# On the estimation of ligand binding affinities

Genheden, Samuel

2012

# On the estimation of ligand binding affinities

**Samuel Genheden**

Division of Theoretical Chemistry
Lund University, Sweden

LUND UNIVERSITY

**Doctoral Thesis**

The thesis will be publicly defended on Friday 10th of February 2012, 9.15 in lecture hall B, Chemical Centre, Lund

The faculty opponent is:
Thomas Simonson, Ecole Polytechnique, Paris

The grading committee is:
Johan Åqvist (Uppsala University), Lennart Nilsson (Karolinska Institue), Gunnar Nyman (Göteborg University), and Håkan Wennerström (Lund University)

Front cover: Molecules illustrated by Maja Genheden ©2011


Doctoral Thesis

Theoretical Chemistry
Center for Chemistry and Chemical Engineering
Lund University
P.O. Box 124
SE-221 00 Lund
Sweden



ISBN 978-91-7422-291-3
Printed by Media-Tryck, Lund University, Lund

To future students:
for inspiration
for knowledge

| Organization | Document name |
|---|---|
| LUND UNIVERSITY | DOCTORAL DISSERTATION |
| | Date of issue: February 10, 2012 |
| | Sponsoring organization: Research school in pharmaceutical sciences - FLÄK |
| Author(s): Samuel Genheden | |


Title and subtitle

On the estimation of ligand binding affinities


Abstract

A method to accurately estimate the binding affinity of a small molecule to a receptor would be indispensable in numerous fields. For instance, most drugs exert their action by binding to a macromolecule target. Thus, a lot of time and resources could be saved in drug design by predicting affinities by computer programs.

In a series of 15 papers, we have tested, compared, and improved the most popular methods to estimate binding affinities. We have used for instance molecular mechanics with generalized Born and surface area solvation (MM/GBSA), linear interaction energy (LIE), and alchemical perturbation methods. Some of the topics covered are:

* How the precision of MM/GBSA estimates are affected by the simulation protocol
* If semiempirical quantum-mechanical methods can improve affinity estimates
* A comparison of different polar solvation methods in MM/GBSA
* If non-polar solvation methods can model different degrees of active-site hydration
* How to obtain normal-mode entropies accurately and efficiently
* What method is more efficient: LIE or MM/GBSA
* A comparison of several end-point continuum-solvation methods
* What charge model to use in simulations of host–guest complexes
* The performance of end-point methods in a binding-affinity blind test
* How we can make alchemical methods more useful for drug design
* If a single-reference state can be used to simulate several ligands
* What properties calculated from molecular dynamics simulations do converge

Together, these studies clearly show what methods to use and what methods to avoid. We conclude that approximate methods are not very accurate and the results are highly system dependent. On the other hand, using alchemical methods, affinity differences between similar ligands can be accurately estimated both quickly and with a high precision.



Key words: binding affinities,protein-ligand complexes,host-guest complexes, MM/GBSA, MM/PBSA, LIE, PDLD/s-LRA, alchemical methods, sampling, solvation, drug design

Classification system and/or index termes (if any):

| Supplementary bibliographical information: | | Language: English |
|---|---|---|
| ISSN and key title: | | ISBN: 978-91-7422-291-3 |
| Recipient's notes | Number of pages: 330 | Price |
| | Security classification | |

## Populärvetenskaplig sammanfattning

De allra flesta människor har vid något tillfälle tagit ett läkemedel. Till exempel vaccineras nu för tiden alla svenska barn mot bland annat stelkramp. Vad nog de allra flesta däremot inte har gjort är att fundera på vad ett läkemedel är. Kanske sträcker man sig till att ha kollat upp ett läkemedel i FASS, och kanske får man då syn på en figur som består av streck och bokstäver (se Figur 1), något som vagt känns igen från gymnasiekemin. En sådan figur av streck och bokstäver är inget annat än en modell av en molekyl, och alla läkemedel är molekyler. Läkemedelsmolekylen kan vara stor och bestå av flera hundra, kanske tusen atomer, men de allra flesta läkemedel är mycket mindre än så. Deras jobb är att finna sitt mål, vanligtvis någon slags receptor, i kroppen och växelverka, det vill säga interagera, med receptorn. En mer generell benämning på molekyler som binder till en receptor är ligander, därav ligand i titeln på avhandlingen.

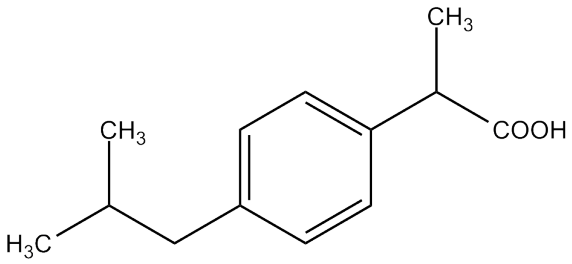


**Figure 1:** Ibuprofen. Den aktiva substansen i till exempel Ipren.

Människor i allmänhet går nog runt med föreställningen att kemister, som är de vetenskapsmän som i huvudsak sysslar med molekyler, är kufar i vita labbrockar som blandar vätskor så det pyser och stänker[1]. Denna föreställning är i många avseenden helt korrekt, men den kemin som behandlas i denna avhandling är av ett helt annat slag. Precis som FASS använder sig av modeller av streck och bokstäver för att beskriva de aktiva substanserna, så handlar denna avhandling om kemiska modeller. Något mer avancerade modeller än streck och bokstäver, men likväl modeller. Dessa modeller används sedan av datorprogram för att simulera det verkliga förloppet när en ligand växelverkar med sin receptor. Resultatet av simuleringarna behandlas sedan av matematiska teorier för att få fram mätvärden. I denna avhandling är vi mest intresserade av storheten bindningsaffinitet. Denna storhet summerar hur starkt liganden binder till receptorn, och kan grovt beskrivas som hur mycket liganden växelverkar med receptorn jämfört med rent vatten. Om en ligand växelverkar mer med receptorn än med vattnet är det bra, men allra mest intressant är att jämföra två ligander. Vilken växelverkar mest med receptorn?

Så vad har dessa modeller att göra med läkemedel? Och varför har jag inte blandat vätskor, utan i stället använt mej av modeller? För att kunna svara på detta behövs lite förståelse för hur ett nytt läkemedel blir till. Läkemedelsutvecklingen börjar med att man identifierar ett mål, det vill säga receptorn, och därefter försöker utvecklarna hitta ligander som har en stark affinitet för receptorn. Detta är den viktigaste egenskapen, men långt i från den enda, som ett bra läkemedel måste ha. I traditionell läkemedelsutveckling behöver dessa molekyler tillverkas och testas i ett labb. Läkemedelskemisterna testar vanligtvis tusentals av olika molekyler, men bara en bråkdel av dessa binder starkt till receptorn. Denna process är väldigt kostsam för att inte tala om tidskrävande — det tar vanligtvis mellan 10 och 20 år att utveckla ett nytt läkemedel. En del av kostnaden och tidsåtgången kan reduceras genom att modellera hela förloppet i en dator. På så sätt behöver inte en massa molekyler tillverkas. Om modellerna som vi använder oss av är tillräckligt noggranna kan de ersätta en del av den traditionella läkemedelsutvecklingen och på så sätt förkorta utvecklingstiden.

Det existerar en hel flora av olika metoder och modeller som kan beräkna affiniteten mellan en ligand och dess receptor. Denna avhandling behandlar noggrannheten och effektiviteten hos flera av dessa metoder, det vill säga hur väl de reproducerar verkligheten och hur fort det går att få fram ett resultat. Så hur långt har vi kommit? Kan läkemedels-

[1] Gör bara en Google-bildsökning på ordet ”chemist”

bolagen ersätta vätskor med datormodeller? Svaret är tyvärr nej. Flera av modellerna kan användas i ett första steg när det gäller att sortera bort flera hundra av dåliga molekyler till ett fåtal lovande. Därefter kan modellerna användas för att få en uppfattning om affiniteten. Tyvärr så dras alla metoderna med flera olika brister som gör att de inte kan användas universellt för alla typer av frågeställningar. Till exempel finns det ingen metod idag som kan ge en väldigt noggrann siffra för en enskild ligand inom en rimlig tid. Däremot finns det metoder som kan ge svar på skillnaden i affinitet mellan två ligander, det vill säga en metod som kan säga vilken utav två ligander som binder bäst. I många avseenden är det den viktigaste informationen. Forskningen som presenteras i den här avhandlingen kan används som en guide till vilka metoder som skall undvikas och vilka metoder som för närvarande är bäst.

# Contents

## List of Papers

This thesis is based on the following papers. The papers will be referred to by their Roman numerals and can be found at the end of the thesis.

I **S. Genheden** and U. Ryde (2010) How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.*, 31:837–846.

II **S. Genheden** and U. Ryde U (2011) Comparison of different initialisation protocols to generate independent molecular dynamics simulations. *J. Comput. Chem.*, 32:187–195

III P. Mikulskis, **S. Genheden**, K. Wichmann, U. Ryde (2011) A semi-empirical approach to ligand affinities: Dependence on Hamiltonian and corrections. *J. Comput. Chem.*, submitted

IV **S. Genheden**, T. Luchko, S. Gusarov, A. Kovalenko, U. Ryde (2010) An MM/3D-RISM approach for ligand-binding affinities. *J. Phys. Chem. B*, 114:8505–8516

V **S. Genheden**, J. Kongsted, P. Söderhjelm, U. Ryde (2010) Non-polar solvation free energies of protein–ligand complexes. *J. Chem. Theory Comput.*, 6:3558–3568

VI **S. Genheden**, P. Mikulskis, L-H. Hu, J. Kongsted, P. Söderhjelm, U. Ryde (2011) Accurate estimation of non-polar solvation free energies requires explicit consideration of binding-site hydration. *J. Am. Chem. Soc.*, 133:13081–13092

VII **S. Genheden** and U. Ryde (2011) The normal-mode entropy in the MM/GBSA method: Effect of system truncation, buffer region, and dielectric constant. Manuscript

VIII **S. Genheden** and U. Ryde (2011) Comparison of the efficiency of the LIE and MM/GBSA methods to calculate ligand-binding energies. *J. Chem. Theory Comput.*, 7:3768–3778

IX **S. Genheden** and U. Ryde (2011) Comparison of end-point continuum-solvent methods for the calculation of protein–ligand binding free energies. *Proteins*, accepted

X **S. Genheden** (2011) MM/GBSA and LIE estimates of host–guest affinities: dependence on charges and solvation model, *J. Comput.-Aided Mol. Design*, 25:1085–1093

XI P. Mikulskis , **S. Genheden**, P. Rydberg, L. Sandberg, L. Olsen, U. Ryde U (2011) Binding affinities of the SAMPL3 trypsin and host–guest blind tests estimated with the MM/PBSA and LIE methods. *J. Comput.-Aided Mol. Design*, in press

XII **S. Genheden**, I. Nilsson, U. Ryde (2011) Binding affinities of factor Xa inhibitors estimated by thermodynamic integration and MM/ GBSA. *J. Chem. Inf. Model*, 51:947–958

XIII **S. Genheden** and U. Ryde (2011) Improving efficiency of protein–ligand free energy calculations by system truncation. *J. Chem. Theory Comput.*, submitted

XIV **S. Genheden**, A. K. Malde, A. E. Mark, U. Ryde (2011) Binding affinities of factor Xa inhibitors estimated by a single-step perturbation approach. Manuscript

XV **S. Genheden** and U. Ryde (2011) Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys. Chem. Chem. Phys.*, submitted

## My contributions to the papers

I, II, VII, VIII, IX, XII, XIII. I performed all of the simulations, all energy calculations and most of the data analysis. I wrote the paper together with Ryde.

III. I performed some of the simulations and did some of the energy calculations. I supervised the rest of the simulations and the calculations. I wrote the paper together with Ryde.

IV. I performed all of the simulations and I did all of the energy calculations except 3D-RISM. I took part in writing the paper.

V. I performed all of the simulations and I did all of the energy calculations except PCM. I took part in writing the paper.

VI. I performed about half of the simulations and I did most of the energy calculations except PCM. I took part in writing the paper.

X. I did everything in this single-author paper.

XI. I performed all of the host–guest simulations and the energy calculations on those systems. I did most of the data analysis. I wrote the majority of the paper.

XIV. I performed all simulations, all energy calculations, and all data analysis. I wrote the majority of the paper.

XV. I performed all of the simulations, energy calculation, and some of the data analysis. I wrote the paper together with Ryde.

## Other papers, not included in the thesis

1. A. Ciancetta, **S. Genheden**, U. Ryde (2011) A QM/MM study of the binding of RAPTA ligands to cathepsin B, *J. Comput.-Aided Mol. Des.*, 25:729-742

2. **S. Genheden**, P. Söderhjelm, U. Ryde (2011) Transferability of conformational dependent charges from protein simulations. *Int. J. Quant. Chem.*, in press, DOI: 10.1002/qua.22967

3. P. Söderhjelm, J. Kongsted, **S. Genheden**, U. Ryde (2010) Estimates of ligand-binding affinities supported by quantum mechanical methods. *Interdiscip. Sci.: Comput. Life. Sci.*, 2:21-37

4. P. Söderhjelm, **S. Genheden**, U. Ryde (2011) Quantum mechanics in structure-based ligand design. In *Protein-ligand interactions*, H. Gohlke, ed., Wiley & Sons, in press

5. C. Diehl, **S. Genheden**, K. Modig, U. Ryde, M. Akke (2009) Conformational entropy changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *J. Biomol. NMR*, 45:157-169.

6. **S. Genheden**, C. Diehl, M. Akke, U. Ryde, (2010) Starting-condition dependence of order parameters derived from molecular dynamics simulations. *J. Chem. Theory Comput.*, 6:2176–2190.

7. C. Diehl, O. Engström, T. Delain, M. Håkansson, **S. Genheden**, K. Modig, H. Leffler, U. Ryde, U. Nilsson, M. Akke (2010) Protein flexibility and conformational entropy in drug/ligand design targeting the carbohydrate recognition domain of Galectin-3. *J. Am. Chem. Soc.*, 132:14577–14589

8. K. Saraboji, M. Håkansson, **S. Genheden**, C. Diehl, J. Qvist, U. Weininger, U. Nilsson, H. Leffler, U. Ryde, M. Akke, D. T. Logan (2011) The carbohydrate-binding site in galectin-3 is preorganized to recognize a sugar-like framework of oxygens: ultrahigh resolution structures and water dynamics. *Biochem.*, in press, DOI: 10.1021/bi201459p

9. C. Diehl, **S. Genheden**, C. Johansson, J. Peterson, G. Krantzl, M. Lepistö. N. Dekker, J. Evenäs, G. Carlström, U. Ryde, M. Akke (2011) Dissecting entropy in drug design: The case of human matrix metalloprotease MMP12, manuscript

10. **S. Genheden**, N. M. F. S.A. Cerqueira, P. A. Fernandes, L. Eriksson, M. J. Ramos (2011) Modelling the biochemistry paradox of the ribonucleotide reductase holoenzyme. Two large pieces fall into place, manuscript

## Preface

So here we are. Approximately four years since I first sat on a train headed for Lund and an interview for a PhD position, and approximately three and a half years since I worked my first day at the department. A lot has happened in those years, and this thesis is sort of a summary of those things. For instance, over twenty papers have been written, and fifteen of them are presented in this thesis.

While I have been working on those research projects, people all around the world have become ill and people have died from numerous diseases. For instance, more than 6 000 HIV infections occurs every day all around the globe, over 30 million people are living with HIV, and 14 million children in sub-Sarahan Africa has lost at least one of their parents to the disease [1]. According to latest statistics from WHO in 2008 [2], almost 16 million died from communicable, maternal, perinatal and nutritional conditions, and 36 million from non-communicable diseases. The largest individual group was cardiovascular diseases with over 17 million deaths. All of these diseases are treated with drugs. According to WHO, pharmaceuticals account for between 15 and 30% of the health spending in transitional economies, and between 25 to 66% in developing countries. WHO has compiled a list of essential drugs, which should be available to every citizen of a country. In 2015, an estimated 10 million deaths a year could be saved by increasing the health intervention, many of which depend on essential drugs [3]. Naturally the pharmaceutical industry is huge, with an estimated value of US$300 billions. The annual sales are about US$10 billions. However, one third of the sales revenue is spent on marketing, and only about half of that is spent on research and

development.

Drugs are nothing more than molecules. Some of them are very large and contain thousands of atoms, but the majority of them is much smaller and usually contains less than one hundred atoms. Their role is to find the intended receptor in the human body and bind to it. This thesis is about drugs, or more generally ligands, and their receptors. It takes about 10 to 20 years to develop a new drug and costs an enormous amount of money. The hope is that computational methods, such as those presented in this thesis, could replace some of the laboratory work. As such, money and time would be saved. This would also be the ultimate goal of green chemistry because less chemicals need to be manufactured.

The theoretical foundation of the methods used during my studies was developed several decades ago. First by John Kirkwood in 1935 [4] and later by Robert Zwanzig in 1954 [5], who developed techniques to calculate free energy differences. However, these theories had to wait for the computer to be useful for anything else but really simple model systems. In 1977, McCammon, Gelin, and Karplus performed the first atomistic molecular dynamics simulation of a protein [6], just a few years after the first empirical force fields for such system were published [7]. In 1984, Tembe and McCammon presented a simple thermodynamic cycle for the calculation of relative binding affinities [8], and in 1986, Wong and McCammon used it to, for the first time, estimate the relative binding affinities for a protein–ligand system [9]. Before that, several groups had used similar techniques to study host–guest systems and to calculate solvation free energies [10]. However, the sampling was too short and the model potentials too inaccurate, for any of these results to be reliable. In the 90's, several advances was made, such as better theories for the calculation of affinities and more accurate potentials, although the computer power at that time did not allow for much sampling. In 2000, Kollman and co-workes declared that a new era of computational chemistry had begun [11]: the era of structure and free energy calculations. That is, due to advances in accurate potential, there was hope that accurate free energies could be calculated on a routine basis. In this thesis, the most popular methods to compute ligand-binding free energies have been used. Therefore, after you have read this, you will see how this era is performing.

The outline of this thesis follows closely the structure of a recent review by Christ, Mark, and van Gunsteren [12]. This is an excellent review from a pedagogical perspective, and therefore I have chosen to follow their outline. In the first chapter, I will introduce the topic at

hand, together with short detours into statistics and evaluation methods. In the three chapters after that, I will describe the three main ingredients of any free energy method, viz., molecular modelling, sampling, and free energy estimation. Then, I will present the research part of the thesis, divided into approximate and rigorous methods. However, my papers will be referenced throughout the thesis in the form of short information boxes. Finally, I will try to summarize my thoughts on where we are now and where we go from here.

I have enjoyed these years and I hope you will appreciate this thesis.

Bring out the equations!

On a train between Lund and Skövde,
September 2011

```
karma police arrest this man he talks in maths he
buzzesLikeAfridge hes like a detuned radio.
```

Thom Yorke

# 1 Introduction

In this text, I will treat a chemical process that can be described by the following reaction

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

where P, L, and PL are the free protein, the free ligand, and the protein–ligand complex, respectively. I will discuss protein–ligand complexes throughout the thesis because these are the complexes I have mainly studied, but Eqn 1.1 can describe the binding of any two molecules. I have for instance studied also host–guest complexes. The reaction in Eqn 1.1 is governed by the association constant, $K_a$, or conversely, the dissociation constant $K_d = 1/K_a$, given by the concentration of the three species at equilibrium

$$K_a = \frac{1}{K_d} = \frac{[\mathrm{PL}]}{[\mathrm{P}][\mathrm{L}]} \tag{1.2}$$

The binding free energy, $\Delta G$, of the reaction is related to $K_a$ through

$$\Delta G = -RT \ln K_a C^{ref} \tag{1.3}$$

where $R$ is the gas constant, $T$ is the absolute temperature in Kelvin, and $C^{ref}$ is the standard concentration. In this thesis, I will use $\Delta G$ as a general notation for free energy, irrespectively if it is a Helmholtz free energy (as in Eqn 1.3) or a Gibbs free energy. The pressure–volume term that should be added to Eqn 1.3 for a correct definition of a Gibbs free energy is negligible under the conditions assumed in this thesis, i.e., a condensed system at room temperature and atmospheric pressure [13].

## 1.1 Thermodynamic basis of ligand-binding

There are several ways to derive a useful expression of $K_a$ and the traditional one use chemical potentials as the base for the derivation [13]. However, I find it more useful to take an approach first shown by Bjerrum [14] and presented in modern form by Sharp and co-workers [15], and Roux and co-workers [16].

Consider the probability, $p_1$, that one of the L molecules is bound to one of the P molecules. Conversely, it is possible to define $p_0$ as the probability that this L molecule is in the bulk, i.e., is unbound. By normalization it follows that $p_1 + p_0 = 1$. Therefore, $[\mathrm{PL}] = p_1[\mathrm{P}]_{\mathrm{tot}}$ and $[\mathrm{P}] = p_0[\mathrm{P}]_{\mathrm{tot}}$, where $[\mathrm{P}]_{\mathrm{tot}}$ is the total protein concentration in the system. Hence,

$$K_a = \frac{p_1[\mathrm{P}]_{\mathrm{tot}}}{p_0[\mathrm{P}]_{\mathrm{tot}}[\mathrm{L}]} = \frac{p_1}{p_0[\mathrm{L}]} \tag{1.4}$$

These probabilities can be calculated by taking the configurational average of an operator $\mathbf{H}$ that is 1 when L is bound, and 0 otherwise. This operator will define two parts of the configurational space of L, *site* and *bulk*, which are the parts of space where L can be considered to be bound and unbound, respectively. Therefore, $p_1$ and $p_0$, can be written as

$$p_1 = \frac{\int_{site} dq \int dS e^{-\beta U}}{\int dq \int dS e^{-\beta U}} \tag{1.5}$$

and

$$p_0 = \frac{\int_{bulk} dq \int dS e^{-\beta U}}{\int dq \int dS e^{-\beta U}} \tag{1.6}$$

where $dq$ and $dS$ are the coordinates of one ligand and the surroundings (protein and solvent), respectively, $1/\beta \equiv RT$, and $U$ is the total potential energy of the system. Now, we can express $K_a$ as

$$K_a = \frac{1}{[\mathrm{L}]} \frac{N \int_{site} dq \int dS e^{-\beta U}}{\int_{bulk} dq \int dS e^{-\beta U}} \tag{1.7}$$

where I have summed up the $N$ ligands because they are indistinguishable. The derivation can be taken one step further by assuming that the bulk region is isostropic and homogenous. This implies that we can fix the ligand at an arbitrary point in the bulk, $\mathbf{r}^*$. Hence, we can write

$$K_a = \frac{1}{[\mathrm{L}]} \frac{N \int_{site} dq \int dS e^{-\beta U}}{V \int_{bulk} dq \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U}} = \tag{1.8}$$

$$\frac{\int_{site} dq \int dS e^{-\beta U}}{\int_{bulk} dq \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U}}$$

where $\delta$ is the Dirac delta function and $V$ the volume of the bulk. The last equality follows from that [L] = $N/V$. This equation implies that $K_a$, and hence the binding free energy, can be calculated as the ratio of the likelihood that the ligand is bound to the protein to the likelihood that the ligand is out in the bulk. The equation also tells us what we need to do in order to compute the free energy. In fact, three main ingredients can be distinguished [12]:

1. We need to evaluate $U$, the total potential of the system. Therefore, we need a mathematical model of the system of interest. This will be discussed in Chapter II.

2. We need to evaluate the configurational integrals, i.e., we need to sample configurations of the system. How to sample the modelled system will be discussed in Chapter III.

3. Finally, we need a way to evaluate the free energy. Unfortunately, we cannot evaluate Eqn 1.8 in practice, because we cannot fully model a bulk system and we cannot sample all possible configurations. Therefore, we need to manipulate the equation such that it can be computed. How to do this will be discussed in Chapter IV.

## 1.2 Methods not based on sampling

Most of the methods to calculate binding affinities used in this thesis, are based on the treatment in the previous section, i.e., they require averaging over many configurations to estimate the affinity. However, this is a very time-consuming and is not feasible in for instance virtual screening where thousands of molecules should be tested in a short time. Therefore, simpler alternatives have been developed that are intended to give a semi-qualitative estimate of the binding affinity [17, 18]. Such methods are usually called scoring functions and can generally be divided into a few classes [18]

1. Energy-based methods

2. Empirical methods

3. Knowledge-based method

4. Consensus methods

The first class of scoring function tries to use a model, like those described in Chapter II, to describe the overall interaction energy between the ligand and the receptor. Additional terms to account for solvation and entropy have been incorporated. Examples of such scoring functions are DOCK [19] and Gold [20] that are based on a classical description of the energies, and the work of Merz and co-workers [21] and Hobza and co-workers [22] that are based on a semiempirical description. The second class of methods considers several terms that are thought to contribute to binding, e.g. hydrogen bonding, solvation, and loss of conformational freedom. Each of the terms are scaled by parameters that are optimized by comparing the scoring function with experimentally determined affinities. Examples of such scoring functions are FlexX [23] and Glide [24]. The third class of functions is based on a statistical pair-potential. This pair-potential is calculated from frequencies of atom contacts in experimental crystal structures and as such is a simplified potential of mean force. Examples of such scoring functions are DrugScore [25] and BLEEP [26]. Finally, consensus methods combine several scoring functions.

The scoring functions have been used with moderate success in virtual screening, i.e., they are fairly good at discriminating between ligands that are very poor binders and ligands that potentially could be good binders. However, scoring functions suffers from a generality problem. A scoring function that performs well on one system, could fail completely on another system [17].

---

In paper XI, a docking program, Glide, was used to predict binding modes of 34 ligands to the protein trypsin, prior to simulations and calculation of binding affinities. This gave us an opportunity to compare the results of the Glide scoring function with simulation-based estimates. In fact, Glide was better than the simulation-based method to discriminate between binders and non-binders. Glide was also slightly better to rank the binders, although the ranking performance of all methods was poor. We, also estimated the binding affinity of host–guest complexes and for these systems, Glide was significantly worse than most of the simulation-based methods. #

---

## 1.3 A primer on statistics

An issue, often overlooked in the ligand-binding community, is a statistical evaluation of the methods developed, tested, and applied to various

receptor–ligand systems. We have tried to be as thorough as possible in the papers to take care of this issue, and therefore, I here introduce some statistical concepts that are used throughout this thesis and especially in the papers.

The results of the simulations, are samples, $x^1, x^2, \ldots, x^N$, from an unknown distribution that measure the true value $y$. A common statistic of such samples is the average, $\langle x \rangle$, which is an estimate of $y$. To describe the reliability of the samples to measure $y$, it is instructive to introduce the mean squared error (MSE) [27]

$$\text{MSE} = \frac{1}{N}\sum(x_i - y)^2 \tag{1.9}$$

The MSE can easily be divided into two parts, one part that depends only on the samples $x$ and one part that depends on $y$ and $x$.

$$\text{MSE} = \sigma^2(x) + \delta^2(x, y) \tag{1.10}$$

The first term on the right-hand-side is the variance (the square-root of the variance is the standard deviation), and the second term is the squared bias. These terms are defined as

$$\sigma^2 = \frac{1}{N}\sum(x_i - \langle x \rangle)^2 \tag{1.11}$$

$$\delta = \langle x \rangle - y \tag{1.12}$$

The bias is a systematic error and stems mainly from the model of the system and is hard to improve by increased sampling. The variance on the other hand, is a statistical error and can in most cases be improved by increased sampling. I will generally talk about accuracy and precision, when discussing the systematic and the statistical error, respectively.

The accuracy is normally discussed in the literature, in relation with the evaluation methods described in the next section. However, the precision is rarely discussed, and in many cases it is not even reported. This is unfortunate, because the reliability of a method cannot be assessed without considering both accuracy and precision. For a simulation-based estimate, it is trivial to compute an uncertainty, but it is less so for scoring functions. However, approaches to estimate uncertainty of such functions have been proposed [28, 29]. The uncertainty is essential when comparing different methods. To see why, I will introduce the concept of confidence intervals.

A confidence interval is an interval in which a sample falls with a specific probability [30]. If I take a 95% confidence interval as example,

and repeat the sampling, then 95% of my samples will fall within this interval. Usually, we assume a Gaussian (normal) distribution for the estimate and then a confidence interval can be computed according to

$$\langle x \rangle \pm \frac{z_{\alpha/2}\sigma}{\sqrt{N}} \tag{1.13}$$

where $z_{\alpha/2}$ is a value taken from the cumulative Gaussian distribution function at the value $\alpha/2$. For a Gaussian distribution and a 95% confidence interval, this value is approximately 1.96.

Anything within the confidence interval is indistinguishable from a statistical point of view. Take two affinity estimates of 2 and 4 kJ/mol, respectively, as an example. The experimental affinity is 2 kJ/mol, and hence, the first estimate is intuitively much better than the second one because it is closer to the experimental value. However, if both measurements have an uncertainty, i.e., a standard error ($\sigma/\sqrt{N}$), of 2 kJ/mol, then neither of the differences between the estimates and the experiment is statistical significant because they fall within the confidence interval.

## 1.4 Evaluating methods and models

To evaluate the success a method, a reference must be established first. For the estimation of affinities, the naturally choice is an experimentally determined affinity. However, for other quantities there might not be an experiment that can give a reference. For instance, in paper V and VI, we wanted to compare several approximate methods to compute the nonpolar solvation free energy. In those papers, a theoretically more rigorous method was used as a reference.

Once the reference has been established, there exist several methods to evaluate the success. For a single affinity, it is simple to report the precision and the bias of the estimate. However, this is not as easily done for a series of compounds. Different metrics exists, and these can be divided into a few classes. For evaluation of absolute estimates, the mean absolute deviation (MAD) is a good metric

$$\text{MAD} = \frac{1}{N} \sum \mid x_i - y_i \mid \tag{1.14}$$

where $x_i$ is the theoretical estimate, and $y_i$ the corresponding experimental value. An underlining assumption in this metric is that the estimates follow a regression line with a slope of one and intercept at the origin. However, if there is a systematic error in all the estimates, this will not be the case. A way to circumvent this is to compute the MAD after removal

of the mean signed deviation (i.e., Eqn 1.14, without taking the absolute value). This metric is called translated MAD or MADtr for short.

$$\text{MADtr} = \frac{1}{N} \sum \mid x_i - y_i - 1/N \sum (x_i - y_i) \mid \tag{1.15}$$

The underlining assumption is still that the estimates follow a regression curve with slope of one, but the intercept can be different from the origin. Another popular metric is the coefficient of determination (or correlation coefficient), $r^2$, which measures the degree of linear correlation between two quantities.

$$r = \frac{\sum (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum (x_i - \langle x \rangle)^2 \times \sum (y_i - \langle y \rangle)^2}} \tag{1.16}$$

A drawback of this metric is that it is sensitive to outliers in the data. Instead of using the exact floating-point numbers in the evaluation, it is possible to just look at the ranking among the estimates. The predictive index, PI [31] is such a method, and has been used in many of the papers. It is defined as follow

$$\text{PI} = \frac{\sum_i \sum_{j<i} w_{ij} C_{ij}}{\sum_i \sum_{j<i} w_{ij}} \tag{1.17}$$

with

$$w_{ij} = \mid y_i - y_j \mid \tag{1.18}$$

and

$$C_{ij} = \begin{cases} -1, & \text{if } (y_j - y_i)/(x_j - x_i) > 0 \\ 1, & \text{if } (y_j - y_i)/(x_j - x_i) < 0 \\ 0, & \text{if } x_j - x_i = 0 \end{cases} \tag{1.19}$$

The drawback of this metric is the same as for $r^2$, PI can become favourable high if there exists outliers in the data. An alternative is Kendall's $\tau$ rank correlation coefficient [30]

$$\tau = \frac{N_a - N_d}{\frac{1}{2} N(N-1)} \tag{1.20}$$

where $N_a$ is the number of pairs where the ranking of the experimental data and predicted data is in agreement, and $N_d$ is the number of pairs where the ranking is in disagreement. If two pairs have a zero difference or if the difference is not statistical significant, $\tau$ can be adjusted by removing such pairs from the calculation.

In paper XI, the range of experimental affinities was very narrow, only 9 kJ/mol. Therefore, only 29 of the 136 possible pairs of ligands had an affinity difference that was significant at the 95% confidence level. Because it is strictly not possible to determine if the ranking is in agreement between experiments and predictions for such pairs, these pairs were excluded from the calculation of $\tau$. We also removed pairs where the predicted difference was not statistical significant. Finally, only between 8 and 29 pairs were included in the calculation of $\tau$. #

The metrics presented above use the estimates, but not the uncertainty of the estimates. Because, the estimates have uncertainties, the metrics also have uncertainties. For some of them, it can be derived analytically, but for a ranked based metric as PI it is much more trickier to find an analytical solution. The approach that has been used throughout this thesis to estimate the uncertainty of the metrics is based on a statistical sampling technique called bootstrapping [32]. The method works by iteratively re-calculating the value of the metric. In each iteration, a new set of estimates are calculated by drawing random numbers from a Gaussian distribution, centred on the original estimations, and with a width equal to the uncertainty of the estimates. The metric is calculated again with this new set of randomly drawn estimates, and the procedure is repeated in a number of iterations. Typically 1000 iterations are sufficient to reach convergence, and the standard deviation of the re-calculated metric is taken as the uncertainty.

# 2 Molecular modelling

A physically sound model of the system of interest is imperative for a successful estimation of the free energy. Basically, three levels of particle-based models can be identified at an increasing degree of coarse graining. The best theory to model molecules is quantum mechanics (QM) that treats both electrons and nuclei. At the next level, molecular mechanics (MM), the electrons are ignored and only the nuclei are treated. Another level can be created by combining several atoms into a single particle, although such methods have never been used to estimate accurate binding affinities. Therefore, only QM and MM are presented in this chapter.

If it is desirable to treat the system at an atomistic level and this will create a too large system, an alternative is to treat part of the system at an atomistic level whereas the rest of the system, e.g., the solvent molecules, can be treated implicitly. Such approaches are also discussed in this chapter.

## 2.1 Quantum mechanics

A quantum mechanical treatment is the most accurate approach to describe molecules. The atoms are described by a wave function, $\Psi$, which is a function of all the electronic, $\mathbf{r}_i$, and nuclear, $\mathbf{R}_i$, positions. The wave function can be calculated from the time-independent Schrödinger equation

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

where $\hat{H}$ is the Hamiltonian operator and $E$ is the total energy of the system. The Hamiltonian is given by [33] (in atomic units)

$$\hat{H} = -\frac{1}{2}\sum \nabla_i^2 - \frac{1}{2}\sum \frac{1}{M}_A \nabla_A^2 - \sum\sum \frac{Z_A}{\mid \mathbf{r}_i - \mathbf{R}_A \mid} + \sum_i \sum_{j>i} \frac{1}{\mid \mathbf{r}_i - \mathbf{r}_j \mid} + \sum_A \sum_{B>A} \frac{Z_A Z_B}{\mid \mathbf{R}_A - \mathbf{R}_B \mid} \tag{2.2}$$

where indices $i$ and $j$ are over electrons, indices $A$ and $B$ are over nuclei, and $M$ and $Z$ are the masses and charges of the nuclei. The terms on the right-hand-side are the kinetic energy of the electrons, the kinetic energy of the nuclei, the electron–nucleus attraction, the electron–electron repulsion, and the nucleus–nucleus repulsion.

Although the Schrödinger equation can describe any non-relativistic system, it cannot be solved for anything but very small systems. Therefore, approximations are necessary and the most basic is the Born–Oppenheimer approximation. It implies that the electrons are moving around fixed nuclei. Because the nuclei are much heavier than the electrons, the electrons can instantaneously adopt to the nuclear coordinates [33]. Applying this approximation, Eqn 2.2 simplifies to

$$\hat{H} = -\frac{1}{2}\sum \nabla_i^2 - \sum\sum \frac{Z_A}{\mid \mathbf{r}_i - \mathbf{R}_A \mid} + \sum_i \sum_{j>i} \frac{1}{\mid \mathbf{r}_i - \mathbf{r}_j \mid} + U(\mathbf{R}_B) \tag{2.3}$$

where $U(\mathbf{R}_B)$ is the potential energy of the nuclei.

### 2.1.1 The Hartee–Fock method

The most basic method to solve Eqn 2.1 with the Hamiltonian in Eqn 2.3 is the Hartee–Fock (HF) method [34]. It is a rather crude method, but is useful in some application, e.g., to calculate charges for ground-state molecules. The HF method treats the interactions between the electrons in an average way and thereby simplifies the many-body problem in Eqn 2.3 to a one-electron problem.

The total wave function must satisfy the Pauli principle that states that no electrons can have the same set of quantum numbers. The way to obey this requirement is to write the total wave function as a Slater determinant

$$\Psi = \mid \chi_1 \chi_2 \cdots \chi_N \rangle \tag{2.4}$$

where $\chi_i$ is a one-electron orbital, describing the motion of a single electron. The energy of each orbital, $\epsilon_i$ is then solved by introducing the

Fock operator, $\hat{\mathbf{f}}_i$. We can then write

$$\hat{\mathbf{f}}_i\chi_i = \epsilon_i\chi_i \tag{2.5}$$

with

$$\hat{\mathbf{f}}_i = -\frac{1}{2}\nabla_i^2 - \sum \frac{Z_A}{\mid \mathbf{r}_i - \mathbf{R}_A \mid} + v_i^{\mathrm{HF}} \tag{2.6}$$

where $v^{\mathrm{HF}}$ is the average potential of all the other electrons. The total energy of the system is then obtained by summing over all orbitals.

The one-electron orbitals are described by an expansion of a known set of functions. Such a set of functions is called a basis function [34]. A complete basis set is one with an infinite expansion, something that cannot be done practically. Therefore, all basis sets are approximations. The most common basis sets used in QM calculations of molecules consist of atomic functions that are combinations of several Gaussian functions. A minimal basis set is a set that only contains the number of functions required to describe all the filled orbitals in each atom. This is a rather crude approximation, and therefore, it is more common to add two or more functions to describe each orbital. Other functions can be added to improve the molecular picture of the orbitals, viz., polarisation function, or functions that better can describe electrons far out from the nucleus, viz., diffuse functions.

---

In almost all papers in this thesis, the HF method has been used for computing the electrostatic potential (ESP) around the ligands. The ESP is then used to derive partial charges for the ligands. This is the standard approach to derive charges in the Amber force field (see below). #

---

### 2.1.2 Semiempirical QM

Solving the HF equations becomes impractical when there are more than a few hundred atoms in the system. Therefore, it is not practical to calculate QM energies of for instance a full protein. An alternative is then to use a semiempirical QM (SQM) method, which makes several approximations to the HF equations [34].

First, only valence electrons are treated and the core electrons are merged into the nuclei. This assumes that only valence electrons are involved in the chemistry of interest, e.g., bond breaking and bond forming. Second, semi-empirical QM methods uses a minimal basis set that

is based upon Slater-type functions. This makes the calculations faster. Third, some of the interactions between electrons are simply ignored or replaced by empirical parameters.

There exist a lot of different SQM methods that differ mainly in which interactions are simplified. In this thesis we have used three different SQM methods, viz., AM1 [35], RM1 [36], and PM6 [37]. All of these methods are different parametrizations of a more primitive model called modified neglect of diatomic overlap (MNDO). In this model, any overlap between orbitals on three or four atoms are ignored. Furthermore, the overlap between orbitals on two atoms, i.e., the diatomic overlap, is treated differently whether the two atoms are the same or not. The interaction between the electron in the first atom and the core of the second atom are treated with an empirical potential, and the core interaction between two different atoms depend on an empirical overlap function. In addition, some polar bonds are treated with special potentials. The AM1 model mainly tries to improve the description of these core–core terms [34], and RM1 and PM6 are more modern parametrisations of AM1.

To improve the efficiency for large systems such as proteins, there have been several implementations of SQM methods that scales linearly with the number of particles in the system [38, 39].

In paper III, we evaluated AM1, RM1, and PM6 to see if they could be useful in the estimation of binding affinities. All of the Hamiltonians gave similar results, but AM1 was best on average. It was also clear that all of the Hamiltonians required additional, empirical terms to account for hydrogen-bonding and dispersion interactions. #

## 2.2 Molecular mechanics

As explained earlier, the Born–Oppenheimer approximation allow us to separate the nuclear motion from the electronic motion and furthermore, in condensed system at room temperature, the QM effects can mostly be ignored if we are only interested in a single Born–Oppenheimer surface. Therefore, the atoms of the system can be described with Newtonian mechanics.

The potential of the system is described by a force field that consists of a combination of a functional form of the potential and parameters. Most force fields for biomolecules ignore many-body effects and approximates the potential as a pair-potential, although many-body effects, e.g.,

polarisation has been included [40, 41]. The most common force fields for biomolecular simulations are four families of force fields, Amber [42], CHARMM [43], OPLS [44], and Gromos [45]. These force fields differ chiefly in their parameters, although they are based on different philosophies.

The force fields differ also in which atoms they treat. The modern versions of Amber, CHARMM, and OPLS treat all atoms, whereas the non-polar hydrogen atoms in Gromos are merged into the carbon atoms [45]. Such united-atom force fields were common in the 80's and early 90's because they require less computing [40, 46], but were largely abandoned when computer efficiency increased. Nowadays they are mainly used to study process at long time scales, e.g., protein folding [40].

The potential of the Amber force field will serve as an example

$$U_{\text{AMBER}} = \sum_{bonds} k_l(l-l_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{torsions} k_\omega(1+cos(n\omega-\delta)) + \sum_i \sum_{j<i} \epsilon \left[ \left( \frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \sum_i \sum_{j<i} \frac{q_i q_j}{4\pi\epsilon_0 \epsilon_{\text{int}} r_{ij}} \quad (2.7)$$

where three first terms are bonded terms and describe bond stretching, angle bending, and torsion rotation, and the two last terms are non-bonded terms and describe repulsion, dispersion, and electrostatics. All these terms will be described more thoroughly below. Amber will mainly be contrasted with Gromos, because these are the two force fields used in this thesis.

Paper XIV contains a comparison between affinities computed with the Amber and the Gromos force fields, using similar methodologies. The results of the two force fields were highly correlated. In fact, they were more correlated with each other than either of them with experiments. #

### 2.2.1 Bonded terms

As described above, the bonded terms consist of bond stretching, angle bending, and torsion rotation (rotation about a bond). The bond term is a sum of simple harmonic terms, where $k_l$ is the force constant, $l$ is

the actual value and $l_0$ is the ideal value. Gromos uses an anharmonic function instead [45]. The angle term is also harmonic, where $k_\theta$, $\theta$, and $\theta_0$ is the force constant, the actual angle, and the ideal angle, respectively. Contrary to the other force fields, Gromos writes the angle term with the cosine of $\theta$ and $\theta_0$ [45]. The torsion term is written as a cosine series, where $k_\omega$ is the force constant, $n$ is the periodicity, and $\delta$ the phase shift. It is possible that several torsion terms are used to describe the rotation about a bond. (In for instance, OPLS this is written explicitly in the functional form of the force field [44].)

CHARMM introduces a Urey–Bradley term that couples bond and angles motions and a potential that improves the description of the torsions of the protein backbone [43, 47]. Both CHARMM and Gromos introduces an additional, harmonic term for so-called improper torsions that is necessary for the geometry of planar and chiral groups [43, 45], in contrast to Amber and OPLS that treats those as regular torsions.

### 2.2.2 Non-bonded terms

The non-bonded terms consist of a Lennard-Jones potential that describes the exchange-repulsion and dispersion (collectively called van der Waals interactions), and a Coulomb term that describes the electrostatics. These potentials are calculated between all pairs of atoms in the system, with a few exceptions. First, pairs of atoms bonded to each other and pairs of atoms separated by another atom are excluded because these interactions are described by bond and angle terms. Secondly, pairs of atoms separated by three bonds (so-called 1-4 interactions) are scaled down because these interactions are partly described by the torsion term. All force fields employ different scaling. For instance, Amber scales all 1-4 interactions in the Lennard-Jones potential with 2.0 and all such interactions in the Coulomb potential with 1.2 [46], whereas Gromos do not scale the Coulomb potential at all and has several rules for scaling the Lennard-Jones potential [45]. Third, pairs of atoms that are separated by more than a certain cut-off distance are excluded and replaced by some other treatment.

The Lennard-Jones potential can be written in slightly different forms

depending on which parameters exists in the force field [34].

$$\begin{aligned} U_{\mathrm{LJ}} &= \sum_i \sum_{j<i} \epsilon_{ij} \left[ \left( \frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{0ij}}{r_{ij}} \right)^{6} \right] \\ &= \sum_i \sum_{j<i} 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \\ &= \sum_i \sum_{j<i} \left[ \left( \frac{A_{ij}}{r_{ij}} \right)^{12} - \left( \frac{B_{ij}}{r_{ij}} \right)^{6} \right] \end{aligned} \tag{2.8}$$

where $\epsilon_{ij}$ is the negative minimum of the potential, $r_{0ij}$ is the separation at the minimum, $\sigma_{ij}$ the separation where the potential is zero, $A_{ij} = 4\epsilon_{ij}\sigma_{ij}^{12}$, $B_{ij} = 4\epsilon_{ij}\sigma_{ij}^{6}$, and $r_{ij}$ the actual separation between the two atoms. The parameters ($\epsilon_{ij}$, $r_{0ij}$, $\sigma_{ij}$, $A_{ij}$, and $B_{ij}$) are calculated by combining atomic parameters ($\epsilon_{ii}$, $r_{0ii}$, $\sigma_{ii}$, $A_{ii}$, and $B_{ii}$) in a specific way. Amber uses arithmetic and geometric average for the $r_{0ij}$ and $\epsilon_{ii}$ parameters, respectively, viz., $r_{0ij} = 1/2(r_{0ii}+r_{0jj})$ and $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$ [42]. Gromos, on the other hand, store $A_{ii}$ and $B_{ii}$ parameters, and combine them both using geometric averaging [45].

The $r^{-6}$ term describes the dispersion in a system and can in principle be derived from theory [48]. The $r^{-12}$ term that describes repulsion was chosen for computational convenience [34]. Physically, the repulsion is better described by an exponential term, but this cost too much in biomolecular simulations. The van der Waals interactions beyond the cut-off distance can be treated with a continuum approach [49].

The parameters in the Coulomb potential, is simply partial charges of the two atoms, $q_i$ and $q_j$. $\epsilon_0$ is the relative permittivity, and $\epsilon_{\mathrm{int}}$ is the dielectric constant of the media, which is usually set to unity in simulations. To primitively model solvent effects it can be made distance-dependent [34]. The electrostatic interactions are very important for correct energies and therefore they must be treated rather accurately also beyond the cut-off distance. In a periodic simulation, the electrostatics are most commonly replaced by a Ewald summation technique that treats the long-range interaction in reciprocal space using fast Fourier transform [34, 50]. However, periodic simulations with the Gromos force field are normally employed with a reaction field [45]. In non-period simulation, a reaction field is the most common approach. In this thesis we have used an approximation called local reaction field (LRF) [51], to calculate long-range electrostatics.

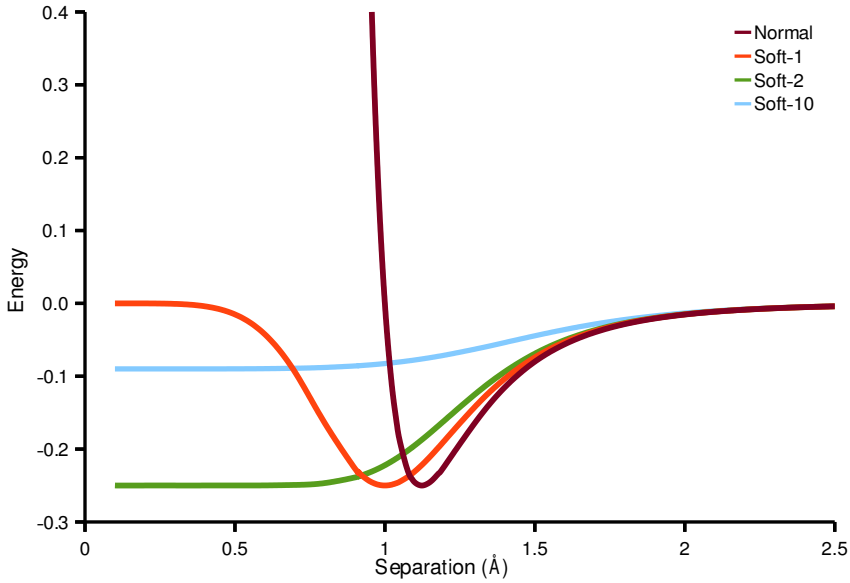


**Figure 2.1:** Lennard-Jones potential. Both the normal and soft-core versions are included. The soft-core potentials are indicated with the value of $\alpha$. The parameters $A$ and $B$ were set to 1, and $\lambda = 0$.

Both the Lennard-Jones and Coulomb potentials create rugged landscapes that may become problematic in free energy calculations. For instance, they cause the so-called van der Waals end-point problem [52, 53], which can lead to computational instabilities. A common approach is to introduce soft-core versions of these potentials [54, 55]. The soft-core Lennard-Jones potential in the Q simulation package is [56]

$$U_{\mathrm{LJ}} = \sum_i \sum_{j<i} \left[ \frac{A_{ij}}{(r^6 + (1-\lambda)\alpha_{ij})^2} - \frac{B_{ij}}{r^6 + (1-\lambda)\alpha_{ij}} \right] \tag{2.9}$$

and I have implemented a corresponding soft-core Coulomb potential in this package

$$U_{\mathrm{Coul}} = \sum_i \sum_{j<i} \frac{q_i q_j}{4\pi\epsilon_0 \sqrt{r^2 + (1-\lambda)\alpha_{ij}}} \tag{2.10}$$

where $\alpha$ is the softness parameter and $\lambda$ is a coupling-parameter (see Chapter 4.1). The value of $\alpha$ is not well-defined, although a recent study tried to optimized it [57]. Soft-core Lennard-Jones potentials are illustrated in Figure 2.1

In papers XII and XIII, soft-core versions of the Lennard-Jones and Coulomb potentials were used. We compared an approach that uses only the soft-core Lennard-Jones potential with an approach that uses soft-core for both van der Waals and electrostatic interactions. It was found that both approaches performed equally well, although the approach with both soft-core potentials seemed to be slightly more stable.

In paper XIV, we tested an approach in which the soft-core potentials are used as a clever way to represent several different atoms with a single unphysical, reference state. The reference state is then perturbed to a real ligand after the simulation. This worked well for non-polar perturbations, but less satisfactorily for polar perturbations. #

### 2.2.3 Parametrisation

When the functional form of the force field has been decided, the parameters in the force field have to be determined. Here the force fields differ again, at least for the non-bonded parameters. The bond and angle equilibrium and force constant parameters are usually taken from crystal structures of small compounds and spectroscopic data, respectively. The torsional parameters are normally fitted to a QM-calculated potential together with the non-bonded interactions [34].

The Lennard-Jones parameters are difficult to parametrise. Historically, these were taken from crystal data for the repulsion term and from atomic polarisabilities for the dispersion term. Nowadays, they are more commonly fitted to experiment liquid properties such as heat of vaporization by running simulations of model compounds [34, 42, 45]. The charges in the Amber and CHARMM force fields are taken from different QM calculations [42, 43], whereas Gromos and OPLS fit the charges so that the force field can reproduce experimental quantities such as heat of vaporization and densities of pure liquids [45, 44].

There usually exist separate sets of parameters for each type of macromolecule, e.g., protein and nucleotides. The parameters for small drug-like compounds are usually taken from the corresponding macromolecular force field, but the latest decade has seen an increase in force fields especially made for small molecules that should work with the macromolecular potential [58, 59, 60].

### 2.2.4 Potentials for liquid water

Interactions with solvent dominate the cost of standard MM simulations. Therefore, the solvent is often treated specially. Because I have used simulations in liquid water throughout the thesis I will shortly describe some different water potentials. The simplest water potential used in this thesis is a so-called three-point water molecule. It contains three charge sites, one on each atom, but only a single Lennard-Jones site on the oxygen. Two different parametrisations have been used, TIP3P [61] and SPC [62]. A four-site potential has also been used, in which the charge of the oxygen is moved away from the hydrogen atoms to mimic the lone-pairs on the oxygen. The parametrization used was TIP4P-Ewald [63], which is a re-parametrization of the original TIP4P [61] to work better in periodic simulations.

## 2.3 Implicit solvent models

One of the most common methods to reduce the complexity of the system and improve the convergence is to treat the water molecules implicitly. The most popular approach, which originates from the work of Max Born, is to treat the solvent as a dielectric continuum [64]. Such methods can be used with both QM and MM methods, and are based on several parameters. Too proceed, it is instructive to introduce a phenomenological approach to solvation [65]. In this approach, the solvation process is divided into two phases. First, a cavity is created in the solvent that precisely can accommodate the solute. Second, the solute is introduced in the cavity and the interaction between the solute and the solvent is turned on. If we assume a MM description of the interaction, we can write the solvation free energy as

$$G_{\text{solv}} = G_{\text{cav}} + E_{\text{rep}} + E_{\text{disp}} + E_{\text{ele}} \tag{2.11}$$

where the first term is the free energy of creating the cavity in the solvent, and the next three terms are the repulsion, dispersion, and electrostatics interaction energy. The last three terms on the right-hand-side of Eqn 2.11 are written as energies rather than free energies, which stems from an assumption that the introduction of solvent–solute interactions does not perturbe the solvent structure. While this could be a decent assumption for the repulsion and dispersion interactions, this is most likely not the case for the electrostatic interaction [65]. Most of the continuum methods separate the calculations in a polar, $G_{\text{pol}}$, which is taken as the last term

of 2.11 and a non-polar, $G_{\text{np}}$ term, which is taken as the first three terms of the right-hand-side of Eqn 2.11 such that

$$G_{\text{solv}} = G_{\text{pol}} + G_{\text{np}} \tag{2.12}$$

Other approaches to implicit modeling of solvent exists; The three-dimensional reference interaction site model (3D-RISM) [66], is a model based on classical density theory, and is therefore not a continuum model. Another model that is intermediate to explicit and continuum methods is the Langevin dipoles (LD) model [67].

### 2.3.1 Polar solvation

Any treatment of a dielectric continuum starts by considering the Poisson equation for this medium [68]

$$\nabla \left[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})\right] = -4\pi\rho(\mathbf{r}) \tag{2.13}$$

where $\epsilon$ is the dielectric constant, $\phi$ the electrostatic potential and $\rho$ the charge density. A Boltzmann factor can be added to the equation to account for ionic strength, giving the Poisson–Boltzmann (PB) equation. This is is seldom done in biomolecular calculations [64] but approaches based on Eqn 2.13 is nonetheless referred to as PB methods. Eqn 2.13 is usually solved by finite difference methods by discretising the charge distribution and the dielectric constant on a grid [69].

The solvation free energy is obtained by calculating the potential at two dielectric constants of the solvent, 1 and 80, representing the vacuum and water environment, respectively. The difference between these potentials is called a reaction field, $\phi_{\text{reac}}$, and can be used to compute the solvation free energy. If the charge distribution is represented by a set of point charges as in a force field, the equation is [69]

$$G_{\text{pol}} = \frac{1}{2}\sum q_i\phi_{\text{reac}}(\mathbf{r}_i) \tag{2.14}$$

In the case of a single ion of radius $a$ in pure solvent, this reduces to the Born formula

$$G_{\text{Born}} = -\frac{q^2}{2a}\left(1 - \frac{1}{\epsilon_{\text{solv}}}\right) \tag{2.15}$$

where $\epsilon_{\text{solv}}$ is the dielectric constant of the solvent. The generalised Born (GB) methods attempt to generalise this expression by picturing a molecule as a set of spheres with charges and radii. If the separation,

$r_{ij}$, between any two atoms is large compared to their radii, it is possible to write down a sum of individual Born terms and pair-wise Coulomb terms [70]

$$G_{\text{pol}} = \sum \frac{q_i^2}{2a_i}\left(\frac{1}{\epsilon_{\text{solv}}} - 1\right) + \frac{1}{2}\sum_i \sum_{i \neq j} \frac{q_i q_j}{r_{ij}}\left(\frac{1}{\epsilon_{\text{solv}}} - 1\right) \tag{2.16}$$

The GB models then proceed one step further by merging the Born and Coulomb terms

$$G_{\text{pol}} = \frac{1}{2}\left(\frac{1}{\epsilon_{\text{solv}}} - 1\right) \sum \sum \frac{q_i q_j}{f_{\text{GB}}} \tag{2.17}$$

where $f_{\text{GB}}$ is an approximate function of the separation between atoms and their "effective" radii. The different GB methods mainly differ in how this function is calculated [70].

Another set of continuum methods do not start from the Poission equation but rather from an apparent surface charge density, $\sigma(\mathbf{r})$ [64, 71]. It can be shown that this density is the only source of the solvent reaction field and hence determines fully the solvation free energy. This is the basis of the methods most commonly used in QM calculations, viz., the polarised continuum model (PCM) [72] and conductor-like screening model (COSMO) [73].

Because these methods are based on the charge density on the surface of the solute, these methods start with a determination of surface segments called tesserae.The surface charge density can thereafter be derived by considering the dielectric boundary conditions at the surface [71] and it is those boundary conditions that differ between the various methods. In the original PCM implementation (also called dielectric-PCM), a dielectric continuum is assumed outside the solute cavity, whereas in COSMO (and the related conductor-PCM method), a conductor, i.e., a medium with an infinite dielectric constant, is assumed. In the latter, it is therefore necessary to scale the calculated charge density such that it represents a dielectric medium. This approximation is most accurate at high dielectric constants. In passing, it should also be mentioned that there exists other variants of these approaches [71, 74].

In both PCM and COSMO, the surface charges can be found iteratively or by matrix manipulations, and once they have been determined the electrostatic potential due to the charges can be calculated by summing over all tesserae. This potential is then added as a perturbation to the system [71].

In paper V, a special implementation of the PCM model was used that is able to calculate the solvation free energy of proteins.

In paper III, the COSMO method was used as the solvent model because it is implemented in the SQM software used. #

### 2.3.2 Non-polar solvation

In the PCM method, the cavitation, dispersion, and repulsion contributions to the non-polar solvation free energy are calculated by three separate terms. [75]. The cavitation free energy is calculated from scaled-particle theory [71, 76]

$$G_{\text{cav}} = \sum_{\text{atoms}} \frac{A_i}{4\pi R_i} G^{\text{HS}}(R_i) \tag{2.18}$$

with

$$G^{\text{HS}}(r) = RT[K_0 + K_1(r+R_{\text{solv}}) + K_2(r+R_{\text{solv}})^2 + K_3(r+R_{\text{solv}})^3] \tag{2.19}$$

where $R_i$ is the radius of an atom, $A_i$ the total area of all tesserae on that atom, $R_{\text{solv}}$ the radius of the solvent molecule and $K_x$ are terms that depends on the radius of atom i to the x:th power. Hence, this expression depends on the radius of the atoms to the powers 0 to 3.

The dispersion and repulsion energies are calculated from volume integrals [77, 78]

$$U_{\text{LJ,cont}} = \sum \rho \int U_{\text{LJ}}(\mathbf{r}) d\mathbf{r} \tag{2.20}$$

where the sum is over all solute atoms and the integration is over all solvent-occupied volume, $\rho$ is the uniform solvent number density, and the Lennard-Jones potential is evaluated between the solute atom and the "solvent atom" (the oxygen in case of water). The volume integral can then be cast into an integration over the solvent-accessible surface through the use of the divergence theorem

$$U_{\text{LJ,cont}} = \rho \sum_a \sum_b \int_{S_b} \left[ \frac{A_{ab}}{6r_{sa}^{12}} - \frac{B_{ab}}{3r_{sa}^{6}} \right] r_{sa} \mathbf{n}_s d\sigma_s \tag{2.21}$$

where the double sum goes over a pair of atoms, $a$ and $b$, the integration is over all the solvent-accessible surface of atom $b$, $A$ and $B$ are the usual

Lennard-Jones parameters, $r$ is the separation between atom $a$ and the surface, and $\mathbf{n}_s$ is a normal vector to the surface.

It has been noted that the cavitation and repulsion energies show a dependence on the solvent-accessible surface-area or surface-volume. Therefore, a recent approach merged these into a single term, giving the cavity-dispersion (CD) method [78]

$$G_{\mathrm{CD}} = G_{\mathrm{cav}} + E_{\mathrm{disp}} = \gamma_{\mathrm{CD}}\mathrm{MS} + b_{\mathrm{CD}} + E_{\mathrm{disp}} \tag{2.22}$$

where MS is some kind of molecule surface, and $\gamma_{\mathrm{CD}}$ and $b_{\mathrm{CD}}$ are empirical parameters fitted to results from rigorous simulations. For small molecules, it does not matter what kind of surface is used, but for larger molecules the solvent-accessible surface-area (SASA) has been recommended [78, 79].

Although the PCM are more rigorous from a theoretical perspective, the most popular methods to compute the non-polar solvation free energy is to relate the entire free energy to the SASA [80].

$$G_{\mathrm{np}} = \gamma_{\mathrm{SASA}}\mathrm{SASA} + b_{\mathrm{SASA}} \tag{2.23}$$

where $\gamma_{\mathrm{SASA}}$ and $b_{\mathrm{SASA}}$ are empirical parameters fitted to experimental solvation free energies of hydrocarbons [81]. There exists several sets of parameters, and sometimes atom-specific parameters are used [82].

In papers V and VI, we evaluated the three different non-polar solvation methods described here, on their ability to predict the change upon ligand-binding. It was found that none of the methods were able to give accurate predictions on a wide range of systems with different active site hydration. On average, SASA performs best, but only because the predictions are smaller than the other two methods. The PCM method could be improved by introducing explicit interactions from the simulations, but the set of test cases was to small too conclude if this is a general approach. #

### 2.3.3 3D-RISM

In 3D-RISM, the solvent is not assumed to be a structureless dielectric medium. Rather, the solvent structure is described by the probability density $\rho_\gamma g_\gamma(\mathbf{r})$ of finding a solvent site, $\gamma$, in the 3D space at position $\mathbf{r}$, where $\rho_\gamma$ is the average number density and $g_\gamma$ is the normalized distribution function [66]. If we are interested in water, $\gamma$ is one of the hydrogen

atoms or the oxygen atom. If we introduce the total correlation function $h_\gamma + 1 = g_\gamma$, we can write down the 3D-RISM integral equation

$$h_\gamma(\mathbf{r}) = \sum \int c_\alpha(\mathbf{r} - \mathbf{r}')\chi_{\gamma\alpha}(\mathbf{r}')d\mathbf{r}' \tag{2.24}$$

where $c$ is the so-called direct correlation function, $\chi$ is the chemical susceptibility, which is taken as a parameter of the model and is precomputed. The sum is taken over all solvent sites and the integral over all space. Now we have two unknown functions $h$ and $c$. To solve the equation, a closure, a relationship between $h$ and $c$, is introduced and a common choice in 3D-RISM is the Kovalenko–Hirata closure [83]:

$$g_\gamma(\mathbf{r}) = \begin{cases} \exp(d_\gamma(\mathbf{r})), & \text{for } d_\gamma(\mathbf{r}) \leq 0 \\ 1 + d_\gamma(\mathbf{r}), & \text{for } d_\gamma(\mathbf{r}) > 0 \end{cases} \tag{2.25}$$

$$d_\gamma(\mathbf{r}) = -U_\gamma(r)/RT + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})$$

where $U$ is the potential energy from a force field. The 3D-RISM equation is then solved by discretizing the solute–solvent interaction potential and the correlation functions on a grid. The solvation free energy is finally computed from the correlation functions using statistical-mechanical formulas [66].

In paper IV, the 3D-RISM method was introduced in an approach to calculate binding affinities. In the same paper, we also tested two different PB implementations, and four different GB methods. It was found that many of the methods give relative predictions that are rather similar. However, the choice of solvation method totally dictates the absolute predictions, with differences up to 200 kJ/mol. We also found no benefit of using 3D-RISM instead of the much faster PB and GB methods. #

### 2.3.4 Langevin dipoles

The Langevin dipole (LD) model is a grid-based method. The solvent is represented by rotable dipoles on each grid point that does not overlap with the solute, and the orientation of the dipoles is calculated using the Langevin equation. The solute, e.g., a protein could be represented by dipoles, forming the protein-dipoles–Langevin-dipoles (PDLD) method [67]. This method has been used successfully, but has also drawn considerably criticism. The main problem is the grid that has to be sufficiently

fine so that the calculations do not depend on the placement of the grid [64]. Furthermore, the results can diverge for realistic magnitudes of the dipole of a water molecule.

# 3 Sampling

The are mainly two methods that are used to sample the configurational space of a macromolecule, molecular dynamics (MD) and Metropolis Monte Carlo [84]. Noteworthy, is also the mining minima method [85] of Mike Gilson that recently have been extended to protein–ligand calculations [86]. Metropolis Monte Carlo works by performing a random walk in space, subjected to a Metropolis test to determine whether the new configuration should be rejected or not [87]. This method has only been used in a single paper in this thesis for preparing structures and will therefore not be discussed any further. However, MD has been used in all papers and is described below.

When MD has been presented, I will discuss how to simulate a bulk-like system. Periodic and non-periodic simulations are compared. When discussing the non-periodic simulations, I will also introduce truncation approaches. The chapter ends with a discussion on different sampling strategies.

---

In paper II, a Monte Carlo method was used to sample different rotational, conformational, and protonation states of amino acid side-chains, and rotation of crystal water molecules. The different structures created in the Monte Carlo sampling was then used as starting structures for MD simulations. #

---

## 3.1 Molecular dynamics

MD refers to a sampling technique that employs Newton's second law [88] to propagate the system in time. For a single particle $i$, with a mass $m_i$ at a position $q_i$ (I will use a generalised coordinate here, which could be an $x$, $y$, or $z$ coordinate if we are in 3D Cartesian space), the second law reads

$$F_i = m_i \frac{d^2 q_i}{dt^2} \tag{3.1}$$

where $F_i$ is the force on the particle, which can be obtained by differentiating the potential, viz., $F_i = dU/dq_i$. The position after a finite time, $\Delta t$, can be computed using a simple Taylor expansion [84, 34]

$$q(t + \Delta t) = q(t) + \frac{dq(t)}{dt}\Delta t + \frac{d^2 q(t)}{dt^2}\frac{\Delta t^2}{2} + \cdots \tag{3.2}$$

Hence, the position $q(t)$, the velocity $dq(t)/dt$, and the acceleration $d^2q(t)/dt^2$ are sufficient for propagating the system, if the higher order terms are treated in some approximate way. The acceleration is given by Eqn 3.1.

An MD simulation starts by assigning velocities and positions to all atoms in the system. Thereafter, the force is calculated on each atom, which in turn tells in what direction the atom should move. This information is used to update the position of each atom, and the procedure is repeated.

### 3.1.1 Integration of motion

There exist numerous algorithms for integrating the equations of motions [87, 34]. A common example is the simple Verlet algorithm [89] but the drawback of this method is that it does not explicitly propagate the veclocities. An alternative is to use a leapfrog [90] or velocity Verlet algorithms [91]. However, it is out of the scope of this thesis to discuss these algorithms at any length.

It is however pertinent to discuss the size of the time step, $\Delta t$, i.e., the time that passes between a particle movement. As a rule of thumb, the time step should be so small that all the motions of the system can be described. The fastest motion in macromolecular system is the vibration of bonds to hydrogen atoms, which is on the order of 10 fs [34]. Therefore, the time step is usually set to 0.5 to 1 fs, if such vibrations are allowed. However, there exist algorithms, such as SHAKE [92] and LINCS [93],

that constrain the motion of certain bonds and most macromolecular simulations are performed with such restraints on bonds involving hydrogen atoms. Therefore, the time step is usually set to 2 fs, which of course increases the efficiency of the simulation.

### 3.1.2 Sampling in different ensembles

A thermodynamic ensemble is a collection of all possible systems that have different microscopic states but belong to the same macroscopic or thermodynamic state. An ensemble is usually denoted by the thermodynamic quantities that are constant, e.g., number of particles, volume, and pressure [94]. Because an MD simulation obeys the Newtonian laws it can be shown that such a simulation will sample configurations in an ensemble with constant number of particles, constant volume, and constant total energy (potential plus kinetic energy), i.e., the microcanonical ensemble [87].

The purpose of simulations that estimate binding affinities are to reproduce affinities that can be measured experimentally. An experiment that measures a ligand-binding affinity cannot be performed at constant energy, and hence plain MD simulation would be useless for this application. Normally, the experiments are performed at constant pressure and constant temperature, i.e., the isobaric-isothermal ensemble, and luckily there exists algorithms that keep those quantities constant in an MD simulation.

An algorithm that keeps the temperature constant in an MD simulation is called a thermostat, and there exist a lot of them [95]. In this thesis we have mostly used a Langevin thermostat [96] that actually let us perform Langevin dynamics, rather than Newtonian dynamics. The equation of motion is then modified to the following equation [84]

$$m_i \frac{d^2 q_i}{dt^2} = F_i - \zeta \frac{dq}{dt} + R_i(t) \tag{3.3}$$

where the two additional terms on the right-hand-side are a frictional term ($\zeta$ is a friction constant) and a random-force term. Through a relationship between $\zeta$ and $R$, the temperature is kept constant. Another thermostat that has been used is the Berendsen, or weak-coupling algorithm [97], which is an algorithm to re-scale the velocities such that the temperature is kept constant. However, it has been shown that this thermostat does not produce a fully correct thermodynamic ensemble [87].

An algorithm that keeps the pressure constant in a simulation is called a barostat and chiefly works by adjusting the volume of the simulated system [87]. In this thesis, we have exclusively used a weak-coupling algorithm [97].

## 3.2 The simulated environment

Ultimately, we would like to simulate a bulk-like system because we would like to reproduce experimental quantities that are obtained under such conditions. However, it is not feasible to simulate so many particles and hence we need to make approximations. Nowadays, any biomolecular simulation that is used to estimate affinities solvate the solute in some tens of thousands of water molecules and use cleaver tricks to mimic bulk behaviour.

### 3.2.1 Periodic simulations

To approximate a bulk system, most of the biomolecular systems are today simulated by imposing periodic boundary conditions [84]. This implies that the solute is solvated in a box of solvent molecules. This simulation box is then replicated infinitely in all directions. Therefore, if an atom drifts outside the simulation box it will end up in an image of the box, i.e., it will appear on the opposite side of the box. The minimum-image convention makes sure that interactions are not double counted, by only counting the shortest distance between a pair of atoms, irrespectively if they are in the central box or in an image [34].

The simulation box is not restricted to a specific shape. Whenever periodicity was imposed in this thesis, a truncated octahedron was used. Such a shape reduces the number of solvent molecules that need to be simulated, compared to a simple cubic box.

### 3.2.2 Non-periodic simulations

Another possibility that is used by some research groups is to solvate the solute in a sphere, a droplet, of solvent molecules. Typically, there is vacuum outside the droplet, and therefore, the simulated system must be treated in a special way to reproduce bulk-like behaviour [84].

An early approach was to simulate the outer region using Langevin dynamics, so-called stochastic boundary conditions [98]. The Langevin dynamics imposes friction on the inner region so that the droplet does not evaporate. The outer region is also included in a buffer of fixed atoms

so that the atoms in the outer region are fairly fixed in space. However, such an approach produces artificial density fluctuations and can alter the structure of solute [84].

Another approach is to impose special restraints on the water molecules and two different types of restraints have been used. First, a radial potential is added to the water molecules that prevent them from evaporating. Several empirical potentials have been suggested [99, 100, 101]. Secondly, the orientation of the water molecules is heavily affected by the outside vacuum, and potentials are added to restore bulk-like distributions. One such approach is the SCAAS (surface-constrained all-atom solvent) of Warshel [101], in which a uniform distribution is imposed on the angle between the water dipole vector and the displacement vector from the origin. Essex and Jorgensen proposed a similar method, but they found it also necessary to restrain the vector perpendicular to the plane of the water molecule [100].

Instead of treating the outside as a vacuum, it can be described using the continuum methods described in the previous chapter. Roux and co-workers introduced such a method, called the general solvent boundary potential [102, 103]. Simonson and co-workers introduced a method that is a combination of vacuum simulations and continuum solvation. A simulation is performed in vacuum and the snapshots from the simulation are subsequently corrected for by continuum electrostatics [104].

---

In paper XIII, non-periodic simulations are tested and compared to previous results computed with periodic simulations. It is shown that the non-periodic and periodic results correlate well. The size-dependence is evaluated by making the water droplet smaller and smaller and truncating protein residues outside the droplet. It is shown that a sphere of 15 Å in radius, is sufficient for an accurate free energy estimate.

Non-periodic simulations were also employed in paper XIII, on the test of the method linear interaction energy, which most often is performed in a non-periodic setting. #

---

### 3.2.3 System truncation

Although the most common approach is periodic simulations, it becomes impractical if the system is very large and the interesting chemistry only occurs in a small region of space. Ligand-binding for instance occurs in an active site that typically has a diameter less than 20 Å.

In such a situation, non-periodic simulations can be very effective. Instead of solvating the entire protein in a water droplet, the droplet is centred on the active site and is made smaller than the protein. The residues outside the droplet are then excluded from the simulation.

## 3.3 Sampling strategies

A good simulation should sample the configurational space according to the prescribed distribution, e.g., a Boltzmann distribution if we simulate the canonical ensemble [105]. The $N$ samples extracted from the simulation is then used to compute a time average of some quantity $A$,

$$\langle A \rangle = \frac{1}{N} \sum_{\text{snapshots}} A_i \tag{3.4}$$

As explained in Chapter 1.3, the reliability of this estimate is given by the variance if we for a moment ignore the bias (which is hard to decrease by sampling). To obtain a valid estimate of the variance, the snapshots sampled should be independent of each other [105, 106]. Therefore, we need a way to determine the sampling frequency. The correlation time, $\tau$, is a measure of the simulation time required before the simulation loses "memory" of previous values of $A_i$. The number of independent samples is then, $N_{ind} \sim t_{sim}/\tau$, where $t_{sim}$ is the total simulation time. If $\tau$ is known, it can be used to correct the estimate of the variance [107], or one could simply sample snapshots at intervals of $\tau$.

### 3.3.1 Estimating the correlation time

One approach is to calculate $\tau$ from the normalized autocorrelation function (ACF), but the estimation of the ACF for a finite data series is non trivial [106, 108, 109]. Whenever, $\tau$ has been estimated in this thesis, the method of statistical inefficiency [110, 111] has been used. In this procedure, the following quantity is calculated

$$\Phi = \frac{m\sigma^2(B)_m}{\sigma^2(A)} \tag{3.5}$$

where $\sigma^2(A)$ is the variance of the time series $\{A\}$ (the average of this time series is shown in Eqn 3.4), and $\sigma^2(B)_m$ is the variance of the block average of $\{A\}$, where the block length is $m$. This block average is calculated from

$$B_i = \frac{1}{m} \sum_{j=n-im+1}^{n-(i-1)m} A_j \tag{3.6}$$

That is, $\{A\}$ is divided into a number of non-overlapping series, each with a length of $m$. Once $m$ is so large that the successive values of $B_i$ are statistical independent, $\Phi$ will become a constant and $m$ will be an estimate of the correlation time.

Example uses of the method of statistical inefficiency can be found in papers I and VIII. #

### 3.3.2 Independent simulations

The above procedure as well as the determination of ACF is sensitive to correlations at long time-scales [105]. Therefore, it is very difficult to obtain an accurate estimate of $\tau$. An alternative and effective sampling strategy is then to start $M$ independent simulations and to calculate a statistical ensemble average [105]. The ensemble average of some quantity $A$ can be written

$$\langle A \rangle = \frac{1}{M} \sum_{\text{simulations}} \frac{1}{N'} \sum_{\text{snapshots}} A_i \tag{3.7}$$

Usually, $N > N'$. The ergodic hypothesis of thermodynamics suggests that the two averages in Eqn 3.4 and Eqn 3.7 are identical with perfect sampling [94]. However, it is truly hard to have perfect sampling of a biomolecular system, and hence these two averages can be considerably different. In particular, the precision of the two will be different, because in Eqn 3.4 it is inversely proportional to $\sqrt{N}$, whereas in Eqn 3.7, it is inversely proportional to $\sqrt{M}$, although it also depends on $N'$.

In Paper I, we compared the time-average and the ensemble-average approaches, i.e., we tested whether it was better to run a single long simulation or several shorter. As a test case, we used the avidin protein, which is a tetramer with four identical subunits. A converged free energy estimate should therefore give the same estimate for all four subunits. A single long simulation failed to give this answer, because the precision was underestimated. However, using several short independent simulations the same affinity was obtained within the statistical uncertainty. Therefore, it is better to run several short simulations. It is then also straightforward to improve the precision by running more simulations. #

The most common method to generate independent simulations, is simply to assign different starting velocities to the atoms [112, 113, 114], which here is denoted velocity-induced independent trajectories (VIIT) approach. This must be considered as a rather small perturbation to the system because none of the positions are changed. However, a few studies have used different crystal structures or an NMR ensemble as a starting point for MD [115, 116, 117]. Even a limited conformational search has been used as preparatory step [116]. In Paper II, we suggested three other approaches, which employs the uncertainty in the setting up of macromolecular simulations to increase the sampling of the phase space.

1. Solvent-induced independent trajectories (SIIT)
2. Conformation, rotation, and protonation-induced independent trajectories (CRPIIT)
3. Alternative conformation-induced independent trajectories (ACIIT)

In SIIT, the solute is solvated in different water boxes for each independent trajectory. The water boxes can be created by any means, but in Paper II, they were taken from MD simulations. A similar approach has been used previously [118]. In CRPIIT, different rotational, conformational, and protonation state of amino acid side chains are sampled with MC, and used as starting structures for the MD simulations. Finally, ACIIT takes advantage of the fact that many high-resolution crystal structure contains alternative conformation for several of the amino acids.

---

In paper II, we compared VIIT, SIIT, CRPIIT, and ACIIT. The conclusion was that SIIT and ACIIT gave improved sampling. Therefore, it is recommended to use SIIT because it can always be used. However, ACIIT can only be used if there exists alternative conformations in the crystal structure. The CRPIIT approach was little bit more problematic, because the choice of protonation, rotation, etc. is not arbitrary. However, it was clear the the sampling of active site residues should be avoided, otherwise unwanted effects can be introduced. #

---

# 4 Free-energy estimates

After a discussion on molecular models and sampling, I now return to the core of this thesis, namely free-energy estimates. As will be clear at the end of this chapter, all free-energy estimates are based upon differences between various states. Therefore, this chapter starts by deriving three rigorous approaches for calculating free-energy differences. These techniques can be seen as tools that are employed by approaches to calculate standard (absolute) and relative binding free energies. Such approaches are discussed thereafter.

I will end this chapter with a thorough discussion on approximate free-energy methods and my intention is to highlight similarities and differences. At the end of this chapter, the most popular methods to compute ligand-binding affinities from equilibrium simulations have been covered. Non-equilibrium methods will not be covered at all; a discussion of such methods can be found in [27].

## 4.1 Basic techniques

In this section, I will consider two Hamiltonians, $H_0$ and $H_1$, with the following relationship

$$H_1 = H_0 + \Delta H \tag{4.1}$$

where $\Delta H$ is the difference between the two Hamiltonians. These two Hamiltonians could for instance describe benzene free in solution and phenol free in solution, respectively. $\Delta H$ would then contain, the interaction between the solvent molecules and the hydroxyl group of phenol, and some intramolecular terms. The problem now is how the free-energy

difference between these two systems can be computed. In what follows, I will describe the three most common approaches to solve this issue.

### 4.1.1 Thermodynamic integration

In thermodynamic integration (TI) [4], a composite Hamiltonian, $H$, is introduced that depends on a coupling parameter, $\lambda$. The requirement for this Hamiltonian is that when $\lambda = 0$, $H = H_0$, and when $\lambda = 1$, $H = H_1$. There are different functional forms that satisfy this requirement [107], but in this thesis, I have only used the linear relationship

$$H(\lambda) = (1 - \lambda)H_0 + \lambda H_1 \tag{4.2}$$

The derivation of TI [34] now follows by differentiating the free energy, $G$, with respect to $\lambda$

$$\frac{\partial G(\lambda)}{\partial \lambda} = -RT\frac{\partial \ln Q(\lambda)}{\partial \lambda} = -RT\frac{1}{Q(\lambda)}\frac{\partial Q(\lambda)}{\partial \lambda} \tag{4.3}$$

where $Q$ is the canonical partition function, which naturally depends on $\lambda$ as well:

$$Q(\lambda) = \xi \int_q \int_p d\mathbf{q}d\mathbf{p}e^{-\beta H(\lambda)} \tag{4.4}$$

where the integration is over all coordinates, $q$, and over all momenta $p$, and $\xi$ is a pre-factor that depends on the number of particles in the system. Differentiating $Q$ with respect to $\lambda$ gives

$$\frac{\partial Q(\lambda)}{\partial \lambda} = \xi \int_q \int_p d\mathbf{q}d\mathbf{p}\frac{\partial}{\partial \lambda}e^{-\beta H(\lambda)} = -\xi\beta \int_q \int_p d\mathbf{q}d\mathbf{p}\frac{\partial H(\lambda)}{\partial \lambda}e^{-\beta H(\lambda)} \tag{4.5}$$

and substituting back into Eqn 4.3 we have

$$\frac{\partial G(\lambda)}{\partial \lambda} = \frac{1}{Q(\lambda)}\xi \int_q \int_p d\mathbf{q}d\mathbf{p}\frac{\partial H(\lambda)}{\partial \lambda}e^{-\beta H(\lambda)} \tag{4.6}$$

The last expression can be identified as an ensemble average of $\frac{\partial H(\lambda)}{\partial \lambda}$, hence

$$\frac{\partial G(\lambda)}{\partial \lambda} = \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda \tag{4.7}$$

where the subscript indicates that it is an ensemble average for a fixed $\lambda$. Finally, by integrating from 0 to 1, the free-energy difference can be computed [34]

$$\Delta G = G_1 - G_0 = \int_0^1 \frac{\partial G(\lambda)}{\partial \lambda}d\lambda = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{4.8}$$

This integration is usually computed by a finite-difference method such a trapezoid integration [119], and therefore, simulations are performed at specific values of $\lambda$.

### 4.1.2 Free-energy perturbation

Alternatively, the free-energy difference, $\Delta G$, between two states can be written as a ratio of two partition functions

$$\Delta G = G_1 - G_0 = -RT \ln \frac{Q_1}{Q_0} \tag{4.9}$$

where $Q$ is the canonical partition function, written analogously to Eqn 4.4. Inserting the expression for the partition functions into Eqn 4.9 and ignoring pre-factors, we have [27]

$$\Delta G = -RT \ln \frac{\int_q \int_p d\mathbf{q} d\mathbf{p} e^{-\beta H_1}}{\int_q \int_p d\mathbf{q} d\mathbf{p} e^{-\beta H_0}} = -RT \ln \frac{\int_q \int_p d\mathbf{q} d\mathbf{p} e^{-\beta \Delta H} e^{-\beta H_0}}{\int_q \int_p d\mathbf{q} d\mathbf{p} e^{-\beta H_0}} \tag{4.10}$$

the last expression can be identified as an ensemble average of $e^{-\beta \Delta H}$, hence

$$\Delta G = -RT \ln \left\langle e^{-\beta \Delta H} \right\rangle_0 \tag{4.11}$$

where the subscript indicate that the average should be taken from a simulation of system 0. The contribution from the momenta is cancelled out if the number of atoms in the two systems is identical [27], and therefore, Eqn 4.11 is further simplified to

$$\Delta G = -RT \ln \left\langle e^{-\beta \Delta U} \right\rangle_0 \tag{4.12}$$

where $\Delta U$ is the potential energy difference. Eqn 4.12 is usually called the forward equation, because the perturbation is from 0 to 1. By reversing the process, we can write down the backward equation, as

$$\Delta G = RT \ln \left\langle e^{\beta \Delta U} \right\rangle_1 \tag{4.13}$$

where the simulation should be of system 1.

Now we can ask ourselves if Eqns 4.12 or 4.13 could be used to compute the free energy difference of the system discussed above, i.e., between benzene and phenol free in solution. The general answer is no. $\Delta U$ will in most of the sampled configurations by very large because there is a large energy difference between the benzene and phenol systems. Therefore,

most of exponential terms will be small, leading to an inaccurate estimate of the free energy difference. The solution is to introduce intermediate states, between the benzene and phenol molecules. Because the free energy is a state function, it does not matter how system 0 is perturbed into system 1. Hence, the Hamiltonian can be written as in Eqn 4.2, and the perturbation is made from $\lambda_i$ to $\lambda_{i+1}$, and the FEP formula is employed as a sum over smaller free energies [27, 120]

$$\Delta G = -RT \sum \ln \left\langle e^{-\beta[U(\lambda_{i+1})-U(\lambda_i)]} \right\rangle_{\lambda_i} \tag{4.14}$$

In passing, it should be mentioned that there exist different strategies of how to perturbe from one $\lambda$-value to another [121], not just from $\lambda_i$ to $\lambda_{i+1}$(forward) or from $\lambda_i$ to $\lambda_{i-1}$ (backward).

### 4.1.3 Bennet's acceptance ratio method

The Bennet's acceptance ratio (BAR) method was first derived by Bennet [122] in the 70's but was not popularized until recently [123]. Bennet showed that the free energy difference between two systems is given by

$$\Delta G = RT \left( \ln \frac{\langle f(-\Delta U + C) \rangle_1}{\langle f(\Delta U - C) \rangle_0} \right) + C \tag{4.15}$$

for any function $f(x)$ satisfying $f(x)/f(-x) = \exp(-x)$ and for any offset $C$. However, to minimize the variance of the free-energy estimate, Bennet showed that $f(x)$ should be the Ferm-function, i.e.,

$$f(x) = 1/(1 + \exp(\beta x)) \tag{4.16}$$

and that $C$ should be

$$C = \Delta G + RT \ln \frac{N_0}{N_1} \tag{4.17}$$

where $N_0$ and $N_1$ are the number of samples from the simulation of state 0 and state 1, respectively. Eqns 4.15 and 4.17 is then solved iteratively, until the free energy converges. It is clear from Eqn 4.15 that BAR treats the forward and backward peturbations equally, and the method can be considered as an optimal weighted average of the two perturbations. As with FEP, BAR may also require that the perturbation is divided into smaller steps [120]. An alternative to summing all the small free energies has been proposed recently [124].

Before we conclude this section, a fourth method to compute free-energy differences should be mentioned, viz., weighted-histogram analysis method (WHAM) [125]. This method also combines informations from several states to compute a free energy and can in fact be shown to be equivalent to BAR for only two states [124]. However, it is not as commonly used as TI, FEP, and BAR for ligand binding.

In paper XII, we presented TI results and in paper XIII, we presented FEP results on the same systems. We found that the FEP results had a much better precision than the TI results. This partly comes from the free-energy estimator used, but also from details in the simulation setup. In paper XIII, we also tested BAR, but the differences between the BAR and FEP results were not statistically significant. This is probably because the test system was well-behaved, i.e., a small difference between the FEP estimates in the forward and backward direction. #

## 4.2 Rigorous methods

Now that a few basic techniques have been introduced, I will proceed by showing how the free energy of binding can be computed in practice. In this section, I will discuss rigorous methods that are based on the techniques described above together with the thermodynamic treatment of ligand-binding in Chapter 1.1. Therefore, such methods are exact from a theoretical point of view. However, because perfect sampling is hard to achieve with a biomolecular system and because a force field is only an approximation to the quantum mechanical potential, the practical results are only estimates.

We can distinguish three kinds of methods to compute the binding free energy rigorously. First, there are methods that try to estimate the standard binding free energy, i.e., the free energy that is measured experimentally. Estimates of such methods are also sometimes called absolute free energies, to contrast them with the second class of methods that only try to compute the relative binding free energies between two ligands. These two classes are collectively called alchemical methods because they relay on the unphysical, yet computational feasible, method of transforming one system into another [126]. A third class of methods, less employed and not used in this thesis, is methods that try to mimic the natural process of binding or unbinding by pulling the ligand to or from the binding site.

### 4.2.1 Standard binding free energies

Let us recall the final expression for the association constant $K_a$ from Chapter 1 [16]

$$K_a = \frac{\int_{site} dq \int dS e^{-\beta U}}{\int_{bulk} dq \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U}} \tag{4.18}$$

The methods in this section compute $K_a$ by introducing a series intermediate states into Eqn 4.18 [16, 127]

$$K_a = \frac{\int_{site} dq \int dS e^{-\beta U}}{Z_n} \times \frac{Z_n}{Z_{n-1}} \times \cdots \\ \frac{Z_2}{Z_1} \times \frac{Z_1}{\int_{bulk} dq \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U}} \tag{4.19}$$

The problem is then reduced to finding the intermediate states, i.e., $Z_1$ through $Z_n$. The first proposed method to compute standard binding free energies was the double-annihilation method [128]. In this method, only a single intermediate state is introduced in Eqn 4.19, a state where the ligand is completely decoupled (annihilated) from the environment. Illustrated by the thermodynamic cycle in Figure 4.1a, the final expression for the free energy (by talking the logarithm of $K_a$) is given by

$$\Delta G_{bind} = \Delta G^{\text{bound}}_{\text{annihi}} - \Delta G^{\text{free}}_{\text{annihi}} = \\ \Delta G^{\text{bound}}_{\text{ele}} + \Delta G^{\text{bound}}_{\text{vdW}} - \Delta G^{\text{free}}_{\text{ele}} - \Delta G^{\text{free}}_{\text{vdW}} \tag{4.20}$$

where, $\Delta G^{\text{bound}}_{\text{annihi}}$ and $\Delta G^{\text{free}}_{\text{annihi}}$ are the free energies of decoupling the ligand when it is bound to the protein and when it is free in solution, respectively. Both these free energies can further be decomposed into an electrostatic, $\Delta G_{\text{ele}}$ and a van der Waals term, $\Delta G_{\text{vdW}}$, which give the free energy when the electrostatic and van der Waals interactions are turned off, respectively. Both of these terms are computed with, e.g., TI or FEP.

When the ligand is decoupled from the environment it is free to diffuse around in the simulation box. This may lead to problems with convergence, because there is more space to sample [129]. Also, it is unclear how the standard concentration can be defined unambiguously, although ad-hoc recipes have been suggested [130].

In the double-decoupling method [13, 129], this is solved by introducing two additional intermediate states in Eqn 4.19, by restraining the ligand in the active site of the protein. This approach is illustrated

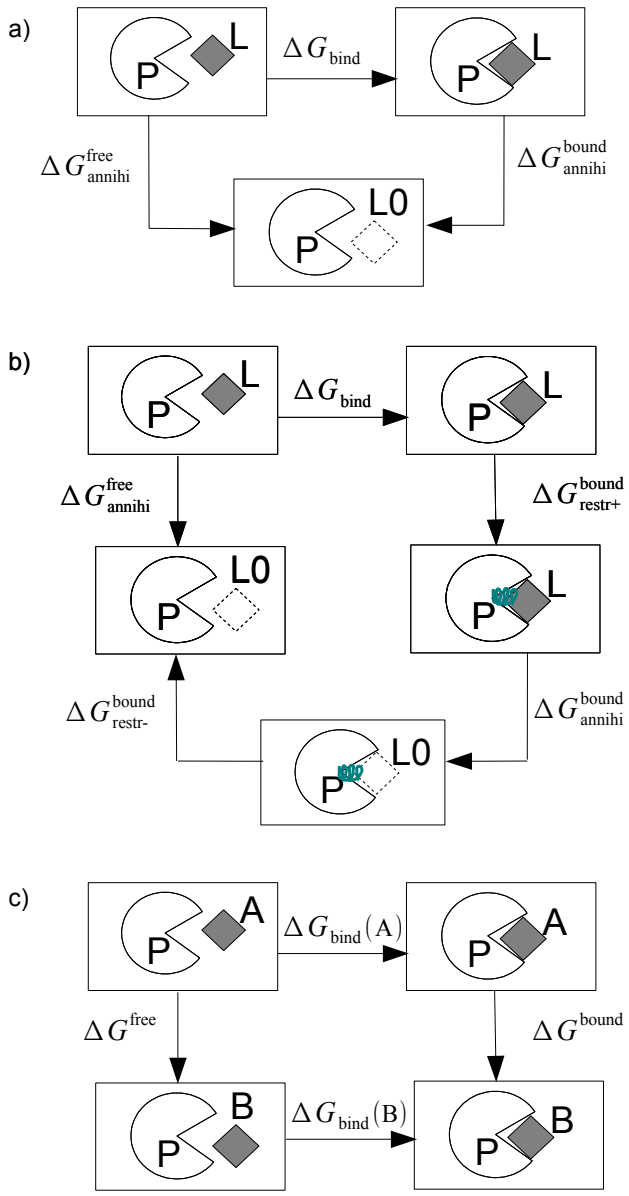


**Figure 4.1:** Thermodynamic cycles to compute binding free energies, a) double-annihilation, b) double-decoupling, and c) relative free energy. L0 describes a non-interacting (decoupled) ligand.

in Figure 4.1b, and the final expression for the free energy is given by [16, 129]

$$\Delta G_{bind} = \Delta G^{\text{bound}}_{\text{restr}+} + \Delta G^{\text{bound}}_{\text{annihi}} + \Delta G^{\text{bound}}_{\text{restr}-} - \Delta G^{\text{free}}_{\text{annihi}} \tag{4.21}$$

- A state in which the ligand is restrained in the binding site has been introduced. The free energy of introducing the restraints is given by $\Delta G^{\text{bound}}_{\text{restr}+}$
- The free energy of decoupling the ligand from its environment, while it is restrained to the binding site is given by $\Delta G^{\text{bound}}_{\text{annihi}}$
- The free energy of removing the restraints of the decoupled ligand is given by $\Delta G^{\text{bound}}_{\text{restr}-}$
- The free energy of decoupling the ligand from its environment when it is free in solution is given by $\Delta G^{\text{free}}_{\text{annihi}}$

$\Delta G^{\text{bound}}_{\text{restr}+}$ can in principle be computed with for instance TI, but can sometimes, in conjunction with $\Delta G^{\text{bound}}_{\text{restr}-}$, be estimated from an approximate analytical formula [131]. $\Delta G^{\text{bound}}_{\text{restr}-}$ can in any case be derived analytically [131]. Several different restraints have been suggested and used [112, 131, 132]. The most basic restraints are rotational and translation restraint but if the ligand is large, it might be advantageous to also restrain the conformational freedom of the ligand.

In paper V and VI, the double-decoupling method was used to estimate the binding affinities of five small ligands. Rather accurate results were obtained, although the precision deteriorated when the ligands became larger. Despite this, the estimates could be used as benchmark for approximate methods. #

### 4.2.2 Relative affinities

The difficulties associated with the methods above can be reduced if its only interesting to determine the difference between the binding free energy of two ligand A and B to the same protein. Employing Eqn 4.18

for $K_a$, we have

$$\begin{aligned}
&\Delta G_{bind}(B) - \Delta G_{bind}(A) = \\
&\quad -RT \left[ \ln \frac{\int_{site} dq^B \int dS e^{-\beta U(B)}}{\int_{bulk} dq^B \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U(B)}} - \right. \\
&\quad \left. \ln \frac{\int_{site} dq^A \int dS e^{-\beta U(A)}}{\int_{bulk} dq^A \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U(A)}} \right] = \\
&\quad -RT \left[ \ln \frac{\int_{site} dq^B \int dS e^{-\beta U(B)}}{\int_{site} dq^A \int dS e^{-\beta U(A)}} - \right. \\
&\quad \left. \ln \frac{\int_{bulk} dq^B \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U(B)}}{\int_{bulk} dq^A \delta(\mathbf{r}_1 - \mathbf{r}^*) \int dS e^{-\beta U(A)}} \right] = \\
&\quad \Delta G_{bound} - \Delta G_{free}
\end{aligned} \tag{4.22}$$

Here, $\Delta G_{bound}$ is the free energy of mutating ligand A to ligand B when they are bound the to protein, and $\Delta G_{free}$ is the corresponding process when the ligands are free in solution. Both these free energies can be calculated with any of the techniques discussed in Chapter 4.1. A thermodynamic cycle (see Figure 4.1c), describing this approach was first suggested by Tembe and McCammon [8].

In papers XII to XV, relative free energies were computed using the method described above, together with either FEP and TI.#

### 4.2.3 Practical considerations

As mentioned before, the approaches described in the two previous sections are collectively called alchemical methods because they transform one system into another. This is possible because the potential is usually described by a force field and as such, the system, can be modified to any extent. It is easy to realize that alchemical methods require sampling of a Hamiltonian that is a mixture of two force fields, and there are two approaches to implement this on a computer. In the single-topology approach, one force field is specified together with parameters for the two end-states. Whenever an atom is disappearing, it is mutated into a dummy atom that does not have any charge and zero Lennard-Jones parameters. In the dual-toplogy approach, two force fields are simultaneously kept in the simulation and mixed appropriately. In such an approach there is no need for dummy atoms [27, 133]. The advantage with

a dual-topology approach is that more complicate differences between ligands can be easily treated at the cost of an increased uncertainty.

Another related problem is which force-field terms that should be perturbed. The electrostatics and van der Waals are obvious, but the treatment of the bonded term is less so. In theory, this should not matter because these parameter are perturbed twice, e.g., once in water and once in the protein, and hence they should cancel out. However, the bonded terms has been treated by different approaches [134]. Putting the bonded terms aside, the question is how to perturb the non-bonded interactions. Perturbing them simultaneously is the obvious answer from an efficiency perspective, but this will cause instabilities unless soft-core potentials are used [54, 55]. Thermodynamically it should not matter whether the electrostatics are perturbed first and then the van der Waals, or vice versa. However, from a technical perspective, the electrostatic perturbation is always performed first, because it is not possible to propagate the motion of a set of atoms with charges but no repulsive term (which determines the size). It is also possible to divide the van der Waals perturbation in two steps [127], although this is seldom done.

We have seen that it is often necessary to introduce several intermediate states in FEP and BAR calculations because the phase space of the two end states do not overlap sufficiently. In TI, such a division comes naturally from the theory. The question is then how many intermediate states that are necessary. As a rule of thumb, for TI the integrand should be smooth, and for FEP, the difference in the important phase space of the two states involved in the perturbations should be small. There have been suggestions how to quantify this overlap [135].

In papers XII and XIII, we investigated how many intermediate states are necessary for an accurate estimate. It turned out that rather few intermediate states were required, between three and five, depending on whether the electrostatic and van der Waals perturbations were attempted simultaneously. #

### 4.2.4 (Un)binding simulations

Although seldom used to estimate binding affinities, the standard binding free energy from ”pulling” simulations can be derived from Eqn 4.19. The intermediate states here are positions of the ligand at different fixed distances from the active site. The binding free energy is thus expressed

as a potential of mean force along the chosen path [127]. However, there is no information that can reveal the correct thermodynamic path a-priori. Unbinding simulations has suggested to be useful for charged ligands or for ligands with a large desolvation energy [129, 136].

## 4.3 Approximate methods

From the previous section, it is clear that the rigorous free energy methods require extensive computations. Simulations of the ligand bound to the protein and free in solution are required, the electrostatic and van der Waals perturbation might have to be computed separately, and the simulations require sampling of several unphysical, intermediate states. To remove the necessity to introduce intermediate states, approximate methods have been developed that require only the simulation of the complex and perhaps the free ligand or the free protein. Such methods are called end-point methods and are described below.

All of these methods are approximate in nature and will not necessarily give the exact result, even with perfect sampling. This is an important point; the hope is that the method will give a good estimate of the binding affinity by error cancellation.

---

In paper XI, all approximate methods described below that are able to compute binding free energies, except PDLD/s-LRA, were used in a blind test. 34 ligands to the trypsin protein and a series of host–guest complexes was used as a test case. Unfortunately, none of the methods gave especially accurate results on the trypsin-case, but rather good results on the host–guest complexes. #

---

### 4.3.1 Linear-response approximation

It is instructive to start with the linear-response approximation (LRA), although it cannot be used on its own to estimate binding affinities because it considers only electrostatic perturbations. We start with two Hamiltonians, $H_0$ in which the ligand interacts fully with its environment, i.e., with both electrostatics and van der Waals interactions turned on, and $H_1$ in which only the van der Waals interactions are turned on. Here, we are only concerned with the potential energy difference between the two systems, $\Delta U$, which is the electrostatic interaction energy between the ligand and the surroundings, here denoted by $E_{\mathrm{ele}}^{\mathrm{L-S}}$. The free

energy difference between these two states can theoretically be calculated with FEP, and expanding the FEP formula in powers of $1/RT$ gives [137]

$$\Delta G_{\text{ele}} = \langle \Delta U \rangle_0 - \frac{1}{2RT} \left\langle (\Delta U - \langle \Delta U \rangle_0)^2 \right\rangle_0 + \frac{1}{6(RT)^2} \left\langle (\Delta U - \langle \Delta U \rangle_0)^3 \right\rangle_0 - \dots \tag{4.23}$$

and reversing the process gives

$$\Delta G_{\text{ele}} = \langle \Delta U \rangle_1 - \frac{1}{2RT} \left\langle (\Delta U - \langle \Delta U \rangle_1)^2 \right\rangle_1 - \frac{1}{6(RT)^2} \left\langle (\Delta U - \langle \Delta U \rangle_1)^3 \right\rangle_1 - \dots \tag{4.24}$$

By truncating after the first term and taking the average of Eqn 4.23 and 4.24 we arrive at the LRA formula [27, 137]

$$\Delta G_{\text{ele}} = \frac{1}{2} \left[ \langle \Delta U \rangle_0 + \langle \Delta U \rangle_1 \right] = \frac{1}{2} \left[ \langle E_{\text{ele}}^{\text{L-S}} \rangle_0 + \langle E_{\text{ele}}^{\text{L-S}} \rangle_1 \right] \tag{4.25}$$

### 4.3.2 Linear interaction energy

The linear interaction energy (LIE) method [138, 139] takes Eqn 4.25 one step further and assumes that $E_{\text{ele}}^{\text{L-S}}$ is zero when sampling at state 1, i.e, when there is no electrostatic interactions between the ligand and the surroundings. However, LIE allows deviation from this assumption by letting 1/2 be a parameter that depends on the nature of the ligand [137]. Furthermore, the LRA treatment must be supplemented by an estimate of the non-polar free energy and LIE assumes that this can be described by scaling the van der Waals interactions with the surroundings.

The LIE approach for ligand-binding can be derived form the same arguments as in the double-annihilation approach (see Eqn 4.20) or from the cycle in Figure 4.2. Hence, the LIE formula for ligand-binding is [139]

$$\Delta G = \alpha \left( \langle E_{\text{ele}}^{\text{L-S}} \rangle_{\text{PL}} - \langle E_{\text{ele}}^{\text{L-S}} \rangle_{\text{L}} \right) + \beta \left( \langle E_{\text{vdW}}^{\text{L-S}} \rangle_{\text{PL}} - \langle E_{\text{vdW}}^{\text{L-S}} \rangle_{\text{L}} \right) \tag{4.26}$$

where $E^{\text{L-S}}$ is the interaction energy between the ligand and the surroundings (either protein or solvent), $\alpha$ and $\beta$ are two parameters, and the subscript of the ensemble averages indicates from which simulation they are calculated.

In paper VIII, the efficiency of the LIE method was investigated and compared to the efficiency of the MM/PBSA method (see below). It was concluded that LIE is somewhat more efficient than MM/PBSA, although the LIE method requires some special setup of the system. #

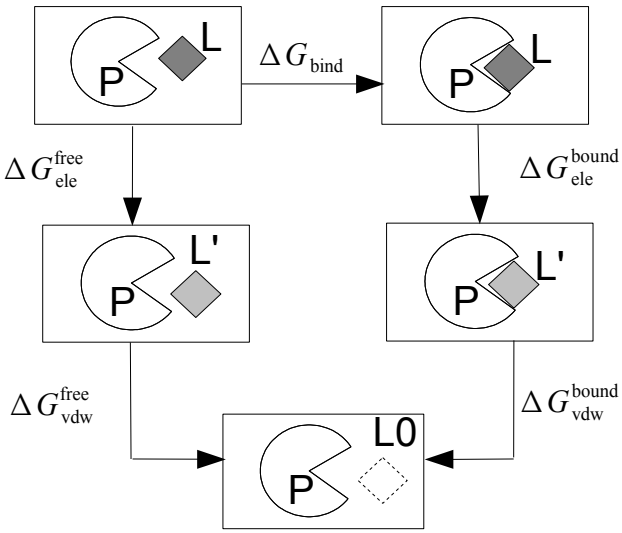


**Figure 4.2:** Thermodynamic cycle to compute binding free energies with LIE. L' describes a ligand with zeroed charges, and L0 describes a non-interacting (decoupled) ligand.

$\alpha$ is a parameter that scales the electrostatic interactions and depends on the nature of the ligand. In the most common parametrisation [137], $\alpha$ depends on the number of hydroxyl groups in the ligand and its charge. If the ligand is charged, $\alpha = 0.5$, if the ligand is neutral and has no hydroxyl groups, $\alpha = 0.43$, and if it is neutral and has one or more than one hydroxyl groups, $\alpha = 0.37$ or 0.33, respectively. An attempt to include other groups has been published recently [140]. The other parameter, $\beta$, which scales the van der Waals interactions, is truly empirical. In the work from the Åqvist lab, it is most commonly set to 0.18 [141, 142], but many groups treat it as a fitting parameter [142]. There have been attempts to derive an expression for the optimal parameter [143]. In passing, it should be noted that in the literature, $\alpha$ is sometimes used as the parameter that scales the van der Waals interactions, and $\beta$ is therefore used as the parameters that scales the electrostatics. A third parameter, $\gamma$, has also been suggested that depends on some kind of molecular surface of the ligand or the protein [144]. However this will only affects the absolute estimates, and it is unclear if it improves the results [145]. Additional terms such as hydrogen bonding terms have also been added to Eqn 4.25 [146], but then the method becomes more of a statistical method with terms partially derived from simulations.

### 4.3.3 PDLD/s-LRA

Similarly to LIE and the double-annihilation methods, the LRA expression for ligand binding is obtained by taking the free energy of the bound state and subtracting the free energy of the free state, i.e.,

$$\Delta G_{\text{ele}} = \frac{1}{2}\left(\left\langle \Delta U^{\text{bound}} \right\rangle_{\text{PL}} + \left\langle \Delta U^{\text{bound}} \right\rangle_{\text{PL}'} - \left\langle \Delta U^{\text{free}} \right\rangle_{\text{L}} - \left\langle \Delta U^{\text{free}} \right\rangle_{\text{L}'}\right) \tag{4.27}$$

Although the LRA method could be used to compute the electrostatic part of the binding free energy, it suffers from the fact that interaction energies with explicitly modelled solvent can be hard to converge [147]. Therefore, the semi-macroscopic protein-dipole–Langevin-dipole method was introduced in the LRA framework, giving the PDLD/s-LRA method. This approach tries to obtain more stable energies by using a simplified water model and scaling the electrostatic interactions. Therefore, this approach does not sample, $\Delta U^{\text{bound}} = \Delta U^{\text{free}} = E_{\text{ele}}^{\text{L-S}}$, but instead samples effective potentials derived from the thermodynamic cycles in Figure 4.3, viz.,

$$\Delta U^{\text{bound}} = \left(G_{\text{pol}}^{\text{PL}} - G_{\text{pol}}^{\text{PL}'}\right)\left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}}\right) + G_{\text{pol}}^{\text{L}}\left(1 - \frac{1}{\epsilon_{\text{int}}}\right) + \frac{E_{\text{ele}}^{\text{L}}}{\epsilon_{\text{int}}} + \frac{E_{\text{ele}}^{\text{L-P}}}{\epsilon_{\text{int}}} \tag{4.28}$$

and

$$\Delta U^{\text{free}} = G_{\text{pol}}^{\text{L}}\left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}}\right) + G_{\text{pol}}^{\text{L}}\left(1 - \frac{1}{\epsilon_{\text{int}}}\right) + \frac{E_{\text{ele}}^{\text{L}}}{\epsilon_{\text{int}}} \tag{4.29}$$

In addition to simulations of the complex and the free ligand, this method requires also simulation of the complex and the free ligand with the ligand charges zeroed (PL′ and L′). Hence, this method does not assume that these energies are negligible as in the LIE method. $G_{\text{pol}}$ is the polar solvation energy, $E_{\text{ele}}^{\text{L}}$ and $E_{\text{ele}}^{\text{L-P}}$ are the intramolecular and intermolecular electrostatic energies, respectively, $\epsilon_{\text{int}}$ is an effective dielectric constant of the solute, and $\epsilon_{\text{ext}}$ is the dielectric constant of the solvent.

Naturally, this method needs to be combined with an estimate of the non-polar free energy. A common strategy is to borrow it from the LIE method, giving the PDLD/s-LRA/$\beta$ method [149, 150].

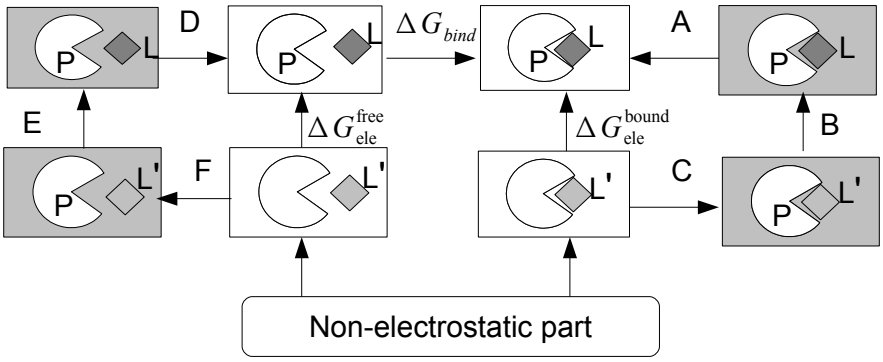


**Figure 4.3:** Thermodynamic cycle used to derive the PDLD/s-LRA terms. The binding free energy is calculated by the central cycle, which is broken down into electrostatic and non-electrostatic contributions, $\Delta G_{bind} = \Delta G_{\text{ele}}^{\text{bound}} - \Delta G_{\text{ele}}^{\text{free}} + \Delta G_{\text{non-ele}}$. The non-electrostatic term is calculated by other approaches, see text, whereas the electrostatic terms are further broken down into the two outer paths. $\Delta G_{\text{ele}}^{\text{bound}}$ is described by the right, outer cycle: $\Delta G(A) = G_{\text{pol}}^{\text{PL}} \left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}}\right)$, $\Delta G(B) = G_{\text{pol}}^{\text{L}} \left(1 - \frac{1}{\epsilon_{\text{int}}}\right) + \frac{E_{\text{ele}}^{\text{L}}}{\epsilon_{\text{int}}} + \frac{E_{\text{ele}}^{\text{L-P}}}{\epsilon_{\text{int}}}$, and $\Delta G(C) = G_{\text{pol}}^{\text{PL}'} \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}}\right)$. $\Delta G_{\text{ele}}^{\text{free}}$ is calculated by the left, outer cycle: $\Delta G(D) = (G_{\text{pol}}^{\text{P}} + G_{\text{pol}}^{\text{L}}) \left(\frac{1}{\epsilon_{\text{int}}} - \frac{1}{\epsilon_{\text{ext}}}\right)$, $\Delta G(E) = G_{\text{pol}}^{\text{L}} \left(1 - \frac{1}{\epsilon_{\text{int}}}\right) + \frac{E_{\text{ele}}^{\text{L}}}{\epsilon_{\text{int}}}$, and $\Delta G(F) = G_{\text{pol}}^{\text{P}} \left(\frac{1}{\epsilon_{\text{ext}}} - \frac{1}{\epsilon_{\text{int}}}\right)$. Eqn 4.28 is the sum of the free energies A, B, and C and Eqn 4.29 is the sum of the free energies D, E, and F. L is the ligand with full charge, whereas the ligand charges have been zeroed in L'. The boxes with white and shaded backgrounds indicate an enviroment with the solvent dielectric constant equal to $\epsilon_{\text{ext}}$ and $\epsilon_{\text{int}}$, respectively.

In paper IX, the PDLD method was replaced with an Amber force field and Possion–Boltzmann or generalised Born solvation, forming the MMPB/s-LRA and MMGB/s-LRA methods. These methods were then compared to the MM/PBSA method. #

### 4.3.4 MM/PBSA

Molecular mechanics with Possion–Boltzmann and surface-area solvation (MM/PBSA) [11, 151, 152] is a method that also treats the polar part of the binding free energy with continuum electrostatics. In this method, the free energy of a state, either the protein–ligand complex, the free protein, or the free ligand, is calculated from [11]

$$G = E_{\text{int}} + E_{\text{ele}} + E_{\text{vdW}} + G_{\text{pol}} + G_{\text{np}} - TS \tag{4.30}$$

where the three first terms are the molecular mechanics energies, viz., the internal (bonds, angles, torsions), electrostatic, and van der Waals energies, $G_{\text{pol}}$ and $G_{\text{np}}$ are the polar and non-polar solvation free energy, and the last term is the absolute temperature times an entropy estimate. The polar solvation energy could in principle be calculated by any continuum method and a common choice, apart from PB, is the generalised Born (GB) method, giving the MM/GBSA method. In this thesis, I will use MM/PBSA as the common name for this method. Furthermore, the MM energies could in principle be calculated by QM methods instead.

Sometimes it is useful to separate the MM/PBSA energy terms into a polar and a non-polar part. Whenever this is the case, the polar part is defined as the sum of the $E_{\text{ele}}$ and $G_{\text{pol}}$ term, and the non-polar part is defined as the sum of the other four terms.

To compute the free energy of binding, an ensemble average of the free energy of the complex is subtracted from the analogous averages for the free protein and free ligand

$$\Delta G = \langle G(\text{PL}) \rangle - \langle G(\text{P}) \rangle - \langle G(\text{L}) \rangle \tag{4.31}$$

Each of these averages should rigorously be calculated from three different simulations, one for each species, an approach that I will call the three-average (3A-MM/PBSA) approach. However, it is much more common to simulate only the complex and then obtain the other species by removing atoms [142]. Such an approach will be called the one-average (1A-MM/PBSA) approach. The rationale is that such an approach will

reduce the noise of the estimates. In addition, the $E_{\text{int}}$ term will cancel exactly.

The theoretical foundation of the MM/PBSA approach has been criticised, although attempts have been made to connect it to thermodynamic theories [15, 153]. One of the most criticised terms is the entropic term [150], which is usually computed from normal-mode frequencies [11].

---

In papers I through VII, the MM/PBSA method was thoroughly examined, and all of the terms in Eqn 4.29 were tested separately. In paper IX, the 3A-MM/PBSA was compared to the 1A-MM/PBSA method. Furthermore, in paper XII, the MM/PBSA results were compared to thermodynamic integration calculations. All these tests are more thoroughly described in Chapter 5. #

---

### 4.3.5 Continuum linear interaction energy

It is possible to replace the explicit ligand–water interactions in the LIE method with estimates from continuum electrostatics. Such methods have been presented several times before, but to my knowledge only for the polar part [154, 155, 156]. Here, I will show how the non-polar part can also be replaced with a continuum treatment. This method will be denoted continuum LIE (CLIE). This will be useful when we compare the different approximate methods in the next section.

---

In papers XI, we used a method that we called continuum LIE. However, this uses a continuum description only for the electrostatic interactions with the solvent. The van der Waals interactions with the solvent were treated with explicit water. #

---

We start with the electrostatics part of Eqn 4.26, $\alpha(\langle E_{\text{ele}}^{\text{L-S}}\rangle_{\text{PL}} - \langle E_{\text{ele}}^{\text{L-S}}\rangle_{\text{L}})$. The electrostatic interactions between the free ligand and the surrounding water molecules are simply $\langle E_{\text{ele}}^{\text{L-S}}\rangle_{\text{L}} = 2G_{\text{pol}}^{\text{L}}$, because the continuum solvation methods are also based on linear response theory [156]. However, the electrostatic interactions between the bound ligand and the surrounding water molecules are slightly more tricky to obtain. The $\langle E_{\text{ele}}^{\text{L-S}}\rangle_{\text{PL}}$ term represents the free energy of turning off the ligand charges when it is bound to the protein. This free energy can be shown to be a sum of three terms [156]; the electrostatic interaction between the ligand and the

protein, $E_{\rm ele}^{\rm L-P}$, the polar solvation energy of the complex, $2G_{\rm pol}^{\rm PL}$, and the polar solvation energy of the complex when the ligand charges is zeroed, $2G_{\rm pol}^{\rm PL'}$. (Here we assume that the internal energy of the ligand cancels). Hence, $\langle E_{\rm ele}^{\rm L-S}\rangle_{\rm PL}$ can be replaced with $\left\langle 2\left(G_{\rm pol}^{\rm PL}-G_{\rm pol}^{\rm PL'}\right)+E_{\rm ele}^{\rm L-P}\right\rangle_{\rm PL}$.

It should be mentioned that an alternative form of the electrostatic term has been used in CLIE [154]. To understand this, we write the total continuum solvation free energy of the complex within the PB formalism

$$G_{\rm pol}^{\rm PL} = \frac{1}{2}\sum_i q_i\phi(\mathbf{r}_i) = \frac{1}{2}\sum_p q_p\phi(\mathbf{r}_p) + \frac{1}{2}\sum_l q_l\phi(\mathbf{r}_l) = \frac{1}{2}\sum_p q_p(\phi^{\rm P}(\mathbf{r}_p)+\phi^{\rm L}(\mathbf{r}_p)) + \frac{1}{2}\sum_l q_l(\phi^{\rm P}(\mathbf{r}_l)+\phi^{\rm L}(\mathbf{r}_l)) \tag{4.32}$$

where $\phi(\mathbf{r})$ is the reaction field (RF) of both the protein and the ligand, which can be divided into a RF of the protein, $\phi^{\rm P}(\mathbf{r})$, and a RF of the ligand, $\phi^{\rm L}(\mathbf{r})$. The sum over $i$ takes into account all the charges in the system, which can be divided into ligand charges denoted by $l$ and protein charges denoted by $p$. Likewise, $G_{\rm pol}^{\rm PL'} = 1/2\sum_p q_p\phi^{\rm P}(\mathbf{r}_p)$ and therefore,

$$G_{\rm pol}^{\rm PL} - G_{\rm pol}^{\rm PL'} = \frac{1}{2}\sum_p q_p\phi^{\rm L}(\mathbf{r}_p) + \frac{1}{2}\sum_l q_l(\phi^{\rm P}(\mathbf{r}_l)+\phi^{\rm L}(\mathbf{r}_l)) \tag{4.33}$$

whereas the alternative [154] is only to evaluate the last two terms of Eqn 4.32, which we will denote by $G_{\rm pol}^{\rm L(P)}$, and therefore ignore the interaction between the protein charges and the part of the RF created by the ligand. The motivation is that this represents the interaction between the ligand and the total RF and therefore the interaction between the ligand and the solvent. To obtain $G_{\rm pol}^{\rm L(P)}$, we need to evaluate Eqn 4.32 but also $G_{\rm pol}^{\rm PL'}$ and $G_{\rm pol}^{\rm P'L} = 1/2\sum_l q_l\phi^{\rm L}(\mathbf{r}_l)$ . $G_{\rm pol}^{\rm L(P)}$ is then obtained by combining the different calculations, viz., $G_{\rm pol}^{\rm L(P)} = \frac{1}{2}\left(G_{\rm pol}^{\rm PL} - G_{\rm pol}^{\rm PL'} + G_{\rm pol}^{\rm P'L}\right)$. Herein, I will only use the expression in Eqn 4.33.

Next, we consider the non-polar part of Eqn 4.26, $\beta(\langle E_{\rm vdW}^{\rm L-S}\rangle_{\rm PL} - \langle E_{\rm vdW}^{\rm L-S}\rangle_{\rm L})$. The van der Waals interactions can be replaced by the corresponding continuum estimates. As described in Chapter 2.3.2, this is simply a integral over solute surface segments. For the free ligand, Eqn 2.20 or 2.21 can be applied directly, but again, for the bound ligand, the situation is a little bit more complicated. However it is easy to realize that the sum in Eqn 2.20 can be divided into a sum over all ligand atoms

and a sum over all protein atoms, hence

$$E_{\mathrm{cvdW}}^{\mathrm{PL}} = \sum_l \rho \int U_{\mathrm{LJ}}(\mathbf{r}) d\mathbf{r} + \sum_p \rho \int U_{\mathrm{LJ}}(\mathbf{r}) d\mathbf{r} = E_{\mathrm{cvdW}}^{\mathrm{L(P)}} + E_{\mathrm{cvdW}}^{\mathrm{P0}} \tag{4.34}$$

where cvdW is used as an abbreviation for continuum van der Waals, and the term we are interested in is $E_{\mathrm{cvdW}}^{\mathrm{L(P)}}$ because this is exactly the interaction between the ligand and the solvent.

Now, we can put everything together and the continuum LIE method can be written as

$$\begin{aligned}\Delta G = \alpha \left[ \left\langle E_{\mathrm{ele}}^{\mathrm{L-P}} \right\rangle_{\mathrm{PL}} + 2 \left\langle G_{\mathrm{pol}}^{\mathrm{PL}} - G_{\mathrm{pol}}^{\mathrm{PL'}} \right\rangle_{\mathrm{PL}} - 2 \left\langle G_{\mathrm{pol}}^{\mathrm{L}} \right\rangle_{\mathrm{L}} \right] + \\ \beta \left[ \left\langle E_{\mathrm{vdW}}^{\mathrm{L-P}} \right\rangle_{\mathrm{PL}} + \left\langle E_{\mathrm{cvdW}}^{\mathrm{PL}} - E_{\mathrm{cvdW}}^{\mathrm{P0}} \right\rangle_{\mathrm{PL}} - \left\langle E_{\mathrm{cvdW}}^{\mathrm{L}} \right\rangle_{\mathrm{L}} \right]\end{aligned} \tag{4.35}$$

### 4.3.6 Comparison of approximate polar methods

It is instructive to compare the various approximate approaches to estimate binding affinities. Although the main equations look rather different, all methods have some things in common. In this section, I will compare the polar parts of 1A-MM/PBSA, PDLS/s-LRA, and continuum LIE.

I start the comparison by setting the solute dielectric constant in the PDLD/s-LRA method to unity, because that is the constant usually adopted in MM/PBSA calculations. However, it is easy to use a different constant in both MM/PBSA and CLIE. It should be noted that the internal dielectric constant is less of a physical constant than a parameter of the method [147, 157]. Assuming an internal dielectric constant of unity, PDLD/s-LRA reduces to

$$\begin{aligned}\Delta G &= \frac{1}{2}\zeta \left( \langle G_{\mathrm{pol}}^{\mathrm{PL}} - G_{\mathrm{pol}}^{\mathrm{PL'}} \rangle_{\mathrm{PL}} + \langle G_{\mathrm{pol}}^{\mathrm{PL}} - G_{\mathrm{pol}}^{\mathrm{PL'}} \rangle_{\mathrm{PL'}} \right) \\ &+ \frac{1}{2} \left( \langle E_{\mathrm{ele}}^{\mathrm{L}} \rangle_{\mathrm{PL}} + \langle E_{\mathrm{ele}}^{\mathrm{L}} \rangle_{\mathrm{PL'}} \right) \\ &+ \frac{1}{2} \left( \langle E_{\mathrm{ele}}^{\mathrm{L-P}} \rangle_{\mathrm{PL}} + \langle E_{\mathrm{ele}}^{\mathrm{L-P}} \rangle_{\mathrm{PL'}} \right) \\ &- \frac{1}{2}\zeta \left( \langle G_{\mathrm{pol}}^{\mathrm{L}} \rangle_{\mathrm{L}} + \langle G_{\mathrm{pol}}^{\mathrm{L}} \rangle_{\mathrm{L'}} \right) \\ &- \frac{1}{2} \left( \langle E_{\mathrm{ele}}^{\mathrm{L}} \rangle_{\mathrm{L}} + \langle E_{\mathrm{ele}}^{\mathrm{L}} \rangle_{\mathrm{L'}} \right)\end{aligned} \tag{4.36}$$

**Table 4.1:** Comparison of approximate methods to compute polar binding free energy

| Interaction | 1A-MM/PBSA | PDLD/s-LRA | CLIE |
|---|---|---|---|
| $E^{\mathrm{L-P}}$ | $\left\langle E^{\mathrm{L-P}}_{\mathrm{ele}} \right\rangle_{\mathrm{PL}}$ | $\frac{1}{2}\left(\left\langle E^{\mathrm{L-P}}_{\mathrm{ele}} \right\rangle_{\mathrm{PL}} + \left\langle E^{\mathrm{L-P}}_{\mathrm{ele}} \right\rangle_{\mathrm{PL'}}\right)$ | $\alpha \left\langle E^{\mathrm{L-P}}_{\mathrm{ele}} \right\rangle_{\mathrm{PL}}$ |
| $E^{\mathrm{L}}$ | | $\frac{1}{2}\left(\langle E^{\mathrm{L}}_{\mathrm{ele}} \rangle_{\mathrm{PL}} + \langle E^{\mathrm{L}}_{\mathrm{ele}} \rangle_{\mathrm{PL'}} - \langle E^{\mathrm{L}}_{\mathrm{ele}} \rangle_{\mathrm{L}} - \langle E^{\mathrm{L}}_{\mathrm{ele}} \rangle_{\mathrm{L'}}\right)$ | |
| Protein solvation | $\left\langle G^{\mathrm{PL}}_{\mathrm{pol}} - G^{\mathrm{P}}_{\mathrm{pol}} \right\rangle_{\mathrm{PL}}$ | $\frac{1}{2}\zeta\left(\left\langle G^{\mathrm{PL}}_{\mathrm{pol}} - G^{\mathrm{PL'}}_{\mathrm{pol}} \right\rangle_{\mathrm{PL}} + \left\langle G^{\mathrm{PL}}_{\mathrm{pol}} - G^{\mathrm{PL'}}_{\mathrm{pol}} \right\rangle_{\mathrm{PL'}}\right)$ | $2\ \alpha \left\langle G^{\mathrm{PL}}_{\mathrm{pol}} - G^{\mathrm{PL'}}_{\mathrm{pol}} \right\rangle_{\mathrm{PL}}$ |
| Ligand solvation | $\left\langle G^{\mathrm{L}}_{\mathrm{pol}} \right\rangle_{\mathrm{PL}}$ | $\frac{1}{2}\zeta\left(\left\langle G^{\mathrm{L}}_{\mathrm{pol}} \right\rangle_{\mathrm{L}} + \left\langle G^{\mathrm{L}}_{\mathrm{pol}} \right\rangle_{\mathrm{L'}}\right)$ | $2\alpha\left(\left\langle G^{\mathrm{L}}_{\mathrm{pol}} \right\rangle_{\mathrm{L}}\right)$ |

where $\zeta = (1 - 1/\epsilon_{\text{ext}}) = 0.99$ (in water).

In paper IX, we tested the influence of the internal dielectric constant in the MM/PBSA and PDLD/s-LRA methods. Depending on the system and the method, an increased dielectric constant could be advantageous or disadvantageous. #

After inserting Eqn 4.30 into Eqn 4.31, the polar part of the 1A-MM/PBSA approach is

$$\Delta G = \langle E_{\text{ele}}^{\text{L-P}} \rangle_{\text{PL}} + \langle G_{\text{pol}}^{\text{PL}} - G_{\text{pol}}^{\text{P}} \rangle_{\text{PL}} - \langle G_{\text{pol}}^{\text{L}} \rangle_{\text{PL}} \tag{4.37}$$

By identifying four different types of interactions in the equations, protein–ligand interactions, ligand intramolecular electrostatic energy, protein solvation, and ligand solvation, the three methods are compared in Table 4.1. The protein–ligand interactions are treated similarly in 1A-MM/PBSA and CLIE, although CLIE scales the interaction. PDLD/s-LRA on the other hand average the interaction over two ensemble averages, and the final energy would be somewhere between the 1A-MM/PBSA and CLIE energy. PDLD/s-LRA is also the only one that treats changes in intramolecular energy, although this could be achieved with the 3A-MM/PBSA methods. The protein solvation term is treated rather differently in the three methods. MM/PBSA considers an unbound cavity that is void of ligand and therefore filled with continuum solvent, whereas PDLD/s-LRA and CLIE consider an unbound cavity that is filled with a non-interaction ligand. In addition, PDLD/s-LRA takes contribution from two ensembles, and both this method and CLIE scales the solvation, although it is clear that the effective scaling is close to unity. The ligand solvation is also treated rather differently. 1A-MM/PBSA takes this contribution from the complex simulation, although this could be improved by using the 3A-MM/PBSA method. PDLD/s-LRA averages over two ensembles and also scales the solvation. CLIE uses in principle the pure solvation free energy of the ligand. In conclusions it seems that the one of the largest difference is that MM/PBSA is the only method that does not scale the energies, most clearly seen for the protein–ligand interaction energy. Recently is was suggested that scaling could improve the results [150] and a test has been made in that direction [158].

**Table 4.2:** Comparison of approximate methods to compute non-polar binding free energy

| Interaction | 1A-MM/PBSA | CLIE |
|---|---|---|
| $E^{\mathrm{L-P}}$ | $\left\langle E^{\mathrm{L-P}}_{\mathrm{vdW}} \right\rangle_{\mathrm{PL}}$ | $\beta \left\langle E^{\mathrm{L-P}}_{\mathrm{vdW}} \right\rangle_{\mathrm{PL}}$ |
| Protein non-polar solvation | $\left\langle E^{\mathrm{PL}}_{\mathrm{cvdW}} - E^{\mathrm{P}}_{\mathrm{cvdW}} \right\rangle_{\mathrm{PL}}$ | $\beta \left\langle E^{\mathrm{PL}}_{\mathrm{cvdW}} - E^{\mathrm{P0}}_{\mathrm{cvdW}} \right\rangle_{\mathrm{PL}}$ |
| Ligand non-polar solvation | $-\left\langle E^{\mathrm{L}}_{\mathrm{cvdW}} \right\rangle_{\mathrm{PL}}$ | $-\beta \left\langle E^{\mathrm{L}}_{\mathrm{cvdW}} \right\rangle_{\mathrm{L}}$ |
| Cavitation energy | $\langle \Delta\Delta G_{\mathrm{cav}} \rangle_{\mathrm{PL}}$ | |
| Entropy | $\langle \Delta TS \rangle_{\mathrm{PL}}$ | |

### 4.3.7 Comparison of approximate non-polar methods

In this section, I will compare the non-polar terms of 1A-MM/PBSA and CLIE. I start the comparison by choosing the PCM method for the non-polar solvation free energy (see chapter 2.3.2) and expand the 1A-MM/PBSA approach, which gives

$$\Delta G = \langle E^{\mathrm{L-P}}_{\mathrm{vdW}} + E^{\mathrm{PL}}_{\mathrm{cvdW}} - E^{\mathrm{P}}_{\mathrm{cvdW}} - E^{\mathrm{L}}_{\mathrm{cvdW}} \rangle_{\mathrm{PL}} + \langle \Delta\Delta G_{\mathrm{cav}} - T\Delta S \rangle_{\mathrm{PL}} \tag{4.38}$$

Similar to the comparison of polar methods, I here introduce a number of interactions, protein–ligand van der Waals interactions, protein non-polar solvation, ligand non-polar solvation, cavitation energy, and entropy. The comparison is made in Table 4.2

The interaction energy between the protein and the ligand is identical, but CLIE scales this term (and usually rather extensively). The protein non-polar solvation differ in which unbound cavity the methods assume: 1A-MM/PBSA assumes that the unbound cavity is completely void of ligand and therefore filled with solvent, whereas the CLIE method assumes that it is filled with a non-interacting ligand. 1A-MM/PBSA calculates the ligand non-polar solvation from a complex ensemble, although this could be improved with the 3A-MM/PBSA. Again, CLIE scales this term. Furthermore, CLIE lacks an explicit cavitation term, although such

a term has been suggested. CLIE also lacks an entropy term, but it has been argued that it is implicitly in the $\beta$ parameter [159]. In the standard MM/PBSA approach, a SASA term is used that should in principle model protein and ligand solvation, as well as cavitation energy.

From this and the preceding section, it is clear that approximate continuum solvent-based methods differ mostly in the non-polar treatment, whereas the polar treatment is rather similar.

# 5 Summary of thesis work

In this chapter I will shortly summarize the most important findings of this thesis. As the papers are only a few pages away, there is no need to repeat the abstracts but I rather try to provide a quick overview of what have been done. I have followed a logical rather than a chronological order to make a more clear connection between the papers. Furthermore, I have divided the fifteen papers into two sections. First, I will discuss the evaluation of the MM/PBSA method and the comparison with the LIE and PDLD/s-LRA/$\beta$ methods. Thereafter, I will discuss three studies in which rigorous methods rather than approximate methods have been used as the main tool.

## 5.1 Approximate methods

Most of the papers in this thesis are dedicated to the evaluation of the MM/PB(GB)SA method. In paper I, we start with investigating how to obtain MM/GBSA estimates with a statistical precision of 1 kJ/mol. Typically, MM/PBSA gives poor precision and thus could not be compared to other methods, although this is a topic rarely discussed in the literature. We tested if it is best to run a single long simulation or if it is advantageous to run several short simulations. As a test case, the tetrameric protein avidin was used. Because the four binding sites are equivalent, a reliable method should give four identical estimates within the statistical uncertainty. This could not be obtained by running a single long simulation because the estimated standard error was too low. However, by running several simulations and average the results, we could

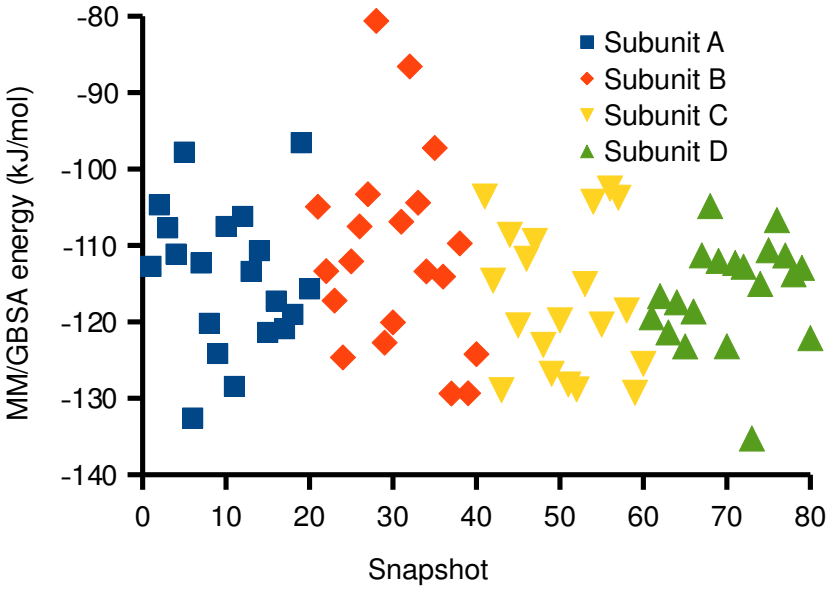


**Figure 5.1:** MM/GBSA results for 20 simulations and 4 subunits of the avidin–biotin complex.

obtain converged results. As shown in Figure 5.1, the estimates from the different simulations could differ by up to 70 kJ/mol, illustrating the imprecise nature of the method. By averaging over 20 to 25 simulations, the goal of a precision of 1 kJ/mol, could be reached.

In the second study, paper II, we investigated several methods to obtain statistically independent simulations. Four different approaches were tested

1. Velocity-induced independent-trajectories (VIIT)
2. Solvent-induced independent-trajectories (SIIT)
3. Conformation, rotation, and protonation-induced independent-trajectories (CRPIIT)
4. Alternative conformation-induced independent-trajectories (ACIIT)

These approaches were introduced in Chapter 3.3.2 and all of them take advantage of some ambiguities when setting up a biolmolecular system for simulations. We found that SIIT and ACIIT could improve the sampling,

compared to the simpler and more common VIIT approach. Therefore, it is recommended to always use at least SIIT. (ACIIT can naturally only be used when the crystal structure contains alternative conformations.) CRPIIT could also be used to improve the sampling, but care must be taken so that the chemistry of the active site do not change.

In the following five papers, we evaluated each of the terms in the MM/PBSA method, viz., the energy model, polar and non-polar solvation, and entropy. In paper III, we replaced the MM energies (electrostatics and van der Waals) with a semi-empirical QM method and the polar solvation method with COSMO. Three different Hamiltonians were tested, viz., AM1, RM1, and PM6. These Hamiltonians were supplemented with empirical corrections for hydrogen-bonds and dispersion interactions. It was found that all of the Hamiltonians perform more or less equally well compared to experiments, although AM1 was on average the best method. However, it was clear that a dispersion correction was required to obtain any correlation with experiment. The benefit of the hydrogen-bond correction was not clear but it can be applied by default because it is very cheap. However, even with the best method and both corrections, the results were not significantly better than with the regular MM/PBSA method.

In paper IV, we introduced the 3D-RISM method into the MM/PBSA framework, by replacing the polar and non-polar terms. This also gave us a chance to compare this method to the usual GB and PB methods. In total, we tested two variants of 3D-RISM, two implementations of PB, and four different GB methods. The binding-affinity estimates when using each of these methods are shown in Figure 5.2. It is clear that many of the methods gave similar relative affinities, although there are differences of about 40 kJ/mol. However, the absolute affinities varied significantly more depending on the solvation method with differences up to 200 kJ/mol. Therefore, it is meaningless to discuss absolute MM/PBSA results. Also, we saw no advantage of using 3D-RISM even though it is more rigorous.

In papers V and VI, we evaluated three continuum estimates of the non-polar solvation free energy by comparing them to rigorous thermodynamic integration calculations. Three different continuum estimates were considered, viz., SASA, CD, and PCM (see Chapter 2.3.3). In paper V, we studied the binding of benzene to the engineered cavity of T4-lysozyme. This cavity is void of any water molecules in the unbound state, something that the continuum methods cannot handle because they fill all unoccupied space with continuum water.

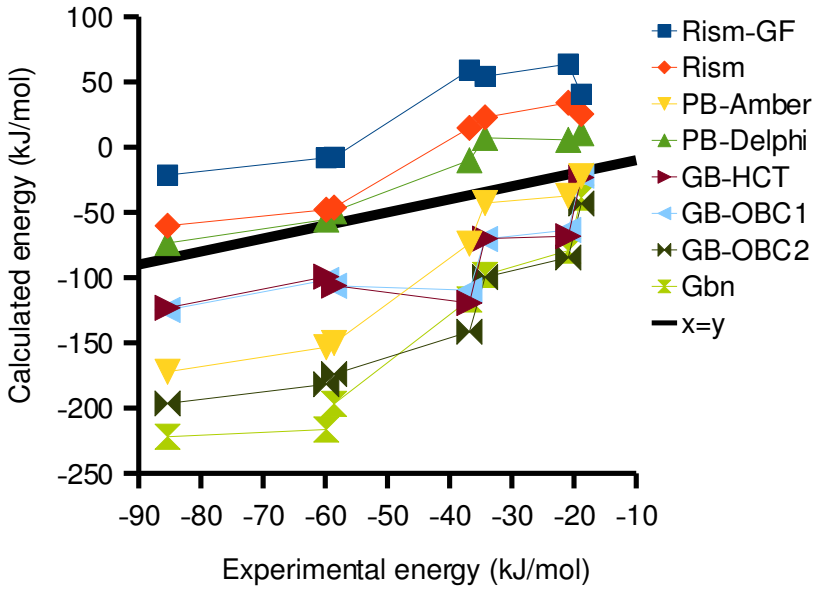


**Figure 5.2:** MM/GBSA results for seven avidin ligands using various polar solvation methods.

However, this behaviour could be corrected for by introducing a non-interacting ligand in the unbound state. In paper VI, this study was extended to four other protein systems with a varying degree of solvation in the unbound protein. It was found that neither of the continuum method could give an accurate estimate on the wide range of protein systems as illustrated in Figure 5.3. This is because they neglect the microscopic structure of water molecules, the interaction energy between the active site water molecules and the protein, and the entropy change when the water molecules are expelled by the ligand. The SASA performed best overall but only because it is a restrictive method, always giving small estimates. We showed how the PCM method could be improved by supplementing it with energies from explicit simulations, but the set of test cases was too small to determine if this is a general approach.

Finally, in paper VII, we evaluated the normal-mode entropy usually employed in the MM/PBSA method. We showed that removing residues within 8 to 16 Å of the ligand but including a buffer region of residues and water molecules changes the absolute entropies by 1 to 5 kJ/mol on average. However, the error introduced is systematic and relative entropies

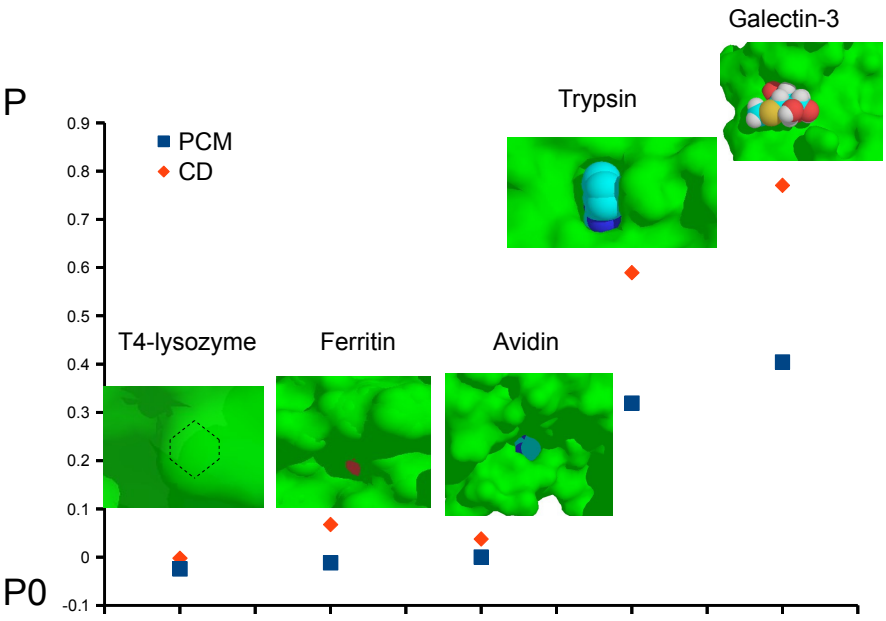


**Figure 5.3:** Non-polar solvation free energy using either PCM or CD for five complexes. The results are plotted on a relative scale; at 0 the continuum estimates are most accurate if the unbound site is void of water (P0), and at 1 the continuum estimates are most accurate if the unbound site is filled with water (P). The figures indicate the solvent exposure of the active site.

deviate not more than 1.5 kJ on average, which is within the statistical uncertainty. Hence, we have showed that it is more advantageous to use a truncated protein because it is much more efficient and gives better precision than calculations of a full protein.

In paper VIII, we compare MM/PBSA with LIE. Because we used only one test we could not make any definite conclusion about the accuracy of the method. Therefore, we concentrated on the precision of the two methods and the computer time required to obtain the estimates. It was found that LIE was more efficient than MM/PBSA. For a truncated simulation, LIE is about 2–7 times more efficient and for a full protein it is about 1.0–2.4 times more efficient. However, LIE requires a special setup of the simulation, e.g., neutralisation of amino-acid residues and this seems to affect the results. Moreover, LIE cannot easily be used to study four binding sites of avidin simultaneously.

We continued the comparison of approximate methods in Paper IX, by comparing variants of the MM/GBSA, MM/PBSA, and PDLD/s-LRA/$\beta$ method. We tested different ways to average energies from simulations, i.e., a one-average or three-average MM/PBSA or a LRA treatment which require simulations of the protein and free ligand both when the ligand is fully charged and when its charges have been zeroed. Furthermore, we tested if the methods could be improved by scaling the electrostatics energies with a dielectric constant. The study showed that the results are highly system dependent.

In paper X, I step away from protein system and look instead at host–guest complexes. Such complexes are interesting test systems because they are smaller whereas the binding is dictated by the same forces as in a protein system. I evaluated MM/PBSA, MM/GBSA, and LIE on their ability to reproduce experimental results for two sets of host–guest complexes. Because there is no tailor-made force field for such systems, I investigate the effect of three different charge schemes. The methods were very robust regarding the choice of charges and hence it is sufficient to use the cheap AM1-BCC charges. MM/GBSA reproduced experimental affinities well on both test systems, whereas LIE required optimisation of the non-polar part for one of the test systems.

In paper XI, we present our predictions to the Sampl3 challenge. The Sampl3 challenge is a blind test of affinity predictions for 34 ligands to the trypsin protein. We participated with estimates using MM/PB(GB)SA, (C)LIE, and Glide score. (In this study, we only used continuum electrostatics in CLIE). Unfortunately, none of the methods were able to give accurate results. However, many of the experimental affinity differences were not statistical significant and the range of experimental affinities were narrow. Removing pairs of ligands with differences that are not statistical significant, rather accurate ranking of the remaining ligands pairs could be obtained. Half of the ligands were considered non-binders because their affinity could not be measured experimentally. CLIE or LIE were either equal or worse than a random guess too discriminate between binders and non-binders. However, MM/PBSA and MM/GBSA were significantly better than chance. The Sampl3 challenge, also included a test of host–guest affinities. Here, MM/PBSA, MM/GBSA, LIE and CLIE performed well, whereas Glide score was slightly worse. This highlight that the methods are highly system-dependent.

## 5.2 Rigorous methods

In Paper XII, we move away from the approximate methods. The main objective is to determine an optimal TI protocol. As a test case, we use nine inhibitors to the factor Xa enzyme. We wanted to know how many intermediate states were necessary for an accurate estimate, how long simulations that are required, and if it was best to separate the electrostatic and van der Waals perturbation or not. We found that rather few intermediate states were required, five if we performed the electrostatic and van der Waals perturbation separately, and only three if the two perturbations were done simultaneously. For the free ligand simulations, about 2 ns of sampling was required, whereas the complex simulations only required 1 ns to converge. This protocol gave accurate results, except if we attempted a perturbation from a ligand with +2 net charge to a ligand with a +1 net charge. In such perturbation, we obtained errors compared to experiments on the order of 30 kJ/mol. Finally, we compared the TI results to MM/GBSA and found that TI is slightly more accurate and actually the more efficient method. The latter comes from the fact that TI is inherently more precise, whereas MM/GBSA requires averaging over many independent simulations to reach a similar precision.

We continued with the development of a rigorous protocol in paper XIII. Here, we investigate if we could go from a periodic system to a non-periodic system, and then truncate the non-periodic system. A comparison of the periodic and non-period simulations revealed a high degree of correlation and none of the differences were statistical significant. Therefore, we started to truncate the water droplet, and we could do that successfully for all ligands down to 15 Å radius, without introducing an error in the affinity estimate of more than 1 kJ/mol on average (see Figure 5.4). By truncating the system we could reduce the computational cost by a factor of 500.

Next, in paper XIV, we study another approach to improve the efficiency of rigorous calculations. In this approach, only a single reference molecule was simulated instead of all the other ligands. The idea is that the reference molecule should encompass all the other ligands of interest, i.e., it should be able to sample configurations important to many ligands and therefore it contains soft-core sites. After the simulation of the reference molecule, it is perturbed to all the real ligands. This approach was tested on the factor Xa inhibitors, and it was found that for non-polar perturbations the approach worked satisfactorily. However, for polar perturbations, the approach introduced large errors. The solution is to make several reference states.

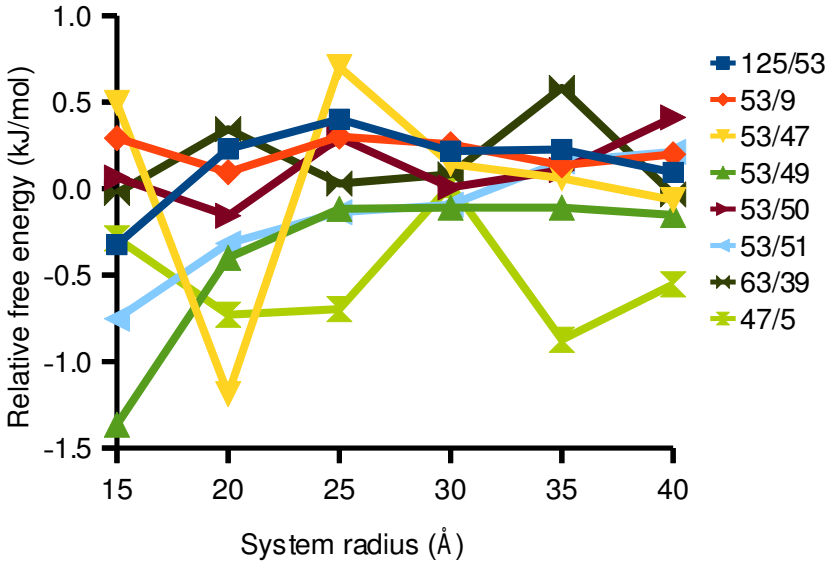


**Figure 5.4:** Free energy of transform one ligand into another when they are bound to the protein. Free energies are shown for different system radii, relative to a system with the entire protein.

The final paper, paper XV, is more about entropies than free energies. However, it compares statistical properties of entropy estimates with the properties of free energy estimates and therefore, it is included in this thesis. The title of the paper tells the objective: we want to find out if molecular dynamics simulations of protein will converge to an equilibrium state. This was motivated by our studies on conformational entropies and related quantities, where we have observed slow convergence of absolute entropy estimates. In this paper, we present a simple protein model that describes this behaviour and we perform 500 ns simulations of three protein systems. Because the entropy did not converge, we asked ourselves if other properties from MD simulations do converge. Therefore, we also computed MM/GBSA estimates for two ligands, and perform the first step in a FEP estimate of the electrostatic and van der Waals perturbation. We show that MM/GBSA fluctuates by 3 to 15 kJ/mol due to conformational changes. In contrast, the FEP estimate fluctuates less than 1 kJ/mol. Hence, from a statistical and thermodynamical perspective, an FEP approach is preferred.

## Concluding remarks

A little more than a decade ago, Kollman and co-workers declared that a new era of biomolecular simulations had begun: the era of structure and free-energy calculations. Owing to advances in simulation techniques such as particle-mesh Ewald summation, improved theories for computing free energies such as double-decoupling, and much better potentials for macromolecules, it was envisioned that simulations could be used on a routine basis to estimate ligand binding affinities. As a decade has past, it is pertinent to ask: Has this vision come through? Where are we now? And how do we move on from here? I believe that the research presented in this thesis can cast some light on these questions.

In the same paper where Kollman and co-workers declared the new era they, for the first time, published a unified presentation of the MM/PBSA method (although it had been used previously). Looking at the results of the research in this thesis, it is clear that MM/PBSA was not the answer to all problems. Although it is probably the most popular simulation-based method to estimate affinities, it suffers from huge problems. Perhaps the most clear and alarming problem was presented in paper VI, on the study of non-polar solvation free energies. The conclusion drawn in this paper is that MM/PBSA is unable to accurately predict the relative affinity between two ligands, if the structure of the ligand differs significantly. This immediately precludes one of the best uses of MM/PBSA, i.e., to predict relative affinities without concern of structural similarities of the ligands that is a requirement for many of the rigorous methods.

In my view, the LIE and PDLD/s-LRA$\beta$ methods suffer from similar

problems, although it is less obvious, because the non-polar solvation free energy is partly parametrised. Ultimately this come down to whether the method can handle a variable number of water molecules in the active site, both when different ligands are bound and when the protein is unbound. Because neither LIE nor PDLD/s-LRA$\beta$ requires a simulation of the unbound protein, they cannot consistently treat water expulsion. Additional terms that are based on a surface term have been suggested to improve LIE, but we have shown that such a method predicts the non-polar solvation free energy poorly.

Rigorous free energy methods based on TI or FEP can theoretically treat a variable number of water molecules, although it requires more sampling. Starting with methods that attempts to compute absolute free energies, because they could potentially remove the need for approximate methods such that MM/PBSA and LIE, the outlook for accurate and efficient estimates are promising at best. At the moment, they require an enormous amount of computer resources to reduce the uncertainty. In papers V and VI, we used such methods with very small ligands and already a ligand of more than 20 atoms, we had problems with the precision. This can of course be solved by prolonging the simulations, but waiting a month or more to obtain a precise estimate probably scares off most researchers. However, with increased computer resources, this could be feasible in the future.

However, using TI or FEP to compute relative affinities seems to be more promising. We have shown that the efficiency can be improved without loosing accuracy or precision. But of course these methods suffer from limitations themselves. One of them is that only small perturbations can be practically handled. Therefore, if it is of interest to study larger changes, several intermediates ligands have to be created. Another drawback of the TI and FEP methods, is that they cannot easily be used study changes in total charge of the system. For instance, it was not possible to accurately predict the relative affinity of a ligand with a charge +2 to a ligand with a charge +1. Hence, relative affinity methods work efficiently, but are restricted in their use.

The above concerns are mainly technical and some of them will probably be solved in the future. However, it is also of interest to note how well the methods do their job, i.e., how accurate are the available methods? And perhaps more interesting: What accuracy can we expect? What kind of systems can be studied? This thesis also has some guidelines to this end.

Looking at the Sampl3 challenge in paper XV, it is clear that none

of the approximate methods tested were able to accurately predict the affinities. However, we must consider the narrow range of the experimental affinities ($\sim$ 9 kJ/mol). A successful method on such a test case must have a high accuracy but also a very a high precision. The precision would need to be better then the experimental precision, which of course indicates that for such a test case the problem is only partly computational. As shown in that paper, only $\sim$ 20% of the experimental pairs had a free energy difference that were statistical significant.

In many projects, we have used avidin that is much better test case from a computational perspective. The free energy difference between many of the ligands is so large that we do not have to worry about the experimental uncertainty. However, it is less interesting from a real-life perspective because it would be a truly rare compound series of putative drugs that have a range of over 70 kJ/mol.

To conclude, I would say that the vision of Kollman and co-workers has not been fully realised. For some problems, free energy calculations can be routinely used, but it is a long way until we could put any two ligands in the computer, press a button, go home, sleep, and read off an accurate and precise estimate of the affinity the next morning.

# References

[1] World Health Organization, 2009, *AIDS epidemic update*, Geneva, Switzerland

[2] World Health Organization, 2011, *Causes Of Death 2008 Summary Tables*, Geneva, Switzerland

[3] World Health Organization, 2011, Essential medicines

[4] Kirkwood, J. G., 1935, Statistical mechanics of fluid mixtures, *J. Chem. Phys.*, 3:300–313

[5] Zwanzig, R. W., 1954, High-temperature equation of state by a perturbation method. I. Nonpolar gases, *J. Chem. Phys.*, 22:1420–1427

[6] McCammon, J., Gelin, J. B., Karplus, M., 1977, Dynamics of folded proteins, *Nature*, 267:585–590

[7] Karplus, M., 2006, Spinach on the ceiling: A theoretical chemist's return to biology, *Ann. Rev. Biophys. Biomol. Struct.*, 35:1–47

[8] Tembe, B. L., McCammon, J. A., 1984, Ligand–receptor interactions, *Comp. Chem.*, 8:281–283

[9] Wong, C. F., McCammon, J. A., 1986, Dynamics and design of enzymes and inhibitors, *J. Am. Chem. Soc.*, 108:3830–3832

[10] Beveridge, D. L., DiCapua, F. M., 1989, Free energy via molecular simulation: Applications to chemical and biomolecular systems, *Ann. Rev. Biophys. Biophys. Chem.*, 18:431–492

[11] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., Cheatham, T. E., 2000, Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models, *Acc. Chem. Res.*, 33:889–897

[12] Christ, C. D., Mark, A. E., van Gunsteren, W. F., 2010, Basic ingredients of free energy calculations: A review, *J. Comput. Chem.*, 31:1569–1582

[13] Gilson, M., Given, J., Bush, B., McCammon, J., 1997, The statistical-thermodynamic basis for computation of binding affinities: a critical review, *Biophys. J.*, 72:1047–1069

[14] Bjerrum, N., 1926, Untersuchungen über Ionenassoziation. I. Der Einfluss der Ionenassoziation auf die Aktivität der Ionen bei Mittleren Assoziationsgraden, *K Dan Vidensk Selsk*, 7:1–48

[15] Luo, H., Sharp, K., 2002, On the calculation of absolute macromolecular binding free energies, *Proc. Nat. Ac. Sci. U.S.A*, 99:10399–10404

[16] Woo, H.-J., Roux, B., 2005, Calculation of absolute protein–ligand binding free energy from computer simulations, *Proc. Nat. Ac. Sci. U.S.A*, 102:6825–6830

[17] Gohlke, H., Klebe, G., 2002, Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors, *Ang. Chem. Int. Ed.*, 41:2644–2676

[18] Huang, S.-Y., Grinter, S. Z., Zou, X., 2010, Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions, *Phys. Chem. Chem. Phys.*, 12:12899–12908

[19] Meng, E. C., Shoichet, B. K., Kuntz, I. D., 1992, Automated docking with grid-based energy evaluation, *J. Comput. Chem.*, 13:505–524

[20] Jones, G., Willett, P., Glen, R. C., Leach, A. R., Taylor, R., 1997, Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.*, 267:727–748

[21] Raha, K., Merz, K. M., 2004, A quantum mechanics-based scoring function: Study of zinc ion-mediated ligand binding, *J. Am. Chem. Soc.*, 126:1020–1021

[22] Rezac, J., Fanfrlik, J., Salahub, D., Hobza, P., 2009, Semiempirical quantum chemical PM6 method augmented by dispersion and H-bonding correction terms reliably describes various types of noncovalent complexes, *J. Chem. Theory Comput.*, 5:1749–1760

[23] Rarey, M., Kramer, B., Lengauer, T., Klebe, G., 1996, A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.*, 261:470–489

[24] Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., Mainz, D. T., 2006, Extra precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protei–ligand complexes, *J. Med. Chem.*, 49:6177–6196

[25] Gohlke, H., Hendlich, M., Klebe, G., 2000, Knowledge-based scoring function to predict protein-ligand interactions, *J. Mol. Biol.*, 295:337–356

[26] Mitchell, J. B. O., Laskowski, R. A., Alex, A., Thornton, J. M., 1999, BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential, *J. Comput. Chem.*, 20:1165–1176

[27] Chipot, C., Pohorille, A., 2007, *Free Energy Calculations*, Springer, Berlin

[28] Merz, K. M., 2010, Limits of free energy computation for protein–ligand interactions, *J. Chem. Theory Comput.*, 6:1769–1776

[29] Faver, J. C., Benson, M. L., He, X., Roberts, B. P., Wang, B., Marshall, M. S., Kennedy, M. R., Sherrill, C. D., Merz, K. M., 2011, Formal estimation of errors in computed absolute interaction energies of protein–ligand complexes, *J. Chem. Theory Comput.*, 7:790–797

[30] Salkind, N., 2007, *Encyclopedia of Measurement and Statistics*, Sage, Thousand Oaks, CA

[31] Pearlman, D. A., Charifson, P. S., 2001, Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system, *J. Med. Chem.*, 44:3417–3423

[32] Efron, B., 1981, Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods., *Biomet.*, 68:589–599

[33] Atkins, P., Friedman, R., 2005, *Molecular quantum mechanics*, Oxford University Press, Oxford, UK, 4th edition

[34] Leach, A., 2001, *Molecular modelling, principles and applications*, Pearson Education Ltd, Harlow, UK

[35] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., Stewart, J. J. P., 1985, Development and use of quantum mechanical molecular models. AM1: A new general purpose quantum mechanical molecular model, *J. Am. Chem. Soc.*, 107:3902–3909

[36] Rocha, G. B., Freire, R. O., Simas, A. M., Stewart, J. J. P., 2006, RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I, *J. Comput. Chem.*, 27:1101–1111

[37] Stewart, J., 2007, Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements, *J. Mol. Model.*, 13:1173–1213

[38] Dixon, S., Merz, K., 1997, Fast, accurate semiempirical molecular orbital calculations for macromolecules, *J. Chem. Phys.*, 107:879–893

[39] Stewart, J., 1997, Calculation of the geometry of a small protein using semiempirical methods, *J. Mol. Struct.-THEOCHEM*, 401:195–205

[40] Mackerell, A. D., 2004, Empirical force fields for biological macromolecules: Overview and issues, *J. Comput. Chem.*, 25:1584–1604

[41] Ponder, J. W., Case, D. A., 2003, Force fields for protein simulations, *Adv. Proteins Chem.*, 66:27–85

[42] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., Kollman, P. A., 1995, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.*, 117:5179–5197

[43] MacKerell, A. D., Bashford, D., Bellott, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998, All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B*, 102:3586–3616

[44] Jorgensen, W. L., Tirado-Rives, J., 1988, The OPLS [optimized potentials for liquid simulations] potential functions for proteins,

energy minimizations for crystals of cyclic peptides and crambin, *J. Am. Chem. Soc.*, 110:1657–1666

[45] Oostenbrink, C., Villa, A., Mark, A. E., Van Gunsteren, W. F., 2004, A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53A5 and 53A6, *J. Comput. Chem.*, 25:1656–1676

[46] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Weiner, P., 1984, A new force field for molecular mechanical simulation of nucleic acids and proteins, *J. Am. Chem. Soc.*, 106:765–784

[47] Mackerell, A. D., Feig, M., Brooks, C. L., 2004, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, *J. Comput. Chem.*, 25:1400–1415

[48] Israelachvili, J., 2010, *Intermolecular and surface forces*, Academic Press, USA

[49] Shirts, M. R., Mobley, D. L., Chodera, J. D., Pande, V. S., 2007, Accurate and efficient corrections for missing dispersion interactions in molecular simulations, *J. Phys. Chem. B*, 111:13052–13063

[50] Darden, T., York, D., Pedersen, L., 1993, Particle mesh Ewald - an N.Log(N) method for Ewald sums in large systems, *J. Chem. Phys.*, 98:10089–10092

[51] Lee, F. S., Warshel, A., 1992, A local reaction field method for fast evaluation of long-range electrostatic interactions in molecular simulations, *J. Chem. Phys.*, 97:3100–3107

[52] Simonson, T., 1993, Free energy of particle insertion, *Mol. Phys.*, 80:441–447

[53] Boresch, S., Bruckner, S., 2011, Avoiding the van der Waals endpoint problem using serial atomic insertion, *J. Comput. Chem.*, 32:2449–2458

[54] Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., van Gunsteren, W. F., 1994, Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations, *Chem. Phys. Lett.*, 222:529–539

[55] Zacharias, M., Straatsma, T. P., McCammon, J. A., 1994, Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration, *J. Chem. Phys.*, 100:9025–9031

[56] Åqvist, J., 2004, *Manual for the molecular dynamics package Q*, Uppsala, Sweden

[57] Steinbrecher, T., Joung, I., Case, D. A., 2011, Soft-core potentials in thermodynamic integration: Comparing one- and two-step transformations, *J. Comput. Chem.*, 32:3253–3263

[58] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., Case, D. A., 2004, Development and testing of a general Amber force field, *J. Comput. Chem.*, 25:1157–1174

[59] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., Mackerell, A. D., 2010, CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields, *J. Comput. Chem.*, 31:671–690

[60] Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., Oostenbrink, C., Mark, A. E., 2011, An automated force field topology builder (ATB) and repository: Version 1.0, *J. Chem. Theory Comput.*, in press

[61] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., Klein, M. L., 1983, Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.*, 79:926–935

[62] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Hermans, J., 1981, *Intermolecular forces*, Reidel, Dordrecht, 2nd edition

[63] Horn, H., Swope, W., Pitera, J., Madura, J., Dick, T., Hura, G., Head-Gordon, T., 2004, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew, *J. Chem. Phys.*, 120:9665–9678

[64] Klamt, A., 2005, *COMSO-RS. From Quantum Chemistry to fluid phase thermodynamics and drug design*, Elsevier

[65] Westergren, J., Lindfors, L., Höglund, T., Lüder, K., Nordholm, S., Kjellander, R., 2007, In silico prediction of drug solubility: 1. Free energy of hydration, *J. Phys. Chem. B*, 111:1872–1882

[66] Hirata, F., 2004, *Molecular theory of solvation*, Springer, Dordrecht

[67] Warshel, A., Russel, S., 1984, Calculations of electrostatics interactions in biological-systems and in solutions, *Quart. Rev. Biophys.*, 17:283–422

[68] Sharp, K. A., Honig, B., 1990, Electrostatic interactions in macromolecules: Theory and applications, *Ann. Rev. Biophys. Biophys. Chem.*, 19:301–332

[69] Fogolari, F., Brigo, A., Molinari, H., 2002, The Poisson–Boltzmann equation for bimolecular electrostatics: A tool for structural biology, *J. Mol. Recog.*, 15:377–392

[70] Bashford, D., Case, D. A., 2000, Generalized Born models of macromolecular solvation effects, *Ann. Rev. Phys. Chem.*, 51:129–152

[71] Cossi, M., Barone, V., 2007, *In: European summerschool in quantum chemistry 2007 book III*, Lund University, Lund, Sweden

[72] Miertuš, S., Scrocco, E., Tomasi, J., 1981, Electrostatic interaction of a solute with a continuum. A direct utilizaion of Ab Initio molecular potentials for the prevision of solvent effects, *Chem. Phys.*, 55:117–129

[73] Klamt, A., Schuurmann, G., 1993, COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, *J. Chem. Soc., Perk. Trans. 2*, pp. 799–805

[74] Tomasi, J., Mennucci, B., Cammi, R., 2005, Quantum mechanical continuum solvation models, *Chem. Rev.*, 105:2999–3094

[75] Barone, V., Cossi, M., Tomasi, J., 1997, A new definition of cavities for the computation of solvation free energies by the polarizable continuum model, *J. Chem. Phys.*, 107:3210–3221

[76] Pierotti, R. A., 1976, A scaled particle theory of aqueous and nonaqueous solutions, *Chem. Rev.*, 76:717–726

[77] Floris, F., Tomasi, J., 1989, Evaluation of the dispersion contribution to the solvation energy. a simple computational model in the continuum approximation, *J. Comput. Chem.*, 10:616–627

[78] Tan, C., Tan, Y.-H., Luo, R., 2007, Implicit nonpolar solvent models, *J. Phys. Chem. B*, 111:12263–12274

[79] Wang, J., Cai, Q., Ye, X., Hsieh, M.-J., Tan, C., Luo, R., 2010, *Amber Tools User's Manual, Version 1.4, 143-150*

[80] Hermann, R. B., 1972, Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area, *J. Phys. Chem.*, 76:2754–2759

[81] Sitkoff, D., Sharp, K. A., Honig, B., 1994, Accurate calculation of hydration free energies using macroscopic solvent models, *J. Phys. Chem.*, 98:1978–1988

[82] Marenich, A. V., Cramer, C. J., Truhlar, D. G., 2009, Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions, *J. Phys. Chem. B*, 113:6378–6396

[83] Kovalenko, A., Hirata, F., 1999, Self-consistent description of a metal-water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model, *J. Chem. Phys.*, 110:10095–10112

[84] Adcock, S. A., McCammon, J. A., 2006, Molecular dynamics: Survey of methods for simulating the activity of proteins, *Chem. Rev.*, 106:1589–1615

[85] Head, M. S., Given, J. A., Gilson, M. K., 1997, "Mining Minima": Direct computation of conformational free energy, *J. Phys. Chem. A*, 101:1609–1618

[86] Chen, W., Gilson, M. K., Webb, S. P., Potter, M. J., 2010, Modeling protein–ligand binding by mining minima, *J. Chem. Theory Comput.*, 6:3540–3557

[87] Frenkel, D., Smit, B., 2002, *Understanding molecular simulations*, Academic Press, San Diego, USA

[88] Newton, I., 1687, *Philosophiae Naturalis Principia Mathematica*

[89] Verlet, L., 1967, Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules, *Phys. Rev.*, 159:98–103

[90] Swope, W. C., Andersen, H. C., Berens, P. H., Wilson, K. R., 1982, A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters, *J. Chem. Phys.*, 76:637–649

[91] Hockney, R. W., 1970, The potential calculation and some applications, *Methods Comput. Phys.*, 9:136–211

[92] Ryckaert, J. P., Cicotti, G., Berendsen, H. J. C., 1977, Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes., *J. Comput. Phys.*, 23:327–341

[93] Hess, B., Bekker, H., Berendsen, H. J. C., Fraaije, J. G. E. M., 1997, LINCS: A linear constraint solver for molecular simulations, *J. Comput. Chem.*, 18:1463–1472

[94] Chandler, D., 1987, *Introduction to modern statistical mechanics*, Oxford University Press, New York USA

[95] Hünenberger, P. H., 2005, Thermostat algorithms for molecular-dynamics simulations, *Adv. Polym. Sci.*, 173:105–149

[96] Wu, X., Brooks, B. R., 2003, Self-guided Langevin dynamics simulation method, *Chem. Phys. Lett.*, 381:512–518

[97] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A., Haak, J. R., 1984, Molecular-dynamics with coupling to an external bath, *J. Chem. Phys.*, 81:3684–3690

[98] Brünger, A., Brooks III, C. L., Karplus, M., 1984, Stochastic boundary conditions for molecular dynamics simulations of ST2 water, *Chem. Phys. Lett.*, 105:495–500

[99] Marelius, J., Kolmodin, K., Feierberg, I., qvist, J. A., 1998, Q: a molecular dynamics program for free energy calculations and empirical valence bond simulations in biomolecular systems, *J. Mol. Graph. Model.*, 16:213–225

[100] Essex, J. W., Jorgensen, W. L., 1995, An empirical boundary potential for water droplet simulations, *J. Comput. Chem.*, 16:951–972

[101] King, G., Warshel, A., 1989, A surface constrained all-atom solvent model for effective simulations of polar solutions, *J. Chem. Phys.*, 91:3647–3661

[102] Im, W., Berneche, S., Roux, B., 2001, Generalized solvent boundary potential for computer simulations, *J. Chem. Phys.*, 114:2924–2937

[103] Banavali, N., Im, W., Roux, B., 2002, Electrostatic free energy calculations using the generalized solvent boundary potential method, *J. Chem. Phys.*, 117:7381–7388

[104] Simonson, T., Archontis, G., Karplus, M., 1997, Continuum treatment of long-range interactions in free energy calculations. application to protein–ligand binding., *J. Phys. Chem. B*, 101:8349–8362

[105] Grossfield, A., Zuckerman, D. M., 2009, Quantifying uncertainty and sampling quality in biomolecular simulations, *Ann. Rep. Comput. Chem.*, 5:23–48

[106] Madsen, H., 2008, *Time Series Analysis*, Chapman & Hall, CRC, New York USA

[107] Steinbrecher, T., Mobley, D. L., Case, D. A., 2007, Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations, *J. Chem. Phys.*, 127:214108

[108] Zwansig, R., Ailawadi, K., 1969, Statistical error due to finite time averaging in computer experiments, *Phys. Rev.*, 1982:280–282

[109] Lu, C.-Y., Vanden Bout, D. A., 2006, Effect of finite trajectory length on the correlation function analysis of single molecule data, *J. Chem. Phys.*, 125:124701

[110] Bishop, M., Frinks, S., 1987, Error analysis in computer-simulations, *J. Chem. Phys.*, 87:3675–3676

[111] Yang, W., Bitetti-Putzer, R., Karplus, M., 2004, Free energy simulations: Use of reverse cumulative averaging to determine the equilibrated region and the time required for convergence, *J. Chem. Phys.*, 120:2618–2628

[112] Lawrenz, M., Baron, R., McCammon, J. A., 2009, Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: Determinants of H5N1 avian influenza virus

neuraminidase inhibition by peramivir, *J. Chem. Theory Comput.*, 5:1106–1116

[113] Fujitani, H., Tanida, Y., Ito, M., Jayachandran, G., Snow, C., Shirts, M., Sorin, E., Pande, V., 2005, Direct calculation of the binding free energies of FKBP ligands, *J. Chem. Phys.*, 123:084108

[114] Sadiq, S. K., Wright, D. W., Kenway, O. A., Coveney, P. V., 2010, Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases, *J. Chem. Inf. Model.*, 50:890–905

[115] Loccisano, A. E., Acevedo, O., DeChancie, J., Schulze, B. G., Evanseck, J. D., 2004, Enhanced sampling by multiple molecular dynamics trajectories: carbonmonoxy myoglobin 10 $\mu$s$\mathrm{A}_0 \rightarrow \mathrm{A}_{1-3}$ transition from ten 400 picosecond simulations, *J. Mol. Graph. Model.*, 22:369–376

[116] Worth, G. A., Nardi, F., Wade, R. C., 1998, Use of multiple molecular dynamics trajectories to study biomolecules in solution: The YTGP peptide, *J. Phys. Chem. B*, 102:6260–6272

[117] Li, Y., Liu, Z., Wang, R., 2010, Test MM-PB/SA on true conformational ensembles of protein-ligand complexes, *J. Chem. Inf. Model.*, 50:1682–1692

[118] Grossfield, A., Feller, S. E., Pitman, M. C., 2007, Convergence of molecular dynamics simulations of membrane proteins, *Proteins*, 67:31–40

[119] Bruckner, S., Boresch, S., 2011, Efficiency of alchemical free energy simulations. II. Improvements for thermodynamic integration, *J. Comput. Chem.*, 32:1320–1333

[120] Bruckner, S., Boresch, S., 2011, Efficiency of alchemical free energy simulations. I. A practical comparison of the exponential formula, thermodynamic integration, and Bennett's acceptance ratio method, *J. Comput. Chem.*, 32:1303–1319

[121] Jorgensen, W. L., Thomas, L. L., 2008, Perspective on free-energy perturbation calculations for chemical equilibria, *J. Chem. Theory Comput.*, 4:869–876

[122] Bennet, C. H., 1976, Efficient estimation of free energy differences from Monte Carlo data, *J. Comput. Phys.*, 22:245–268

[123] Shirts, M. R., Bair, E., Hooker, G., Pande, V. S., 2003, Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods, *Phys. Rev. Lett.*, 91:140601

[124] Shirts, M. R., Chodera, J. D., 2008, Statistically optimal analysis of samples from multiple equilibrium states, *J. Chem. Phys.*, 129:124105

[125] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., Kollman, P. A., 1992, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method, *J. Comput. Chem.*, 13:1011–1021

[126] Shirts, M. R., Mobley, D. L., Chodera, J. D., 2007, Alchemical free energy calculations: ready for prime time?, *Ann. Rep. Comput. Chem.*, 3:41–59

[127] Deng, Y., Roux, B., 2006, Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant, *J. Chem. Theory Comput.*, 2:1255–1273

[128] Pranata, J., Jorgensen, W. L., 1991, Monte-Carlo simulations yield absolute free energies of binding for guanine-cytosine and adenine-uracil base pairs in chloroform., *Tetrahedron*, 47:2491–2501

[129] Deng, Y., Roux, B., 2009, Computations of standard binding free energies with molecular dynamics simulations, *J. Phys. Chem. B*, 113:2234–2246

[130] General, I. J., 2010, A note on the standard state's binding free energy, *J. Chem. Theory Comput.*, 6:2520–2524

[131] Wang, J., Deng, Y., Roux, B., 2006, Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials, *Biophys. J.*, 91:2798–2814

[132] Boresch, S., Tettinger, F., Leitgeb, M., Karplus, M., 2003, Absolute binding free energies: A quantitative approach for their calculation, *J. of Phys. Chem. B*, 107:9535–9551

[133] Michel, J., Essex, J., 2010, Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations, *J. Comput.-Aided Mol. Des.*, 24:639–658

[134] Boresch, S., Karplus, M., 1998, The role of bonded terms in free energy simulations: 1. Theoretical analysis, *J. Phys. Chem. A*, 103:103–118

[135] Wu, D., Kofke, D. A., 2005, Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation, *J. Chem. Phys.*, 123:054103

[136] Lee, M. S., Olson, M. A., 2006, Calculation of absolute protein-ligand binding affinity using path and endpoint approaches, *Biophys. J.*, 90:864–877

[137] Åqvist, J., Hansson, T., 1996, On the validity of electrostatic linear response in polar solvents, *J. Phys. Chem.*, 100:9512–9521

[138] Åqvist, J., Medina, C., Samuelsson, J.-E., 1994, A new method for predicting binding affinity in computer-aided drug design, *Proteins Eng.*, 7:385–391

[139] Hansson, T., Marelius, J., Åqvist, J., 1998, Ligand binding affinity prediction by linear interaction energy methods, *J. Comput.-Aided Mol. Des.*, 12:27–35

[140] Almlöf, M., Carlsson, J., Åqvist, J., 2007, Improving the accuracy of the linear interaction energy method for solvation free energies, *J. Chem. Theory Comput.*, 3:2162–2175

[141] Carlsson, J., Boukharta, L., Åqvist, J., 2008, Combining docking, molecular dynamics and the linear interaction energy method to predict binding modes and affinities for non-nucleoside inhibitors to HIV-1 reverse transcriptase, *J. Med. Chem.*, 51:2648–2656

[142] Foloppe, N., Hubbard, R., 2006, Towards predictive ligand design with free-energy based computational methods?, *Curr. Med. Chem.*, 13:3583–3608

[143] Wang, W., Wang, J., Kollman, P. A., 1999, What determines the van der Waals coefficient $\beta$ in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations?, *Proteins*, 34:395–402

[144] Carlson, H. A., Jorgensen, W. L., 1995, An extended linear response method for determining free energies of hydration, *J. Phys. Chem.*, 99:10667–10673

[145] Almlöf, M., Brandsdal, B. O., Åqvist, J., 2004, Binding affinity prediction with different force fields: Examination of the linear interaction energy method, *J. Comput. Chem.*, 25:1242–1254

[146] Ostrovsky, D., Udier-Blagović, M., Jorgensen, W. L., 2003, Analyses of activity for factor Xa inhibitors based on Monte Carlo simulations, *J. Med. Chem.*, 46:5691–5699

[147] Warshel, A., Sharma, P. K., Kato, M., Parson, W. W., 2006, Modeling electrostatic effects in proteins, *Biochim. Biophys. Acta*, 1764:1647–1676

[148] Muegge, I., Tao, H., Warshel, A., 1997, A fast estimate of electrostatic group contributions to the free energy of protein-inhibitor binding, *Proteins Eng.*, 10:1363–1372

[149] Sham, Y. Y., Chu, Z. T., Tao, H., Warshel, A., 2000, Examining methods for calculations of binding free energies: LRA, LIE, PDLD-LRA, and PDLD/S-LRA calculations of ligands binding to an hiv protease, *Proteins*, 39:393–407

[150] Singh, N., Warshel, A., 2010, Absolute binding free energy calculations: On the accuracy of computational scoring of protein–ligand interactions, *Proteins*, 78:1705–1723

[151] Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., Case, D. A., 1998, Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate–DNA helices, *J. Am. Chem. Soc.*, 120:9401–9409

[152] Massova, I., Kollman, P., 2000, Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding, *Pers. Drug Discov. Des.*, 18:113–135

[153] Swanson, J. M., Henchman, R. H., McCammon, J. A., 2004, Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy, *Biophys. J.*, 86:67–74

[154] Zhou, R., Friesner, R. A., Ghosh, A., Rizzo, R. C., Jorgensen, W. L., Levy, R. M., 2001, New linear interaction method for binding affinity calculations using a continuum solvent model, *J. Phys. Chem. B*, 105:10388–10397

[155] Huang, D., Caflisch, A., 2004, Efficient evaluation of binding free energy using continuum electrostatics solvation, *J. Med. Chem.*, 47:5791–5797

[156] Carlsson, J., Andér, M., Nervall, M., Åqvist, J., 2006, Continuum solvation models in the linear interaction energy method, *J. Phys. Chem. B*, 110:12034–12041

[157] Schutz, C. N., Warshel, A., 2001, What are the dielectric "constants" of proteins and how to validate electrostatic models?, *Proteins*, 44:400–417

[158] Söderhjelm, P., Kongsted, J., Ryde, U., 2010, Ligand affinities estimated by quantum chemical calculations, *J. Chem. Theory Comput.*, 6:1726–1737

[159] Carlsson, J., Å qvist, J., 2006, Calculations of solute and solvent entropies from molecular dynamics simulations, *Phys. Chem. Chem. Phys.*, 8:5385–5395

## Acknowledgments

This might be the most difficult part to write because I cannot consult any books but have to draw the acknowledgements from memory. Therefore, to allow for any blanks, I start by expressing my gratitude to all people that I have stumble upon the last three and a half year; in some way you have shaped who I am, and thereby this thesis.

Den första person som jag skulle vilja nämna vid namn är **Ulf**, min oumbärlige handledare. Jag vill tacka dig för du trodde på mig så mycket att du väntade ett halvår på att jag skulle avsluta mina masterstudier, jag tror inte du ångrar ditt val. Jag är också tacksam för att du har fått mig att inse att det här med att vara forskare är det bästa jobb som finns i denna värld.

Jag skulle också vilja tacka min bihandledare, **Mikael**. Det har varit ett sant nöje att få lära känna hur en experimentallist arbetar och tänker. Det är en erfarenhet jag kommer bära med mig väldigt länge.

Next, I would like to mention past and present members of the bio group: **Paulius** for being a great friend, for all the fun both privately and professionally, for room-sharing at last. **Pär** för att du har varit en oavsiktlig förebild och för alla goda idéer. **Jimmy** för att du introducerade mig på avdelningen. **Antonella**, **Jacob**, **LiHong**, **Jon**, and **Mehdi** for nice collaborations, and **Marie-Celine** for being a great friend and initiating Thursday-meetings. **Markus**, **Maryam**, **Esko**, **Jilai**, and **Jenny** for support and for nice discussions.

**Carl** för alla goda samarbeten, både de som blev av och de som fortfarande ligger stilla, för allt stöd, och för kul på FLÄK-mötena. **Asbjörn** and **Jonas**, för fantastiskt flummiga, pseudointellektuella diskussioner

vid luncher och fikapauser. **Marta K**, for a lot of great laughs. **Svante** för ditt språkliga kunnande och för din passion för sällskapsspel. **Mickael** for being a great friend, too bad you moved away. **Martin**, **Björn**, **Maxime**, **Segad**, **Anıl**, **Ryan**, **Fei**, **Camille**, **Axel**, **Björn**, and **Chris** for being there. **Johanna** för att du gjorde en våtlabb relativt rolig och för alla kakorna. **Gleb** for nice chats about science and everything else.

**Ingrid**, **Eva**, och **Bodil** för all hjälp med administrativa ting. **Marie** för att du är en frisk fläkt, för en kort tid av rumsdelande. **Gunnar**, **Bo** och **Per-Åke** för alla utmanande frågor under mina avdelningsseminarier, lärorika kurser och historieberättande på fikapauser. **Magnus** för hjälp med lunarc-relaterade saker. **Torbjörn**, **Per-Olof**, **Petter**, **Jan**, **Mikael**, och **Valera** för hjälp genom åren och för bra föreläsningar. **Olov** på slutet.

All the students at the FLÄK research school for actually quite a lot of fun at Brösarp and Grand Hotel. Tack till **Per**, för dessa möjligheter. Tack till FLÄK för finansiering.

**Ulf N** för support och goda idéer för diverse galectin-project, för gott samarbete med kursen i läkemedelskemi. **Kristoffer** för hjälp med NMR teorier, **Derek** och **Saraboji** för samarbetet med vattnet.

During the course of these years, I have not been isolated to Lund, and I would like to send out regards to the MD group at University of Queensland. To **Alan**, who took me in and believed in my capabilities so much that you mentioned PostDoc the first day. **Alpesh** for suggesting a collaboration, your guidance and your knowledge. **David** and **Meg** for showing me a PostDoc-life, and for nice discussions at the lunch break.

**Marta S** for coming to Sweden and brighten the mood at department, for broadening my interest in molecular association. **Jose** for a good collaboration.

Jag har också haft det stora nöjet ett arbeta nära med personer på AstraZeneca både i Mölndal, Södertälje och den numera nedlagda Lundasiten. Tack **Ingemar** för ditt intresse och för din enorma kunskap om intressanta system. **Lars** för goda samarbeten med host–guest systemen, och för trevliga samtal. **Ola** för feedback. **Göran**, **Johan**, och **Matti** för gott samarbetet och för feedback.

På andra sidan sundet skulle jag vilja tacka **Patrik** för gott samarbete och för att du fanns där på min första konferens, utan dig skulle jag varit vilse. Tack **Lars** for gott samarbete.

I Uppsala skulle jag vilja tacka **Johan Å** för att du redde ut saker angående LIE och Q. Tack till **Göran** för hjälp med Q och för veckan i Les-Diablerets. Tack till **Johan S** och **Lars** för alla svar angående LIE

och Q.

Without a lot of computer resources this thesis would not be half as long, so thanks people at **Lunarc** in Lund, **HPCN** in Umeå, **C3SE** in Göteborg, and **NSC** in Linköping for allocation and support. Thanks to **Andvare**, **Balmung**, **Docenten**, **Milleotto**, **Akka**, **Platon**, **Beda**, **Neolith**, and **Lychee** for keeping up with me.

Jag skulle vilja tacka **Leif** för att du fick mig att tro på vetenskap igen och för att du vidarebefordrade ett mejl om en viss utlysning i Lund. Tack också **Ylva** för att du fick mig att läsa kemi på gymnasiet.

Tack till **Karin** för att du hjälpte till att bredda min kunskap om folkhälsa.

Tack till min släkt för allt stöd. Ett särskilt tack går ut till dig **Farmor**, för att du alltid har varit nyfiken på vetenskap och försökt förstå vad jag gör. Jag vill också tacka **svärföräldrarna** för stöd och nyfikenhet. Tack bonusföräldrarna, **Helene** och **Morgan**. Tack **mor** och **far**.

Till sist **Maja**. Som har gjort denna resa tillsammans med mig. Jag älskar dig.