq and ff19SB.[64,65] The QM/MM-FEP calculations with the native AMBER QM/MM routine may be improved by implementing the PM7 method,[66] which is designed for reducing errors in noncovalent interactions involving halogens. Also, better methods than IE and C2 for estimating the entropies are needed.

References

[1] G. L. Patrick, *Medicinal Chemistry*. Oxford, 2001.

[2] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, and S.-Q. Liu, "Insights into protein–ligand interactions: mechanisms, models, and methods," *International Journal of Molecular Sciences*, vol. 17, no. 2, p. 144, 2016.

[3] S. de Beer, N. P. Vermeulen, and C. Oostenbrink, "The role of water molecules in computational drug design," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, pp. 55–66, 2010.

[4] P. Ball, "More than a bystander," *Nature*, vol. 478, no. 7370, pp. 467–468, 2011.

[5] M. B. Smith, *Biochemistry: An Organic Chemistry Approach*. CRC Press, 2020.

[6] A. R. Voth, F. A. Hays, and P. S. Ho, "Directing macromolecular conformation through halogen bonds," *Proceedings of the National Academy of Sciences*, vol. 104, no. 15, pp. 6188–6193, 2007.

[7] L. Brammer, E. A. Bruton, and P. Sherwood, "Understanding the behavior of halogens as hydrogen bond acceptors," *Crystal Growth & Design*, vol. 1, no. 4, pp. 277–290, 2001.

[8] M. R. Scholfield, M. C. Ford, A.-C. C. Carlsson, H. Butta, R. A. Mehl, and P. S. Ho, "Structure–energy relationships of halogen bonds in proteins," *Biochemistry*, vol. 56, no. 22, pp. 2794–2802, 2017.

[9] E. V. Anslyn and D. A. Dougherty, *Modern physical organic chemistry*. University science books, 2006.

[10] J. C. Ma and D. A. Dougherty, "The cation-π interaction," *Chemical reviews*, vol. 97, no. 5, pp. 1303–1324, 1997.

[11] R. Grünberg, M. Nilges, and J. Leckner, "Flexibility and conformational entropy in protein-protein binding," *Structure*, vol. 14, no. 4, pp. 683–693, 2006.

[12] L. S. Vedula, G. Brannigan, N. J. Economou, J. Xi, M. A. Hall, R. Liu, M. J. Rossi, W. P. Dailey, K. C. Grasty, M. L. Klein, *et al.*, "A unitary anesthetic binding site at high resolution," *Journal of Biological Chemistry*, vol. 284, no. 36, pp. 24176–24184, 2009.

[13] Schrödinger, LLC, New York, *Maestro*, 2016.

[14] M. L. Verteramo, O. Stenström, M. M. Ignjatović, O. Caldararu, M. A. Olsson, F. Manzoni, H. Leffler, E. Oksanen, D. T. Logan, U. J. Nilsson, *et al.*, "Interplay between conformational entropy and solvation entropy in protein–ligand binding," *Journal of the American Chemical Society*, vol. 141, no. 5, pp. 2012–2026, 2019.

[15] R. Battistutta, M. Mazzorana, S. Sarno, Z. Kazimierczuk, G. Zanotti, and L. A. Pinna, "Inspecting the structure-activity relationship of protein kinase CK2 inhibitors derived from tetrabromo-benzimidazole," *Chemistry & Biology*, vol. 12, no. 11, pp. 1211–1219, 2005.

[16] F. Jensen, *Introduction to computational chemistry*. John Wiley & Sons, 2007.

[17] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature Reviews Drug discovery*, vol. 3, no. 11, pp. 935–949, 2004.

[18] M. Korth, "Third-generation hydrogen-bonding corrections for semiempirical QM methods and force fields," *Journal of Chemical Theory and Computation*, vol. 6, no. 12, pp. 3808–3816, 2010.

[19] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *The Journal of Chemical Physics*, vol. 132, no. 15, p. 154104, 2010.

[20] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical Review*, vol. 136, no. 3B, p. B864, 1964.

[21] J. L. Bao, L. Gagliardi, and D. G. Truhlar, "Self-interaction error in density functional theory: An appraisal," *The Journal of Physical Chemistry Letters*, vol. 9, no. 9, pp. 2353–2358, 2018.

[22] D. A. Case, R. E. Duke, R. C. Walker, N. R. Skrynnikov, T. E. Cheatham III, O. Mikhailovskii, C. Simmerling, Y. Xue, A. Roitberg, S. A. Izmailov, *et al.*, "Amber 22 reference manual," 2022.

[23] N. Huang, C. Kalyanaraman, K. Bernacki, and M. P. Jacobson, "Molecular mechanics methods for predicting protein–ligand binding," *Physical Chemistry Chemical Physics*, vol. 8, no. 44, pp. 5166–5177, 2006.

[24] K. Vanommeslaeghe, O. Guvench, *et al.*, "Molecular mechanics," *Current Pharmaceutical Design*, vol. 20, no. 20, pp. 3281–3292, 2014.

[25] P. W. Atkins and R. S. Friedman, *Molecular quantum mechanics*. Oxford University Press, 2011.

[26] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "FF14sb: improving the accuracy of protein side chain and backbone parameters from FF99sb," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3696–3713, 2015.

[27] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.

[28] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: TIP4P-EW," *The Journal of Chemical Physics*, vol. 120, no. 20, pp. 9665–9678, 2004.

[29] S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building water models: a different approach," *The Journal of Physical Chemistry Letters*, vol. 5, no. 21, pp. 3863–3871, 2014.

[30] A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme," *Journal of Molecular Biology*, vol. 103, no. 2, pp. 227–249, 1976.

[31] L. Cao and U. Ryde, "On the difference between additive and subtractive QM/MM calculations," *Frontiers in Chemistry*, vol. 6, p. 89, 2018.

[32] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, no. 3, pp. 327–341, 1977.

[33] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An Nlog (N) method for Ewald sums in large systems," *The Journal of Chemical Physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[34] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[35] G. A. Ross, M. S. Bodnarchuk, and J. W. Essex, "Water sites, networks, and free energies with grand canonical Monte Carlo," *Journal of the American Chemical Society*, vol. 137, no. 47, pp. 14930–14943, 2015.

[36] D. Adams, "Chemical potential of hard-sphere fluids by Monte Carlo methods," *Molecular Physics*, vol. 28, no. 5, pp. 1241–1252, 1974.

[37] R. Kjellander, *Statistical Mechanics of Liquids and Solutions: Intermolecular Forces, Structure and Surface Interactions*. CRC Press, 2019.

[38] B. L. Tembre and J. A. Mc Cammon, "Ligand-receptor interactions," *Computers & Chemistry*, vol. 8, no. 4, pp. 281–283, 1984.

[39] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. i. nonpolar gases," *The Journal of Chemical Physics*, vol. 22, no. 8, pp. 1420–1426, 1954.

[40] D. Wu and D. A. Kofke, "Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation," *The Journal of Chemical Physics*, vol. 123, no. 5, p. 054103, 2005.

[41] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," *The Journal of Chemical Physics*, vol. 3, no. 5, pp. 300–313, 1935.

[42] C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," *Journal of Computational Physics*, vol. 22, no. 2, pp. 245–268, 1976.

[43] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, "Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods," *Physical Review Letters*, vol. 91, no. 14, p. 140601, 2003.

[44] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *The Journal of Chemical Physics*, vol. 129, no. 12, p. 124105, 2008.

[45] R. Rathore, R. N. Reddy, A. Kondapi, P. Reddanna, and M. R. Reddy, "Use of quantum mechanics/molecular mechanics-based FEP method for calculating relative binding affinities of FBPase inhibitors for type-2 diabetes," *Theoretical Chemistry Accounts*, vol. 131, pp. 1–10, 2012.

[46] M. R. Reddy and M. D. Erion, "Relative binding affinities of fructose-1,6-bisphosphatase inhibitors calculated using a quantum mechanics-based free energy perturbation method," *Journal of the American Chemical Society*, vol. 129, no. 30, pp. 9296–9297, 2007.

[47] K. Świderek, S. Martí, and V. Moliner, "Theoretical studies of HIV-1 reverse transcriptase inhibition," *Physical Chemistry Chemical Physics*, vol. 14, no. 36, pp. 12614–12624, 2012.

[48] M. A. Olsson and U. Ryde, "Comparison of QM/MM methods to obtain ligand-binding free energies," *Journal of Chemical Theory and Computation*, vol. 13, no. 5, pp. 2245–2253, 2017.

[49] V. Luzhkov and A. Warshel, "Microscopic models for quantum mechanical calculations of chemical processes in solutions: LD/AMPAC and SCAAS/AMPAC calculations of solvation energies," *Journal of Computational Chemistry*, vol. 13, no. 2, pp. 199–213, 1992.

[50] J. Gao, "Absolute free energy of solvation from Monte Carlo simulations using combined quantum and molecular mechanical potentials," *The Journal of Physical Chemistry*, vol. 96, no. 2, pp. 537–540, 1992.

[51] J. Gao and X. Xia, "A priori evaluation of aqueous polarization effects through Monte Carlo QM-MM simulations," *Science*, vol. 258, no. 5082, pp. 631–635, 1992.

[52] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, *et al.*, "Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models," *Accounts of Chemical Research*, vol. 33, no. 12, pp. 889–897, 2000.

[53] S. Genheden and U. Ryde, "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities," *Expert Opinion on Drug Discovery*, vol. 10, no. 5, pp. 449–461, 2015.

[54] J. Åqvist, V. B. Luzhkov, and B. O. Brandsdal, "Ligand binding affinities from md simulations," *Accounts of Chemical Research*, vol. 35, no. 6, pp. 358–365, 2002.

[55] J. Michel, "Current and emerging opportunities for molecular simulations in structure-based drug design," *Physical Chemistry Chemical Physics*, vol. 16, no. 10, pp. 4465–4477, 2014.

[56] L. Duan, X. Liu, and J. Z. Zhang, "Interaction entropy: A new paradigm for highly efficient and reliable computation of protein–ligand binding free energy," *Journal of the American Chemical Society*, vol. 138, no. 17, pp. 5722–5728, 2016.

[57] W. M. Menzer, C. Li, W. Sun, B. Xie, and D. D. Minh, "Simple entropy terms for end-point binding free energy calculations," *Journal of Chemical Theory and Computation*, vol. 14, no. 11, pp. 6035–6049, 2018.

[58] T. Lazaridis, "Inhomogeneous fluid approach to solvation thermodynamics. 1. theory," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3531–3541, 1998.

[59] C. N. Nguyen, T. Kurtzman Young, and M. K. Gilson, "Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril," *The Journal of Chemical Physics*, vol. 137, no. 4, p. 044101, 2012.

[60] T. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner, "Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding," *Proceedings of the National Academy of Sciences*, vol. 104, no. 3, pp. 808–813, 2007.

[61] J. Wallerstein, V. Ekberg, M. M. Ignjatović, R. Kumar, O. Caldararu, K. Peterson, S. Wernersson, U. Brath, H. Leffler, E. Oksanen, *et al.*, "Entropy–Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C," *JACS Au*, vol. 1, no. 4, pp. 484–500, 2021.

[62] V. Ekberg, M. L. Samways, M. Misini Ignjatović, J. W. Essex, and U. Ryde, "Comparison of grand canonical and conventional molecular dynamics simulation methods for protein-bound water networks," *ACS Physical Chemistry Au*, vol. 2, no. 3, pp. 247–259, 2022.

[63] V. Ekberg and U. Ryde, "On the use of interaction entropy and related methods to estimate binding entropies," *Journal of Chemical Theory and Computation*, vol. 17, no. 8, pp. 5379–5391, 2021.

[64] K. T. Debiec, D. S. Cerutti, L. R. Baker, A. M. Gronenborn, D. A. Case, and L. T. Chong, "Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model," *Journal of Chemical Theory and Computation*, vol. 12, no. 8, pp. 3926–3947, 2016.

[65] C. Tian, K. Kasavajhala, K. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, *et al.*, "ff19sb: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution," *Journal of Chemical Theory and Computation*, vol. 16, no. 1, pp. 528–552, 2019.

[66] J. J. Stewart, "Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters," *Journal of Molecular Modeling*, vol. 19, pp. 1–32, 2013.

Scientific publications

Author contributions

Paper I: Attempts to improve alchemical relative binding free-energy simulations by quantum-mechanical charges

I performed all of the MD simulations and FEP calculations for all transformations of COX2, HIV-PT and GP, as well as some transformations of fXa, CDK2 and DHFR. I also did the Vac and Ptch MD simulations and FEP calculations for all transformations of NA, Ferritin, Estro and p38.

Paper II: Entropy–Entropy Compensation between the Protein, Ligand, and Solvent Degrees of Freedom Fine-Tunes Affinity in Ligand Binding to Galectin-3C

I performed the MD simulations as well as the subsequent GIST calculations, conformational entropy estimates and MM/GBSA calculations.

Paper III: Comparison of Grand Canonical and Conventional Molecular Dynamics Simulation Methods for Protein-Bound Water Networks

I performed the AMBER MD calculations on ferritin, as well as subsequent GIST calculations on that protein. I took part in assessing the water structure (equilibration, clustering, density maps, etc.) of both ferritin and galectin-3C as well as writing the manuscript.

Paper IV: Halogen Bond Interactions and Solvation in Protein–Ligand Binding: Progressive Changes in Binding Thermodynamics Across a Series of Halogen-Substituted Ligands

I performed the MD simulations and GIST and FEP calculations on the complexes with the extra point charge on the halogen atom.

Paper V: On the Use of Interaction Entropy and Related Methods to Estimate Binding Entropies

I performed the MD simulations on galectin-3C, lysozyme and ferritin as well as the RMSD calculations in fig. S4, S6 and S8.

Paper VI: QM/MM binding-affinity calculations in proteins with the reference-potential approach

I performed all of the simulations and calculations on CK2. I took part in writing the manuscript.